



Proceedings

iPres 2022 Glasgow

12—16 September 2022

www.ipres2022.scot



iPRES 2022
GLASGOW

18th International Conference
On Digital Preservation



iPres 2022 was proudly hosted by the Digital Preservation Coalition (DPC)



Proceedings of the 18th International Conference on Digital Preservation 2022.

The iPres 2022 conference proceedings are published under a Creative Commons license. With the exception of any logos, emblems, trademarks or other nominated third-party images/ text, this work is available for re-use under a Creative Commons Attribution 4.0 International license (CC-BY 4.0). Further details about CC BY licenses are available at <https://creativecommons.org/licenses/by/4.0/>.

All external links were active at the time of publication unless otherwise stated.

These proceedings contain the published and peer-reviewed submissions of the 18th International Conference on Digital Preservation. All other materials of the conference will be published on the Open Science Framework (OSF) iPres 2022 Conference pages: <http://doi.org/10.7207/ipres2022-resources>

The OSF proceedings contain all submitted papers, panels, posters, workshops, tutorials, ad-hoc proposals, Digital Preservation Bake Off Challenge contributions, as well as presenters' slides, optional additions and the collaborative notes taken during the conference.

The majority of the presentations at iPres 2022 have been recorded and the three iPres 2022 Keynote presentations have been published on World Digital Preservation Day 2022. They are now available on: <http://doi.org/10.7207/ipres2022-recordings>.

The remaining recordings will be made public in January 2023 and will be accessible on the same page.

The Conference Photo albums are available on Flickr: <https://www.flickr.com/photos/dpconflickr/albums>



The DPC is very grateful to the following individuals who contributed their time to support the organization of the iPres 2022 Conference

The iPres 2022 Program Committee was about 50% larger than in previous years, with the explicit purpose of including a new generation of leaders.

The Committee was helped by a huge army of reviewers; and they have in turn worked through a massive outpouring of contributions and proposals; and for this we are very grateful.

Conference Organization

Program Committee

General Chair

William Kilbride*, DPC

Administrators

John McMillan, DPC

Ellie O'Leary, DPC

Communications Team

Ali Hayes-Brady, Monash University

Sarah Middleton, DPC

Angela Puggioni, DPC

Tracy Seneca*, University of Illinois at Chicago

Workshops & Tutorials Chairs

Arif Shaon, Qatar National Library

Holly Wright, Archaeology Data Service

Posters & Lightning Talk Chairs

James Doig, National Archives of Australia

Elizabeth Kata, International Atomic Energy Agency

Hannah Smith, Historic Environment Scotland

Papers & Panels Chairs

Micky Lindlar, TIB - German National Library of Science and Technology

James Lowry, Queens College, City University of New York

Tricia Patterson, Harvard University

Michael Popham, DPC

Marcel Ras*, Vrije Universiteit Amsterdam

Legacy Chairs

Lorna Hughes, University of Glasgow

Garth Stewart, National Records of Scotland

Hackathon Chairs

Kieran O'Leary, National Library of Ireland

Klaus Rechert*, University of Freiburg

Leontien Talboom, University College London, The National Archives UK

Remote Participation Chairs

Belinda Chan, National Library Board, Singapore

Alina Karlos, University of Namibia

Shira Peltzman, University of California Los Angeles

First Time Participation Chairs

Caroline Catchpole, The National Archives UK

Alexis Tindall, University of Adelaide Library

Heather Tompkins, Library and Archives Canada

Jaye Weatherburn, University of Melbourne

Committee Members

Euan Cochrane, Yale University Library

Helen Hockx-Yu, University of Notre Dame

Irina Hope, Digital Curation Centre

Ruby Martinez, University of Illinois

Kai Naumann, State Archives of Baden-Württemberg

David Underdown, The National Archives UK

Executive Board

Kate Murray, Library of Congress

Tim Keefe, Chester Beatty

(*Indicates additional membership of the iPres Steering Conference Group)



Local Organizing Committee

Jo Brown, InConference
William Kilbride, DPC
Amy Love, InConference
John McMillan, DPC
Sharon McMeekin, DPC
Sarah Middleton, DPC
Ellie O’Leary, DPC
Angela Puggioni, DPC
Louise Watson, InConference

Reviewers

Jefferson Bailey, Internet Archive
Bertrand Caron, Bibliothèque Nationale De France
Caroline Catchpole, The National Archives (UK)
Zijun Chen, National Science Library, Chinese Academy of Sciences
Gang Chen, Institute of High Energy Physics, Chinese Academy of Sciences
Sally Cholewa, Natwest Group
David Cirella, Yale University
Euan Cochrane, Yale University Library
Louise Curham, Charles Sturt University
Amy Currie, DPC
Steve Daly, The National Archives (UK)
James Doig, National Archives of Australia
Elizabeth England, U.S. National Archives (NARA)
Claire Fox, Yale University Library
Steffen Fritz, DLA Marbach
Andrea Goethals, National Library of New Zealand



Makoto Goto, National Museum of Japanese History
Edith Halvarsson, Bodleian Libraries
Karen Hanson, Portico
Ali Hayes-Brady, Monash University
Heikki Helin, CSC - It Center For Science
Patricia Herterich, DCC, University of Edinburgh
Helen Hockx-Yu, University of Notre Dame
Inge Hofsink, National Library of The Netherlands
Lorna Hughes, University of Glasgow
Perla Innocenti, University of Strathclyde
Neill Jeffries, Bodleian Libraries, University of Oxford
Leslie Johnston, US National Archives and Records Administration
Catherine Jones, UKRI/STFC - Energy Data Center
Peter Judge, Lloyds Banking Group
Alina Karlos, University of Namibia
Elizabeth Kata, International Atomic Energy Agency
Tim Keefe, Chester Beatty
William Kilbride, Digital Preservation Coalition
Wachiraporn Klungthanaboon, Chulalongkorn University

Leo Konstantelos, University of Glasgow
Nick Krabbenhoeft, New York Public Library
Santhilata Kuppli Venkata, The National Archives (UK)
Richard Lehane, International Atomic Energy Agency
Micky Lindlar, TIB - German National Library of Science and Technology
Rowena Loo, National Archives of Australia
James Lowry, Queens College, City University of New York
Ruby Martinez, University of Illinois
Sharon McMeekin, DPC
Anna McNally, University of Westminster
Jenny Mitcham, DPC
Kate Murray, Library of Congress
Kiyonori Nagasaki, International Institute for Digital Humanities
Kai Naumann, Landesarchiv Baden-württemberg
Joshua Ng, Archives New Zealand
Eleanor O'Leary, DPC
Kieran O'Leary, National Library of Ireland
Jack O'Sullivan, Preservica
Kevin Palendat, Library And Archives Canada
Natalie Pang, National University of Singapore



Tricia Patterson, Harvard University
Shira Peltzman, UCLA Library
Michael Popham, DPC
Abigail Potter, Library of Congress
Marcel Ras, Vrije Universiteit Amsterdam
Klaus Rechert, University of Freiburg
Jonas Recker, GESIS - Leibniz-Institut für Sozialwissenschaften
Judith Rog, National Library of The Netherlands
Sarah Romkey, Artefactual Systems Inc
Robin Ruggaber, University of Virginia Library
Daisy Selematsela, University of the Witwatersrand Library
Tracy Seneca, University of Illinois
Arif Shaon, Qatar National Library
Linda Shave
Hania Smerecka, Lloyds Banking Group
Sharon Smith, Library and Archives Canada
Hannah Smith, Historic Environment Scotland
Caylin Smith, Cambridge University Library
Teresa Soleau, J. Paul Getty Trust
Ross Spencer, Ravensburger Ag
Garth Stewart, National Records of Scotland
Paul Stokes, Jisc
Lance Stuchell, University of Michigan Library
Shoaib Sufi, University of Manchester / Software Sustainability Institute
Dhani Sugiharto, National Archives of the Republic of Indonesia
Shigeo Sugimoto, University of Tsukuba
Leontien Talboom, University College London,

The National Archives
Melissa Terras, University of Edinburgh
Grace Thomas, Library of Congress
Nicole Thorne-Vicatos, Townsville Hospital and Health Service
Tyler Thorsted, Church of Jesus Christ of Latter-day Saints
Alexis Tindall, University of Adelaide Library
Heather Tompkins, Library and Archives Canada
David Underdown, The National Archives
Remco van Veenendaal, National Archives of The Netherlands
Tamara van Zwol, Netherlands Institute for Sound and Vision
Natalie Vielfaure, University of Manitoba
Jaye Weatherburn, University of Melbourne
Marcin Werla, Qatar National Library
Jan Whalen, University of Manchester
Paul Wheatley, DPC
Lotte Wijsman, National Archives of The Netherlands
Tobias Wildi, University of Applied Sciences of the Grisons
Kevin Wong, National Archives of Singapore
Holly Wright, Archaeology Data Service
Zhenxin Wu, National Science Library, Chinese Academy of Sciences
Zhongming Zhu, National Science Library, Chinese Academy of Sciences
Eld Zierau, Royal Danish Library





iPRES 2022
G L A S G O W

Technology & Innovation Centre
12–16 SEPTEMBER 2022

The DPC is very grateful to the following organizations who sponsored iPres 2022



国家数字科技文献资源长期保存体系
National Digital Preservation Program





Table of contents

1	Proceedings
2	Acknowledgements
6	Sponsors
7	Table of contents
17	Introduction

Long Papers

28	It Takes a Whole Village to Define a Preservation Strategy: Formalizing Policies on Data Formats Normalization at the National Library of France Mr Bertrand Caron, Alix Bruys, Thomas Ledoux, Jordan de La Houssaye
40	Useable Software Forever. The Emulation as a Service Infrastructure (EaaS) Program of Work Euan Cochrane, Dr Klaus Rechert, Jurek Oberhauser, Seth Anderson, Claire Fox, Ethan Gates
53	How do users discover digital preservation tools? Report on a survey of professionals Dr Amber Cushing, Tess Burchmore, Sarah Conroy, Phoebe Doyle, Niamh Hegarty, Rebecca Kelly, Payton Kufeldt, Morgan McGann, Cian Ormond, Gerard Quine, Martina Reba, Ronan Woods
61	Appraisal and Selection on a Long-term Preservation Repository? Can you repeat that, please? Luis Faria, Miguel Guimarães, Miguel Ferreira
70	A Generic Emulator Interface for Digital Preservation --- Towards a Collaborative Distributed Emulator Registry Rafael Gieschke, Klaus Rechert
81	Preservation Strategies for New Forms of Scholarship Deb Verhoff, Karen Hanson, Jonathan Greenberg
89	Green Goes with Anything: Decreasing Environmental Impact of Digital Libraries at Virginia Tech Alex Kinnaman, Alan Munshower
99	Ain't No Mountain High Enough: Developing a New Skills Framework for Digital Preservation Sharon McMeekin, Dr Amy Currie
108	Resilience of Internet Art Supported by Executable Archive Principles Case-study of Flash & VMRL Artwork Dr Natasa Milic-Frayling, Michael Takeo Magruder
121	"We're all doing the best we can with what we've got": Preservation practices of Data Curation Network members Hoa Luong, Mikala Narlock, Jon Petters
145	Feasible, Adaptable and Shared: A call for a community framework for implementing ML and AI Dr Meghan Ferriter, Abigail Potter, Eileen Jakeway Manchester, Jaime Mears



153 A Digital Preservation Wikibase

Dr Katherine Thornton, Kenneth Seals-Nutt

162 Metadata Quality in Digital Libraries: An Analysis of Survey Response Data

Hannah Tarver, Meredith Hale, Rachel White, Steven Gentry, Madison Chartier, Rachel Wittmann

173 Making Risk Modeling Accessible With DiAGRAM

David Underdown, Alexandra Leigh, Pauline Descheemaeker

184 E-ARK, Ten years and still going strong: Results, Use Cases and Benefits.

Carl Wilson, Janet Anderson, Dr David Anderson, Dr Jaime Kaminski, Dr Diogo Proença, István Alföldi

193 Construction of a Benchmark Model for Long-Term Preservation Value Evaluation of Academic Information on Social Media

Liu Hui, Zhang Dongrong



Short Papers

- 202 Cultivating the Scientific Data of the Morrow Plots: Visualization and Data Curation for a Long-term Agricultural Experiment**
Bethany Anderson, Sandi Caldrone, Joshua Henry, Heidi Imker, Hoa Luong, Kelli Trei, Sarah Williams
- 208 It Takes a Village in Practice: Growing Communities During a Pandemic**
Megan Forbes, Laurie Arp
- 211 Vault: Building an Extensible, Affordable Digital Preservation & Repository Service**
Jefferson Bailey
- 215 Repository Speed Dating A methodology for narrowing the field**
Sven Schlarb, Karin Bredenberg, Carl Wilson
- 220 OAIS-compliant digital archiving of research and patrimonial data in DNA**
Pierre-yves Burgi, Jan Krause, Linda Meiser, Dina Andriamahady, Hugues Cazeaux, Lamia Friha, Basma Makhoul Shabou
- 225 DNA Data storage for long term digital preservation**
Euan Cochrane, Daniel Chadash
- 231 The design and implementation of a necessary and sufficient system for the long-term archival retention of digital documents**
Dr Viv Cothey, Claire Collins
- 235 EMA: Brazilian Cultural Heritage Image Dataset - Towards AI-based metadata annotation of digital collections**
Vagner Inácio de Oliveira, Paula Dornhofer Paro Costa, Dalton Martins
- 240 Optimizing Memory for Legacy DOS Systems**
Dr Denise de Vries
- 244 Metadata Encoding and Transmission Standard (METS) Version 2**
Karin Bredenberg, Aaron Elkiss, Inge Hofsink, Juha Lehtonen, Andreas Nef, Tobias Steinke, Robin Wendler
- 250 Access Quality Metrics for Net Art**
Dragan Espenschied, Lyndsey Moulds, Xiao Ma
- 255 Digital Preservation Pipeline for Data Storage Media At The Cinémathèque Suisse**
Robin François, Rebecca Rochat
- 260 Data curation and agroecology: examining data requirements for short supply chains**
Dr Sarah Higgins, Christopher I. Higgins
- 266 Design Patterns in Digital Preservation**
Dr Andrew Jackson



- 271 Passive Digital Preservation on Paper in Practice**
Vincent Joguin, Jean-Noël Dumont
- 277 Robustifying Links with Zotero**
Martin Klein, Shawn M. Jones
- 282 Evaluating Digital Preservation Capability with Large at-risk Collections: Lessons learnt from preserving the NVA Archive**
Emma Yan, Dr Leo Konstantelos, Clare Paterson
- 287 Archivemata-EPrints Integration: Developing digital preservation capacity for open repositories**
Tomasz Neugebauer, Sarah Lake
- 293 Mapping the Landscape of Digital Preservation Networks The nestor Digital Preservation Community survey**
Micky Lindlar, Svenia Pohlkamp, Monika Zarnitz, Thomas Bähr, Stefan Strathmann
- 300 Developing an approach for archiving Digital Audio Workstation projects: A pilot study**
Valerie Love, Dr Michael Brown
- 305 Going for Gold or Good Enough? Observations on three years of benchmarking with DPC RAM**
Jenny Mitcham, Paul Wheatley
- 311 Monitoring Bodleian Libraries' Repositories with Micro Services**
Edith Halvarsson, James Mooney, Sebastian Lange
- 315 From Ray Cats TO DPC RAM: How Best to Preserve a Digital Memory of the Nuclear Decommissioning Process**
Michael Popham, Jenny Mitcham
- 319 Caring for Born Digital Video Camera Original Formats: Considering Intentional Change**
Crystal Sanchez
- 324 Do We Really Know Our Data? Assessing File Format Policy Compliance and Digital Preservation Tenability via a New Software Tool**
Tom Smyth
- 329 The 2022 Revision of the PREMIS Rights Entity**
Marjolein Steeman, Karin Bredenberg, Bertrand Caron, Leslie Johnston, Michelle Lindlar, Jack O'Sullivan, Sarah Romkey
- 334 "...provide a lasting legacy for Glasgow and the nation" Two years of transferring Scottish Cabinet records to National Records of Scotland**
Garth Stewart
- 340 Seeking Sustainability: Developing a Modern Distributed Digital Preservation System**
Nathan Tallman, Hannah Wang



- 345 "A Tartan Rather Than a Plain Cloth": Building a Shared Workflow to Preserve the Regional Ethnology of Scotland Project Archive**
Sara Day Thomson
- 349 Vanished: Preserving the Carmichael Watson Project Website Offline Using Webrecorder**
Anisa Hawes, Sara Day Thomson
- 353 Macintosh Resource Forks - Choosing File Formats for Preservation**
Tyler Thorsted
- 356 A Decade of Trustworthy Digital Repository Certification: Yet There Was One**
Jessica Tieman, David Walls, Lisa LaPlant
- 359 Act Now, Late or Never: Make Digital Objects (more) archivable early in their life cycle?**
Yvonne Tunnat, Katharina Markus
- 366 These Crawls can Talk. Context Information for Web Collections.**
Susanne van den Eijkel, Daniel Steinmeier
- 371 The CO2 Emissions of Storage and use of Digital Objects and Data. Exploring Climate Actions.**
Robert Gillesse, Arie Groen, Eva Van Den Hurk-Van't Klooster, Tamara van Zwol, Lotte Wijsman
- 374 Improving the archiving and contextualization of electronic messaging in French**
Dr Bénédicte Grailles, Dr Touria El Mekki, Dr Édouard Vasseur
- 378 From Outpost to Community: Strengthening support for the Australasian digital preservation community through regional presence**
Jaye Weatherburn, Alexis Tindall, Michaela Hart
- 383 Preservation Watch: Working Towards A Supra-Organizational Preservation Watch Function Within The Dutch Digital Heritage Network**
Tamara van Zwol, Eva Van Den Hurk - Van't Klooster, Lotte Wijsman
- 388 Open Access Books and Digital Preservation**
Dr Alicia Wise, Dr Mikael Laakso, Dr Ronald Snijder
- 392 Evaluating a Taxonomy for Video Game Development Artifacts: Archival Taxonomies in Highly Innovative Domains**
Dr Marc Schmalz, Kylie Snyder, Corey Cherrington, Lidia Morris, Tara Disher, Dr Jin Ha Lee



Panels

- 397 ARCHIVER: Sustainable Preservation of Scientific Data**
Matthew Addis, Teo Redondo, João Fernandes
- 399 How can bringing together the workflows of publishing and preservation lead to better, longer-term solutions that benefit both?: A panel with COPIM Work Package 7, the Embedding Preservability in New Forms of Scholarship Project (NYU), and Project JASPER**
Dr Miranda Barnes, Karen Hanson, Dr Alicia Wise
- 402 A Labor of Language: Building The Global Preservation Community Through Funded Translation Projects**
Rebecca Fraimow, Lorena Ramírez-López, Juana Suárez, Pamela Vizner
- 404 Right Click to Preserve: Preservation, NFTs, and Distributed Ledgers**
John Bell, Regina Harsanyi, Jon Ippolito
- 406 CoreTrustSeal v3.0 In a Preservation and Community Context**
Dr. Jonthan Crabtree, Hervé L'Hours, Ingrid Dillo
- 409 Digital Storytelling as Preservation: A Screening and Panel Discussion**
Syreeta Gates, Jamie A. Lee, Dr James Lowry
- 410 'Practical' Vs. 'Exemplary' Sustainability: Is There a Right Way to Archive Email?**
Christopher Lee, Christopher Prom, Ruby Martinez
- 413 Lessons Learned During the Implementation of a Digital Preservation Project: Experiences from Europe, USA, and Asia**
Teo Redondo, Nathan Tallman, Jessica Knight, Mark Hobbs, Mui Huay Ho, Driek Heesakkers
- 416 It's All Important of Course, But...**
Paul Stokes, Tamsin Burland
- 418 Computational Access to Digital Material: Exploring topics around engagement, ethics and resources**
Leontien Talboom, Jenny Mitcham, James Baker, Sonia Ranade
- 422 Will DNA form the Fabric of our Digital Preservation Storage? DNA Data Storage: A Panel Discussion**
Paul Wheatley, Daniel Chadash, Sibyl Schaefer, Euan Cochrane
- 424 2021 NDSA Staffing Survey: Digital Preservation Intent vs Reality**
Lauren Work, Elizabeth England, Sharon McMeekin, Shira Peltzman, Juana Suárez



Posters

- 429 Keeping Up With the Data: Reflections on Fixity and Data Visualizations**
Angela Beking
- 431 Supporting Preservation of Veteran Personal Archives: A Workshop on the Use of the Virtual Footlocker Project Curriculum**
Dr Edward Benoit III, Dr Allan Martell
- 433 Exploring Software, Tools and Methods used in Web Archive Research**
Katharina Schmid, Sharon Healy, Helena Byrne
- 436 Towards a Collections Model for Preservation Planning at the British Library**
Michael Day, Maureen Pennock
- 438 Digital Preservation In A Lunchbox: Launching a community of practice**
Émilie Fortin, Mireille Nappert
- 440 AIA/Oliver Witte Collection: A digital preservation workflow**
Elisabeth Genest
- 441 Incorporating Digital Preservation and Access Maturity Models into Wider Assessment Programmes: Archive Service Accreditation and the Levels of Digital Preservation and Born-Digital Access**
Dr Melinda Haunton
- 443 Preserving Collections On Tape At The National Library Of Scotland: From Business Case To Bytes**
Lee Hibberd, Alistair Bell, Alan Russell
- 445 Fostering a Data Infrastructure for the Humanities and Social Sciences: A Case Study in Japan**
Dr Ui Ikeuchi, Shinsuke Ito, Yukio Maeda, Kiyonori Nagasaki, Takeshi Hiromatsu
- 447 Leveraging AI for Video Appraisal: A Case Study at the World Bank Group**
Jeanne Kramer-Smyth, Paloma Beneito Arias
- 449 Progress to Participatory Digital Preservation with Geopark: A Case Study of How Gold Museum in Taiwan Participates in Shui-Chin-Chiu Geopark to Engage with Local Communities**
Yi-Ting Lin
- 451 Strength in Numbers**
Laura Peurt, Rachel MacGregor
- 453 Research Weeks**
Peter May
- 456 Upscaling the MPT**
Peter May, Kevin Davies
- 458 Bringing Transparency and Permeability to Organizational Silos: Improving Workflow and Culture**
Daniel Noonan, Sue Beck



- 460 LIBNOVA Consortium: A successful community project**
Teo Redondo, Miquel Tèrmens, Fernando Aguilar, David Giaretta, Julia Thiele, Ciprian Abaseaca, Mikel Rufian
- 462 Preserving Electronic Theses at the University of St Andrews Libraries and Museums**
Sean Rippington, Janet Aucock
- 465 Preserving Photogrammetry Outputs: A case study at the University of St Andrews Libraries and Museums**
Sean Rippington
- 467 Quality Assurance For Born-Digital Interactive Narratives: The New Media Writing Prize Collection As A Case Study**
Giulia Carla Rossi, Tegan Pyke
- 470 Digital Preservation Capabilities of Universities: Survey in the Light of DPC RAM**
Dr Özhan Sağlık
- 473 Creating Workflows to Scale Out Large Open Access E-book Acquisitions at the Library of Congress**
Kristy Darby, Lauren Seroka, Camille Salas, Andrew Cassidy-Amstutz, Elizabeth Holdzkorn
- 475 Behind the Scenes. 3 Decades of Digital Preservation**
Barbara Sierman
- 477 Concept Model for Development of Preservation Plans**
Asbjørn Skødt
- 479 What does Data Loss Really Cost?**
Paul Stokes, Tamsin Burland, Sarah Middleton
- 481 The CO2 Emissions of Storage and Use of Digital Objects and Data**
Lotte Wijsman, Tamara van Zwol, Robert Gillesse, Arie Groen
- 483 Community Archives and Digital Sustainability**
Audrey Wilson, John Pelan, Sean Rippington
- 485 Lessons from the National Archives of Singapore's Journey Developing a Digital Preservation System for Public Records**
Kevin Wong
- 487 Bit Preservation Using the Open Source BITREPOSITORY.ORG Framework**
Eld Zierau, Mathias Jensen, Rasmus Kristensen



Workshops

489 Preserving Complex Digital Objects Revisited

Patricia Falcao, Caylin Smith, Sara Day Thomson

491 Changing Curriculums for a Changing World? Living in Interesting Times: Digital Preservation Education, Pedagogy and Skills

Prof. Ann Gow, Dr Paul Gooding, Zoe Bartliff, Dr Yunhyong Kim, Dr Kathryn Simpson

493 Welcome to Fedora 6.0: Features, Migration Support & Integrations for Community Use Cases

Arran Griffith, Daniel Bernstein

495 The Climate Crisis and New Paradigms For Digital Access

Rachel MacGregor, Anna McNally, Dr James Baker

497 The Bits In The Bytes: Understanding File Format Identification

Francesca Mackenzie, Andrea Hricikova, Andrey Kotov, Kathryn Phelps

499 Eternalize DBs Workshop: Exchange on sustainability and re-use of database content

Dr Kai Naumann, Kevin McMahon

501 The Value of Catastrophic Data Loss

Paul Stokes, Tamsin Burland, Sarah Middleton

503 Registering our preservation intentions: A collaborative workshop on digital preservation registries

Paul Wheatley, Ross Spencer, Euan Cochrane, Kate Murray, Andrew Jackson, Francesca Mackenzie



Tutorials

- 505 Understanding and Implementing PREMIS: A Tutorial**
Karin Bredenberg, Eld Zierau, Micky Lindlar
- 507 Cyber Resilience**
Greg Hewitson
- 509 Writing Binary by Hand**
Martin Hoppenheit
- 511 Scalable Curation of Email with Open-Source Tools: Review, Appraisal, and Triage of Mail (RATOM)**
Christopher Lee, Kam Woods
- 513 LABDRIVE Tutorial: A Research Data Management and Digital Preservation Platform**
Teo Redondo, Antonio Guillermo Martinez Largo, Maria Fuertes
- 515 Continuous Improvement Tools for Developing Capacity and Skills**
Sharon McMeekin, Jenny Mitcham, Dr Amy Currie
- 517 Using ePADD for Email Preservation: Implementing the ePADD+ Project Enhancements**
Sally DeBauche, Paul Carlyle, Tricia Patterson
- 519 Automated Topic Modelling in Archives Portal Europe**
Dr Marta Musso, Kerstin Arnold, Konstantinos Stamatis



Introduction

It is a pleasure to present the proceedings of iPres 2022, the eighteenth International Conference on Digital Preservation hosted by the Digital Preservation Coalition (DPC) in Glasgow, Scotland from the 12th-16th September 2022.

2022 is the twentieth anniversary of the Digital Preservation DPC. Its mission has remained constant over those two decades, a recognition that digital preservation is not only a technical challenge but a human one too. As iPres has shown over the years, the maintenance and renewal of technical infrastructures are a familiar topic in this community but they are oriented towards a socio-technical challenge. Long term success demands renewal and support of the social infrastructures too. As the posters around Glasgow almost said at the time of the conference, 'People make digital preservation'.

Also constant through the work of the DPC has been the hospitality and generosity of colleagues and partners around the world. On this twentieth anniversary of its foundation the DPC sought to repay that generosity, renew those friendships, and welcome delegates from all over the world to our home city of Glasgow. The result was a conference much larger and much more diverse than could have been anticipated, with 649 delegates joining in-person and online (Figure 1).

In-person vs Online delegates

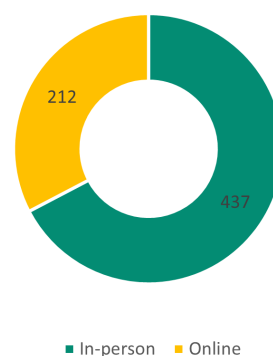


Figure 1: iPres 2022 offered a substantial online offering which was enjoyed by one third of delegates

First time attendees

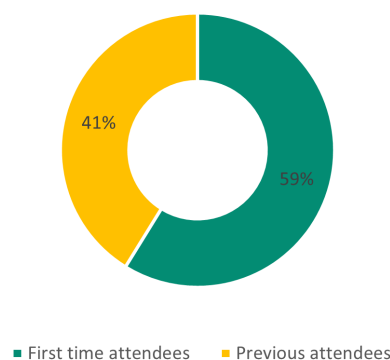


Figure 2: More than half of the iPres 2022 delegates were first-time attendees

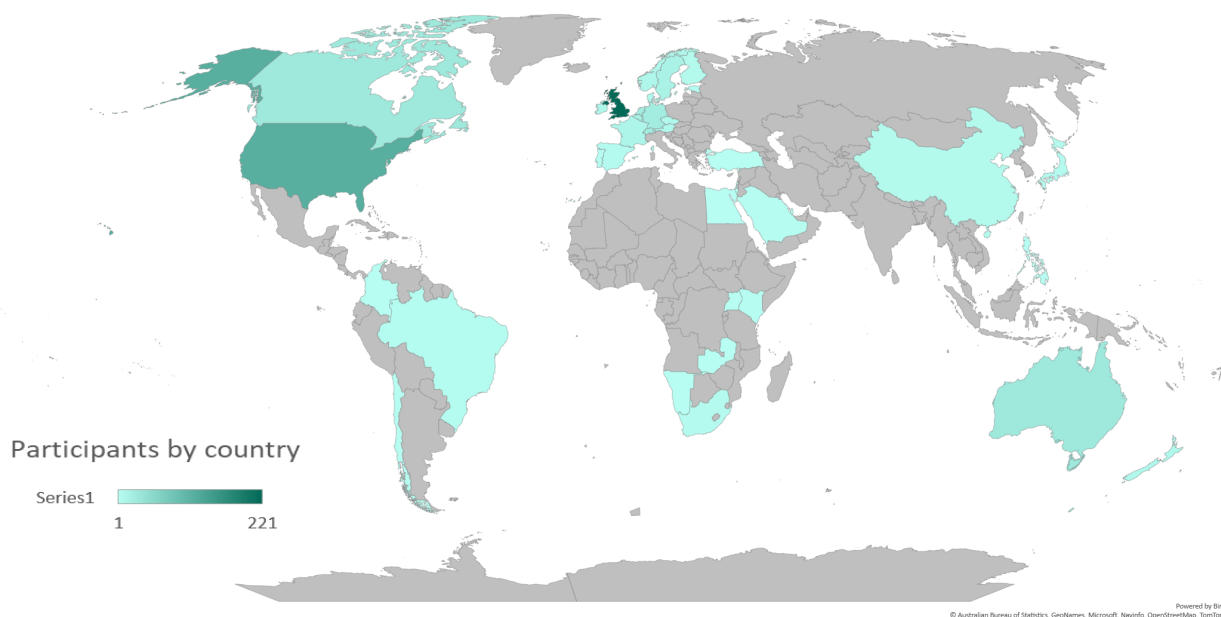


Figure 3: Delegates travelled to iPres 2022 from 38 countries around the world.

Every continent except Antarctica was represented (Figure 3), as well as many professional sectors and

career stages (Figure 4). More than half of the delegates were first time attendees (Figure 2).

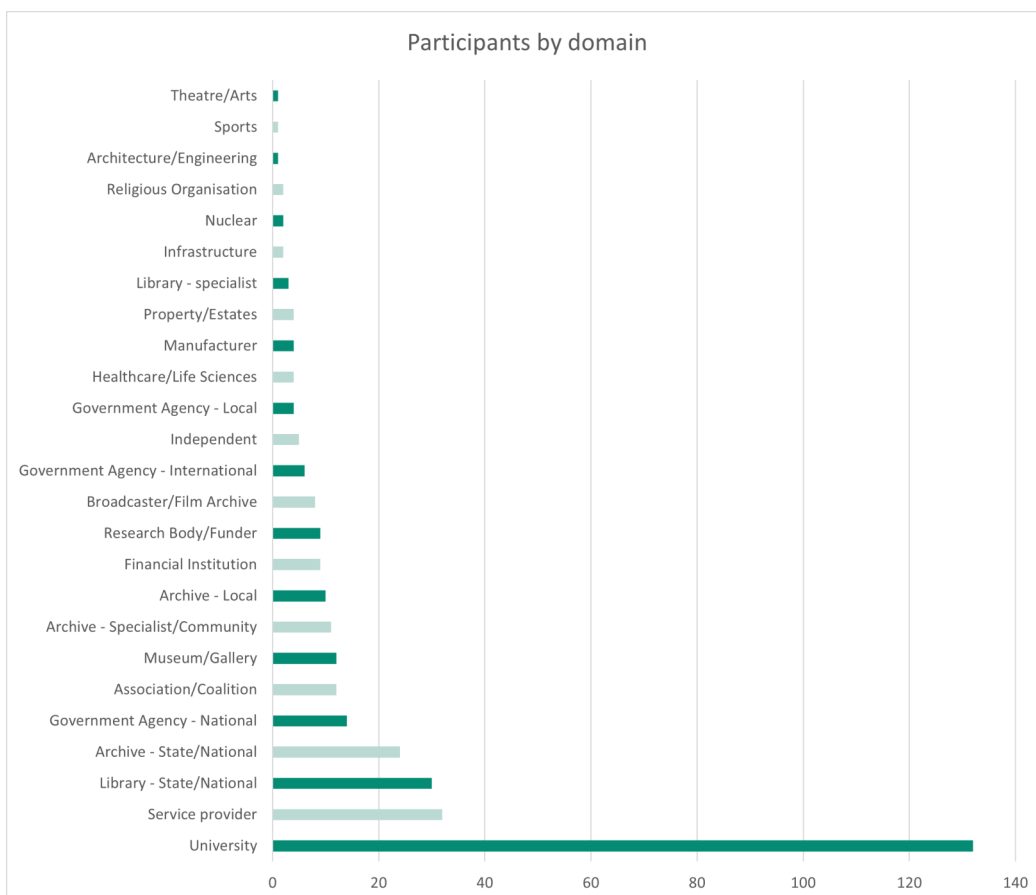


Figure 4: iPres 2022 welcomed digital preservation professionals from a broad range of sectors, demonstrating a diversification of the community.



iPres 2022 saw an unprecedented commitment to accessibility and inclusion. This is most evident in the policy of welcome and inclusion which framed the conference. Reviewers were instructed on the inclusion policy at the outset and reminded of our promise that, anyone generous enough to offer a contribution would be rewarded with supportive comments, even if their contribution was not accepted. Reviewers were not able to make so called 'confidential remarks' and reviewers were informed that all the comments received would be shared with contributors.

The commitment to inclusion was also online participation which has not only allowed us to provide free access to all relevant sessions afterwards; an effort which also allowed the Local Organizing Committee to experiment with subtitling and translation which have not been available at iPres before.

The Local Organizing Committee also oversaw the largest program of grant support and scholarships in the history of iPres: which meant a total of 128 delegates were sponsored or subsidized to attend through the following means:

- DPC supported the participation of 62 delegates at the conference from its Member Fund
- DPC further supported travel, subsistence and registration for 14 more delegates from its Career Development Fund
- Portico sponsored travel, subsistence and registration for 3 delegates from low-middle income countries (none of these were able to complete the immigration process and the sponsorship was returned)
- DPC sponsored travel, subsistence and registration for 3 delegates from low-middle income countries from its member fund (only one delegate completed the immigration process)
- 37 members of the Program Committee members were offered complimentary registration in return for their work
- 10 early career professionals and students were given complimentary places in return for volunteering





As everyone has been, the iPres community was profoundly affected by the Covid 19 pandemic, a fact which helps make sense of the planning, the constraints and the innovations in the program. The organizers of iPres 2021 faced enormous difficulties in bringing their postponed conference to a successful conclusion, holding firm to the idea that digital preservation is fundamentally collaborative. So too the 'Friends of iPres', many of whom were present in Glasgow, accomplished an amazing feat of logistics and planning to deliver the #wemissipres festival in September 2020.

There was a sense that iPres 2022 was always going to be something out of the ordinary because of the moment in global history that it occupied. It convened at a significant moment in local history too, specifically the death of Queen Elizabeth on the eve of the conference. A book of remembrance was established in the foyer of the conference venue, allowing delegates to pay respects collectively and in person.

There is a traditional saying on the death of the monarch: 'The Queen is Dead; Long live the King.' A great deal is packed into these eight words. It is a phrase entirely in the present tense and immediately relevant at iPres 2022. But it also looks to the future and speaks significantly of the past. It more than implies continuity between today, yesterday and tomorrow. 'Today yesterday and tomorrow' could be a subtitle of every single paper in the volume that follows. The digital age, and digital preservation in particular spell this out:

Continuity means a commitment to change; and change means a commitment to learning; and learning means openness to others.

This was the first iPres to have an explicit environmental policy which stated measures taken to limit our carbon footprint; including working with the venue to procure locally sourced food, placing recycling bins around the venue and employing the services of a local printer for all of our branded signage (which used vegetable based inks). The iPres App also allowed us to dispense with the conference bag or booklet or pack: live updates, poster videos, program notes, abstracts, posters and all manner of networking were handled via the App; and conference gifts which could be used for many years to come were favoured over single use giveaways.

Despite all of these measures, travel to the conference was always going to have a significant environmental impact. iPres 2022 completed a benchmarking exercise on carbon consumption which has found that based on transport alone we created a carbon footprint of 346 tonnes (Figure 5).



Mode of transport

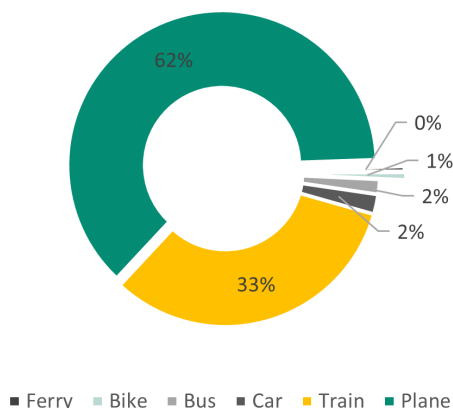


Figure 5: iPres is an international conference which meant that the majority of delegates travelled to Glasgow by plane, significantly contributing to the carbon footprint.

iPres 2022 met under the heading ‘Data for all, for good and for ever’. The call for contributions invited reflection and debate about how digital preservation can support flourishing communities, ecologies, economies and ideas, and it framed these around moments and ideas from the history of the city. It also adapted the motto of our host city, ‘Let Glasgow Flourish’ with the subtitle ‘Let digits flourish.’

The name ‘Glasgow’ means literally the ‘Dear Green Place’: a place to consider the ethical and ecological context of our work. Glasgow is ‘The Workers’ City’: a place therefore to build sustainable communities of practice and professional exchange. Adam Smith wrote ‘The Wealth of Nations’ in Glasgow: a place therefore to model, measure and expand an emerging digital economy, open to all with common purpose for the common good. Glasgow is a ‘city of revolutions’, industrial and otherwise: a place therefore for innovation and radical disruptions. Glasgow is ‘Clyde-Built’, an idiom that means ‘built to last’: a place therefore to engineer for the harshest of conditions, enduring whatever comes.

These five themes are expressed in the main headings of the conference: *Community, Environment, Exchange, Innovation, and Resilience*. Every paper and contribution which follows adapts and interprets these themes, with no shortage of insight, care and creativity. So, while the conference title encouraged delegates to ‘let digits flourish’, this volume demonstrates, through every page and every paper, that iPres is flourishing too.

Keynotes

Amina Shah

Amina Shah is the National Librarian and Chief Executive of the National Library of Scotland. She has more than 25 years’ experience across the library and cultural sector, including both public and academic libraries. Shah has a strong interest in the role libraries, education, literature and culture play in empowering individuals and communities and how organizations can work collaboratively and creatively to maximize their impact and reach.

The mission of the National Library of Scotland is to collect, preserve and make available diverse materials that represent the lives and memories of Scotland’s people. This keynote discusses some of the wonders of those collections and some of the challenges the Library faces in adapting collection and preservation within the context of a rapidly and ever-changing world.



The recording of the presentation ‘Video Killed the Radio Star: preserving a nation’s memory’ has been published on the YouTube Channel of the Digital Preservation Coalition: <https://youtu.be/n00yOiMKFYc>.



Tamar Evangelestia-Dougherty

Tamar Evangelestia-Dougherty is the director of the Smithsonian Libraries and Archives. In addition to her extensive work with rare and distinctive collections, Evangelestia-Dougherty is a published author and public speaker who has presented nationally on topics of inclusivity and equity in bibliography, administration, and primary-source literacy. Her keynote highlighted socio-economic challenges in community archives, calling for more robust digital preservation collaborations to create meaningful pathways toward a holistic digital ecosystem.



Drawing on her own experience as community archives advocate and case studies in North America, Tamar Evangelestia-Dougherty explored socially-engaged techniques to facilitate collaboration and effectively center digital equity and inclusion structures in your engagement efforts to implement multi-stakeholder digital preservation strategies with community archives.

The recording of the presentation ‘Digital Ties That Bind: Effectively Engaging With Communities For Equitable Digital Preservation Ecosystems’ has been published on the YouTube Channel of the Digital Preservation Coalition: <https://youtu.be/IDEWqey559M>

Steven Gonzalez Monserrate

Steven Gonzalez Monserrate is a PhD Candidate in the History, Anthropology, Science, Technology & Society (HASTS) program at the Massachusetts Institute of Technology. He is an ethnographer of data centers and his dissertation surveys the diverse ecological impacts of computing and digital data storage in New England, Arizona, Puerto Rico, and Iceland.

Gonzalez Monserrate is also a speculative fiction writer and filmmaker. This keynote surveys a range of data centers of the future, thinking with artists, futurists, speculative fiction writers and engineers to sketch what sustainable data storage might look like at the end of the decade and beyond. Topics include proposed underwater or extra-terrestrial data centers, 5d memory crystals, data gardens powered by synthetic DNA storage capabilities, and emerging quantum computing technologies.

The recording of the presentation ‘After the Cloud: Rethinking Data Ecologies through Anthropology & Speculative Fiction’ has been published on the YouTube Channel of the Digital Preservation Coalition: <https://youtu.be/pFCqgmLgqzg>.





Peer Reviewed Program

The conference program included sessions of paper presentations, panels, posters and bake-off demonstrations, preceded by workshops and tutorials.

The conference program consisted of up to four concurrent strands each day. One of the strands was hosted entirely on-line and screened in the conference venue. Two strands were webcast from the venue. The fourth was recorded for playback later. The recordings were available to delegates on the platform and will ultimately be available under mostly open access thereafter.

Monday involved Tutorials and Workshops as well as the Digital Preservation Awards. Tuesday and Wednesday

opened with a keynote speaker followed by concurrent strands in sessions of 90 minutes. On Thursday the order was reversed with the keynote at the end of the day. Posters were displayed in two batches on Wednesday and Thursday; and ad hoc activities including the Games Room and the Great Digital Preservation Bake Off carried on throughout.

Following a peer review process iPres 2022 was able to accept a total of 128 submissions, a breakdown of which is shown in the graph at Figure 5 below.

Collaboration is an important theme for the DPC and this continues to be reflected in the conference submissions. The 128 accepted peer reviewed submissions were the work of 331 authors, and the majority of accepted submissions have multiple authors: in some cases up to 12 (Figure 6).

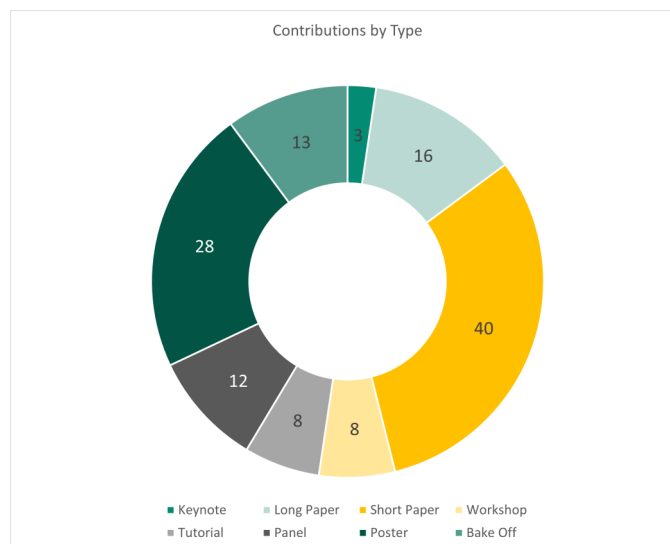


Figure 5: iPres 2022 saw a diverse range of contribution types, with a preference for short papers and posters



Figure 6: The majority of authors chose to collaborate with others on their iPres 2022 submissions



Ad Hoc Program

In addition to the peer reviewed program, iPres 2022 also had a non-peer reviewed program which saw the return of the popular *Digital Preservation Bake Off* on Wednesday, Games and Lightning Talks on the Tuesday and Wednesday, and Professional Visits on the Friday.

The Digital Preservation Bake-Off

Following its debut at iPres 2019 in Amsterdam, the Digital Preservation Bake Off returned to Glasgow this year. The Digital Preservation Bake Off Challenge is an open, light-hearted competition in which solution providers, developers and coders can demonstrate their products and tools while allowing participants to observe the process and verify the claims they make.

Twelve solution providers or ‘bakers’ demonstrated preservation tools and implementations of tools and services in front of a critical audience in a fine setting surrounded by baked goods. Vendors demonstrated their solutions based on a test data-set the conference organizers created for them. By providing a test data-set, demonstrations became more comparable.

Lightning Talks

The Lightning Talks made a welcome return in 2022 with 23 short presentations of three to five minutes each.

Games

Seven digital preservation games were presented by their inventors and played by the delegates, both in person and online.

Career Development

In a new addition to the Ad Hoc program, and in response to calls from the iPres community, the conference offered informal Career Development mentoring. Seven new professionals were paired with experienced members of the community to meet throughout the week and chat together, discuss questions, and share experiences.

Professional Visits

Following the main conference attendees were invited to attend one of 14 professional visits to institutions across Scotland. This Program allowed all iPres 2022 attendees to visit and build relationships with professionals in a range of digital preservation facilities in Scotland. The visits were fully subscribed and attendees reported benefits of practical insights and new partnerships.



Figure 7: Five delegates visited the University of Andrews (Photo courtesy of @SeanRippington)

Social Program



The Social Program saw gatherings and celebrations long overdue. On Monday evening there was a reception and presentation of the Digital Preservation Awards 2022 celebrating those who have served and supported the community with their work over twenty years and more. The Local Organizing Committee also arranged seven social dinners around Glasgow on Monday evening, which proved to be a great and informal way for delegates to meet before the conference got under way. Tuesday saw a more formal civic reception welcoming delegates to the city and marking twenty years of the DPC. On Wednesday all delegates were invited to the conference dinner and ceilidh at the Grand Central Hotel Glasgow. For first time attendees, the Program Committee organized several social events such as pre-conference meet-ups and a virtual coffee corner, and conversation starters like bingo cards and buttons with '1st time attendee'.



Online Sessions

iPres 2022 offered a substantial online program, with delegates participating remotely and in person. Involvement in previous hybrid conferences encouraged the organizers to experiment with a number of features to enhance the experience of online delegates. These innovations were proposed and tested as part of the iPres community consultation in November 2021. Online delegates are typically disadvantaged in two ways: they are less able to network and join the informal discussions on the fringes of the conference venue; and time differences mean they are forced to attend at anti-social times. So, even if a session is webcast they may nonetheless be entirely impractical. To counter this iPres 2022 introduced the following innovations to the online program:

Red Carpet Sessions: brief interviews immediately before and after the keynotes with delegates arriving for the conference, sharing their views on the program and themes.

Sunrise Sessions: a set of 5 90-minute conference sessions at 0730 (local time) each day of the conference for delegates in time zones east (Australasia and Asia) allowing delegates to select and playback the best sessions from the day before and discuss them live as a group, supplemented with commentary from delegates at the venue and a small number of live presentations to be played back at the venue later in the day.

Late Show: a set of 4 90-minute conference sessions at 0800-1930 (local time) for the first 4 days of the conference for delegates in the Americas, playing core content from the previous day, especially keynotes supplemented with commentary from delegates at the venue.

Virtual Coffee Breaks: each day of the conference included an informal 30-minute 'virtual coffee break' between sessions where delegates were invited to chat and introduce themselves, and discuss themes of the conference and share their own insights and discussion points.

Virtual Visits: in addition to in person visits we had hoped to offer 'virtual visits' for delegates online. In the end there was only one offered, and that had to be withdrawn for operational reasons. So despite our efforts, online delegates were not able to participate in the professional visits.

Radio iPres: DPC has an active global community especially in Australasia. The DPC office in Melbourne therefore hosted 'Radio iPres' four days of the conference during the middle of the day (Melbourne time), intended as an informal conversation about digital preservation themes. This was freely available using a different web-conference platform. Allowing us to amplify the messages and themes of the conference to the widest possible audience.



iPres 2022 Prizes

Following iPres conference tradition, iPres 2022 took the opportunity to recognize outstanding contributions and to celebrate these in a set of conference prizes.

This year there were four prizes awarded for: Best Paper, Best Poster, best Contribution by a Newcomer and the Angela Dappert Memorial Award. iPres 2022 recognizes the following outstanding contributions:

Best Paper of iPres 2022 sponsored by nestor

The Best Paper Prize goes to 'Green Goes with Anything: Decreasing Environmental Impact of Digital Libraries at Virginia Tech' by Alex Kinnaman and Alan Munshower of Virginia Tech.

Best Poster of iPres 2022 sponsored by the Digital Repository of Ireland

The Best Poster Prize goes to: 'The CO2 Emissions of Storage and Use of Digital Objects and Data' by Tamara

van Zwol, Lotte Wijsman, Robert Gillesse and Arie Groen of the Dutch Digital Heritage Network.

Best First Time Contribution to iPres 2022 sponsored by the Digital Preservation Coalition

The Best First Time Contribution Award goes to Elisa Rodenburg of the Vrije Universiteit Amsterdam for the game "The Data Horror and Open Science Escape Rooms".

Best combination of research and practice in digital preservation at iPres 2022 sponsored by Adam Farquhar

The Angela Dappert Memorial Prize goes to Andrew Jackson of the British Library for his work on 'Design Patterns in Digital Preservation – Understanding Information Flows.'

IT TAKES A WHOLE VILLAGE TO DEFINE A PRESERVATION STRATEGY

Formalizing Policies on Data Formats Normalization at the National Library of France

Alix Bruys

*Direction of Collections
Bibliothèque nationale de France
Paris, France
alix.bruys@bnf.fr*

Bertrand Caron

*Department of Metadata
Bibliothèque nationale de France
Paris, France
bertrand.caron@bnf.fr*

Thomas Ledoux

*Department of Information Technology
Bibliothèque nationale de France
Paris, France
thomas.ledoux@bnf.fr*

Jordan de La Houssaye

*Department of Information Technology
Bibliothèque nationale de France
Paris, France
jordan.de-la-houssaye@bnf.fr*

Abstract – After publishing its policy on data formats for digital preservation, the National library of France (BnF) had to formalize its method to deal with collected data that did not meet its requirements. This paper describes several significant examples that led BnF from preconceptions to pragmatic decisions upon normalization and preservation strategies for content that could not be ingested as is. Collective intelligence was highly required; this paper is also intended as an attempt to identify which conditions made it possible to emerge between experts, collection managers and process managers.

Described cases tackle issues with PDFs with protection, 48 bits images, PSD files, PDF transformation to JPEG and Final Cut Pro projects. These cases helped define empirically a method, still a work in progress, briefly presented in the last part of the paper.

Keywords – normalization, data formats, preservation strategy, collaboration.

Conference Topics – collaboration; exchange.

I. PREVIOUSLY, ON THE BNF FORMATS WORKING GROUP...

Enters the whole working group, guards standing at the door

The National library of France (BnF) started collecting born-digital content at scale six years ago: donated and acquired texts and still images since 2016¹, ebooks and sound obtained by legal deposit since 2019. Since then, it strives to take the full measure of the differences between digitized and born-digital documents in terms of Quality Assurance (QA), preservation and dissemination.

This is why, since 2018, BnF has reactivated its activity of studying data and metadata formats for the preservation of digital information. As described in an OPF blog post [1], the dedicated working group, named “Groupe Formats de données et de métadonnées pour la préservation numérique (quickly abbreviated “Groupe Formats”, in English “Formats Working Group”) faced in 2017 a need for continued monitoring of data formats in the context of increasing flows of born-digital content.

The working group is composed of around thirty members working in specialized departments

¹ Dates correspond to the ingestion of the first Information Package in the digital preservation repository.

(Engravings and Photography, Performing Arts, Audiovisual, Maps, Music) and in support departments (Information Technology, Preservation, Metadata, Cooperation, Images and Digital Services, Institutional Archives). To gather this team, knowledge of specific content types was requested from different BnF organizational units, expertise was identified in some individuals, and participation from collection departments was demanded.

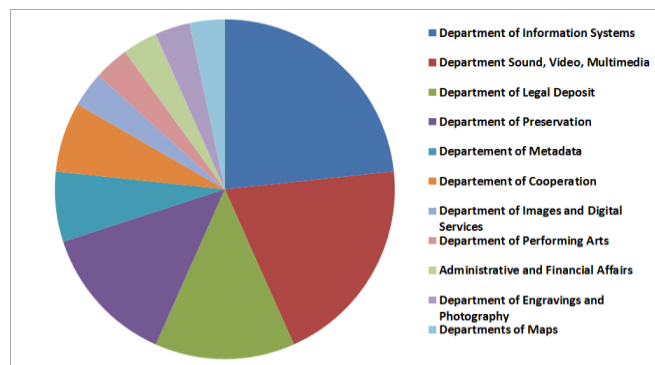


Figure 1 Composition of the Working Group

From 2020 on, the working group's mission has been to publish a revised, justified and accepted formats policy. It was officially released in October 2021, after a five-month review period [2].

This reference document determines and justifies the choices made by BnF in terms of data formats: which data formats it accepts, which properties it extracts from the data, by which tools and how it compares such properties to its requirements. It also started scratching the surface of a difficult question: how does BnF act when it gets data in a format (or with properties) that does not correspond to its standards? This paper reports BnF's efforts to structure its practices and policies one step further.

In this situation, preliminary negotiation with the Producer is preferred, but whenever this is not possible, one among four options must be chosen:

1. Simply refusing the accession of the content;
2. Requesting a new transfer from the Producer;
3. Accepting the content as is and changing the QA, preservation and dissemination environment to take the new data format into account;
4. Transforming the content in order for it to comply with BnF's requirements.

Each of the next five sections will present a real world use case, how it enriched BnF's policy, and/or how this policy in turn informed the Formats working group in order to address the problems at hand.

Because these use cases were far and wide across the range of BnF's activities, the paper intends to show that the diversity of the working group was not only useful, but necessary.

The last section of this paper describes the methodology that emerged in this process.

Note: each section is mischievously introducing actors of the preservation operations in the scenery, identifying them by their first name.

II. REFUSING, IN THE NAME OF THE FORMATS POLICY

Featuring Olivier (collection manager), Alix (process manager), Thomas, Jordan & Bertrand (preservation experts).

In this first use case, Olivier (a collection manager from the Maps and Plans Department) wanted to acquire a simple cartographic document, in the form of 11 PDF files constituting the different parts of an atlas. In the end, he had to give up the acquisition of this resource, despite its value for the BnF collections.

These files were acquired in May 2021, in a context where we couldn't negotiate neither the format nor the rights associated with this set of files. This will rapidly prove important to consider.

At BnF, when documents enter our collections, we try to confront as soon as possible the properties of the files received with the BnF standards. This comparison is first handled by a visual assessment of the documents which exposed no problem. Then an internally developed tool called "Frontin", which retrieves characterization metadata (extracted by Apache Tika and JHOVE, as far as PDF is concerned) and issues an alert in case of properties different from those expected. In this case, Frontin first called Tika, at the time in its 1.12 version (slightly out of date at this time), and reported the following error:

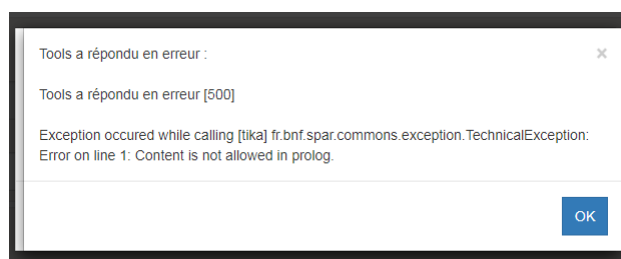


Figure 2 Error reported by Tika

Alix, the Digital Donations and Acquisitions process manager then sought to refine the advice rendered by Frontin, parsing the files with JHOVE (version 1.12.1), which brought up a "Compression

method is invalid or unknown to JHOVE" error. The working group was asked to refine the diagnosis which led Jordan, a preservation expert, to speculate about the presence of TPMs (Technological Protection Measures) as the cause of these errors. It was indeed the case. However, BnF's policy is to accept documents in PDF format as long as they do not contain TPMs (see [3] or [4]). Indeed, TPMs add a layer of complexity to the content that is not tractable in the near future. They jeopardize the accessibility of the content and impede the use of the migration strategy in order to preserve the content.

Subsequently, Thomas (one of the preservation experts) recommended that BnF consider not transforming the file. Indeed, the hypothesis of removing TPMs did not seem clearly authorized, at least as opposed to the case of legal deposit where the absence of TPMs is legally required. The other possibility, which would have consisted in "printing" the file in an image format, would have resulted in the loss of significant properties by going from vector information to a raster image. The re-delivery of the file without TPM by its producer was therefore to be preferred, but this proved impossible due to a lack of respondents. The final decision was therefore to abandon the processing of the document, by eliminating all other possible options in case of data that did not comply with the BnF's format policy.

Note that, as both analysis tools failed to return an explicit error message, the BnF digital preservation "village" joined on this occasion its efforts to the international community, as Bertrand submitted an issue to Apache Tika developers² and supported a similar issue in JHOVE³. In the case of Tika, TPMs were already better recognized by a new version we had not yet implemented. It turned out however that our issue allowed Tika's developers to correct a bug concerning Open Document formats.

This use case reveals several fundamental aspects of the implementation of a format policy, common to all digital entries. It confirms the importance of making a reliable and understandable diagnosis when files arrive. This diagnosis is facilitated by the use of up-to-date and explicit analysis tools that combine identification, characterization and validation tools and synthesize

it in a simplified form. But the diagnosis is only complete after an analysis by a human. It is primarily the responsibility of the process manager, whose role is to ensure that the data entering the process is suitable, natively or after normalization, for access by BnF readers in a permanent manner.

In more complex cases, the process manager solicits and connects different expertises. This use case also makes it possible to evoke the involvement of preservation experts at a very early stage: solicited by the process manager, they refine the diagnosis, evaluate the feasibility of data normalization and accompany the collection manager in the decision regarding the fate of the files received.

III. CHANGING THE ENVIRONMENT INSTEAD OF THE CONTENT?

Featuring Rime (collection manager), Thomas (digital production coordinator), Yannick (product owner), Anne (image signal specialist),

A new challenge was faced when Rime, the collection manager, wanted to process a set of photographs from Brigitte Pougeoise's collection, acquired in 2014. This collection was a mix of ordinary JPEGs as well as TIFFs coded in 48 bits (3 color channels coded with a depth of 16 bits). Our current policy, based on what we can process and what we can give access to, is limited to the more common 8-bit depth. Two approaches were considered: either we expanded our policy or we transformed the content.

Knowing that contemporary practices lean toward better resolution in capturing images, Thomas, the coordinator, explained that the acceptance of these files entailed a revision of our policy, even though this would mean an important evolution of the whole digital environment. Not only the parameters of the assessment tools should be adapted, but more profoundly the whole chain of ingestion and access should be modified in order to take full account of the accuracy of the image (for us, this is a change as important as going from TIFF to JPEG2000 to process images). Yannick, the product owner, was the one who could measure when such

² "Return a more informative error when trying to parse encrypted ODT", issue 3331 on Apache Tika, available at <<https://issues.apache.org/jira/browse/TIKA-3331>> (accessed on March 4th).

³ "Report a more informative error message for encrypted PDFs", issue 640 on JHOVE, available at <<https://github.com/openpreserve/jhove/issues/640>> (accessed on March 4th).

modifications could be carried out and what would be the consequences of this choice.

However, after convening the working group, Anne, the image signal specialist, was able to detect that the use of 16-bit depth was merely an artifact of the post-processing of the images by the photographer. This fact was correlated with the camera model (as described in the images metadata) which would not have been able to encode images in 16-bit depth, as well as the analysis of the color histogram which shows that not all of the space available for encoding had been used. A careful transformation to a more typical 8-bit depth image was then deemed possible by following the usual decision workflow for designing such a transformation⁴, as described in Section V. In taking the decision, the group was helped by the notion of “preservation intent”⁵. The 16-bit depth was not used to capture a richer image, nor was it intended to express a richer image. It was only used in the post-processing of the image, never to be shown. Therefore, should we try to preserve this particular property of the image, our preservation intent would not align to the artist’s intent.

Even though it is not this case that will make us change our processing environment, we are fully aware of the rapid evolution of digital practices, thanks to experts of this domain such as Anne. It is not up to us to avoid it but to be able to take it into account at the right moment and to invest in new formats when they become mainstream. This means that the experts should remain fully vigilant and connected to their communities. The technology watch activity, as described “Preservation Planning” entity of the OAIS [6], is a permanent activity which must enable us to regularly update our policy and leave us sufficient time to make the necessary changes to our processing environment.

Indeed, there is a balance to be found between restricting ourselves to the available formats we already know about and accepting all the particularities that can be thrown at us. It’s not just about having a trustworthy environment that does not distort reality; it’s also about sustainability where we couldn’t cope with the countless forms of creation.

Again, such a decision is only made possible through teamwork where various expertise can be brought together to evaluate the cost of the developments, the content itself and the preservation intention of the creator and of the institution.

IV. TRANSFORMING CONTENT: MULTIPLE CHOICES FOR COMPLEX CONTEXTS⁶

Featuring Sandrine (collection manager), Chloé (collection manager), Rime (collection manager) Bertrand (preservation expert), Anne (image signal specialist).

Three different collections, received as donations by BnF, had in common the fact of having a strong component of digital images intended for consultation in Gallica⁷, the BnF digital library, mostly in formats mastered by BnF (PDF, TIFF and JPEG). These three collections also included some files in PSD format, which is the proprietary format created and used by Adobe for its Photoshop suite. Because this format is proprietary and undocumented, our first intent was to consider a migration for these files. However, because these collections differ in the nature of their content, these PSD files had to be treated differently. Here are the main characteristics of these collections:

- The Philippe Apeloig collection documents the creation of posters by the graphic designer Philippe Apeloig for the book festival in Aix-en-Provence between 1997 and 2015. The collection is hybrid (printed and digital materials) and contains about 300 digital sketches and about fifteen source files of the final printed poster; 3 PSD files are represented among the digital sketches.
- The Amos Gitai collection gathers archives of the film *Rabin, the Last Day* including nearly 2000 photographs of the shooting; 3 PSD files are present among them.
- The Michèle Laurent collection is composed of a hundred photographs of the actor Philippe Caubère’s performances, including some digitizations of book covers; 7 PSD files are present among the scanned images.

After eliminating the other options (request a redelivery, exclude the contents), a study was

⁴ It should be noted that the actual procedure is not yet decided at the time of writing this article.

⁵ This notion is being developed by the digital preservation community for several years. See in particular [5].

⁶ See the BnF blog post [7].

⁷ Available via <https://gallica.bnf.fr>.

initiated to define the preservation strategy for these PSD files, gathering Sandrine, Chloé and Rime, the collection managers concerned, Bertrand, the preservation expert and Anne, the image specialist. To begin, Anne shared her knowledge of the PSD format and the expected uses of the software that produces it. She also revealed the use that had been made of it in the three use cases, according to the properties of the different files after opening them with Photoshop. Subsequently, the working group used the “in-house” method, described in the policy document⁸, consisting in analyzing each use case according to a grid of criteria, structured by three questions:

- Is it necessary to transform the received data?
- If so, in which format?
- Should the source files be retained?

The choice of transforming the data, instead of accepting them as they are, was quickly made for three reasons. First, BnF did not wish to invest in the preservation of a proprietary format. Second, we didn't have the evidence of an intentional technical choice from the data producers. Third, it was necessary to integrate these PSD files into image batches with other formats.

Once this decision was made, the choice of a target format required further investigation, using three criteria relevant to these use cases, taken from the grid defined in the policy document. These three criteria were as follows:

- Format category: identification of a preferred format for the type of content concerned, if applicable.
- Consistency within the information package or the collection: identification of the formats present in the information package or the collection, to be preferred in case of multiple preferred formats.
- Preservation of significant properties or functionalities: definition of a preservation intention, i.e., the set of informational properties and usage modalities of a digital object to be preserved over the long term for a community of users.

In the case of the Apeloig collection, Sandrine, the collection manager, wanted to offer Gallica users the possibility of consulting the information content

of the sketch as part of a batch presenting the successive explorations of the graphic designer. To meet this intention (to show “flattened” image content in Gallica), JPEG was chosen as the target format, even though Anne, the image signal specialist, recommended TIFF as the best option for capturing the maximum amount of information contained in the PSD. In the case of the Gitai collection, JPEG was also chosen, but for slightly different reasons: on the one hand, because Rime wished to privilege access to the visual content like Sandrine, but on the other hand, because Thomas and Bertrand had noted the presence of JPEG files with identical naming, suggesting that JPEGs were the source of PSDs. The proximity of nearly 2000 other photographs in JPEG format also weighed in the decision. For the Laurent collection, there was no doubt that the images were the result of a scanning process. Bertrand therefore advocated the formats retained in the BnF format policy, namely uncompressed TIFF or JPEG 2000. TIFF was finally chosen, because of the exclusive presence of this format in the rest of the collection.

The study also included whether or not to keep the PSD files after they were transformed into the target format. For the Apeloig collection, Chloe, collection manager, wanted to keep all traces of the designer's creative process, including layers and editing history. For the Gitai collection, on the other hand, the files contain layers but are not activated, which makes the PSD format less relevant to these files. For the Laurent collection, the files contained no trace of modifications, which made the PSD even less relevant. Nevertheless, the source files were kept, because they belonged to research-level collections, but also for more pragmatic reasons of prudence and low cost (due to the small number of files involved).

Through three similar and simultaneous use cases, we have experimented with the fact that the choice of a target format is not the result of a miracle recipe. In particular we learned that one cannot simply choose a destination format for a migration based on the source format.

In the field of still images, the BnF's format policy had retained preferred formats for images resulting from digitization or for edited digital photographs,

⁸ See [2], p. 19.

but had not yet pronounced itself, for lack of cases, on images in their production stage.

In the end, these cases did not lead us to change our policy: the presence of PSD files in these collections was too anecdotal, and sometimes not even significant. These cases have taught us how to manage the exception in the search for homogeneity of information packages.

V. TRANSFORMING CONTENT: WHICH METHOD & TOOLS TO USE?

Featuring Sandrine & Bérenger (collection managers), Alix (process manager), Thomas & Bertrand (preservation expert), Anne & Patrick (image signal specialists).

Another case arose with the aforementioned Apeloig collection. The digital assets were of two different kinds:

- Final version of the poster, ready to be printed, in PDF;
- For each poster, several sketches successively made. These files were in different formats: PDF, TIFF and JFIF.

The sketches of the same poster, gathered in the same Information Package, had to be normalized; indeed, BnF policy requires that files with the same use in the same Information Package be in the same format. Sandrine's (the Apeloig collection manager) intention was that only the final version would be reprinted for an exhibition. She considered that the interest of the sketches was limited to documenting the creative process. The informational preservation approach⁹ allowed for a transformation to image format, while retaining the original PDF files.

Thomas noted that the PDFs of the sketches contained some superimposed elements (text, sometimes transparent graphic elements). In order to transform the PDF sketches into JFIF images it was therefore necessary to opt for a rasterization solution instead of a simple image extraction.

A short list of object properties has been determined by Bertrand¹⁰ in order to judge the result of a transformation:

- Definition (width and height of the image in pixels);
- Weight (in bytes);
- Resolution (number of pixels per size unit)
- Dimensions (size in centimeters / inches, depending on the definition and resolution);
- Visual quality (estimated visually by the image signal specialist).

These criteria were completed by some others regarding the software tools¹¹:

- Availability of the tool (free or not, deployment on BnF standard workstations, price);
- Implementation mode (CLI / GUI);
- Possible automation of the tool.

The correct treatment of certain components of the object was also considered:

- Color profile management;
- Presence of internal metadata.

To determine which method and tool would be most effective, the group compared the proposals of several of its members. These proposals came from Thomas and Bertrand, preservation experts, from Anne, an image signal specialist, but also from Bérenger, an audiovisual collection manager. The following tools were evaluated:

- PDFCreator¹², a tool deployed on all BnF workstations and with a GUI;
- XnView¹³, also available on all BnF workstations, with or without the help of Adobe Reader;
- pdftoppm¹⁴, a tool found thanks to Johann van der Knijff's excellent list of PDF processing tools [11], which by coincidence was published at the time of the study;
- PDFBox¹⁵, already used in BnF processes to generate thumbnails from PDFs for digital books;
- Photoshop¹⁶, a tool favored by image signal specialists.

⁹ For a definition of informational vs. artifactual preservation approaches, see [8], p. 15 sqq..

¹⁰ This list is a subset of the properties proposed by [9].

¹¹ The distinction between criteria for evaluating transformation consequences and criteria for transformation process is inspired by [10].

¹² Adobe PDFCreator, PDF converter, <https://www.pdfforge.org/pdfcreator>.

¹³ XnView, free software to view, edit and resize images, <https://www.xnview.com/>.

¹⁴ Poppler pdftoppm, PDF converter to image files, <https://www.mankier.com/1/pdftoppm>.

¹⁵ Apache PDFBox, open-source library for handling PDFs, <https://pdfbox.apache.org/>.

¹⁶ Adobe Photoshop, raster graphics editor, <https://www.adobe.com/fr/products/photoshop>.

Resulting files were examined by Anne and Patrick, our image signal specialists.

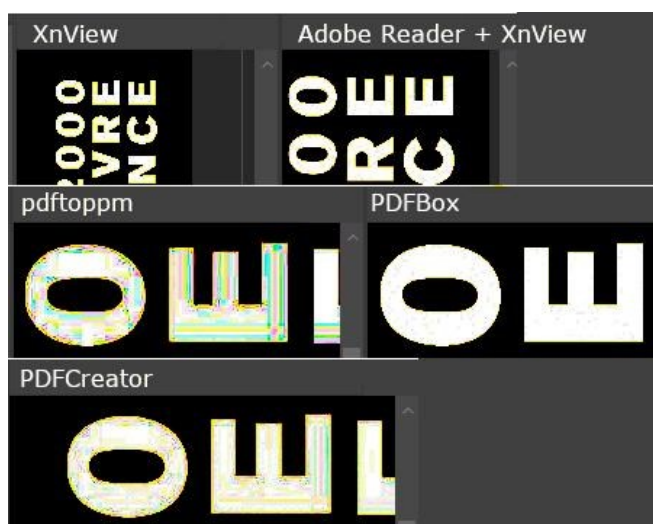


Figure 3 Visual comparison of the transformation to JPEG image.

For each tool, the method was recorded:

- Step-by-step instructions, possibly with screenshots, for GUI-driven tools;
- Command line for CLI-driven tools.

One important result of such a process is a publishable, justifiable and reproducible, though always questionable, policy. In case such a situation occurs, BnF determined that "born-digital" PDF such as those in the Philippe Apeloig collection will be processed by PDFBox, a tool capable of rasterization, while a PDF resulting from a digitization, containing only one image per page, will be processed by an extraction tool such as Apache Tika¹⁷.

PDFBox was then added to our preconditioning tool, Frontin, to handle automated transformations; moreover, shortly after, and following the appearance of a new use case, Thomas studied the automatic distinction between these two types of PDF [12].

This process also showed that comparing results in a working group plenary session had pedagogical virtues. The diversity of the results obtained demonstrates that not all conversions are equal. Moreover, it proves once again that two objects of different nature can be recorded in the same format, and that the strategy adopted will depend on the object nature.

Some organizational issues emerge: the choice of a method cannot omit the "human resources" dimension: depending on whether one chooses a tool with a GUI or only a command line, the personnel capable of implementing the transformation is not the same. This consideration is all the more important as the transformation of born-digital content is time-consuming, for signal specialists as well as for collection managers, who currently tend to consider that these operations are not, or not exclusively, of their responsibility.

VI. WHEN THERE IS NO IDENTIFIED TARGET PRESERVATION FORMAT YET: CREATING DISSEMINATION SURROGATES

Featuring Jean-Yves (audiovisual expert), Rime (collection manager), Bertrand (metadata specialist).

The ultimate challenge arises when we receive material that is not only not currently accepted, but whose formats are either proprietary or require specific hardware. One such recent example comes with the film daily rushes¹⁸ from the FCP (Final Cut Pro) program. This software is one of the classic tool for filmmakers but it is completely tied to the Apple platform and has already undergone one breaking change with version X which chooses an XML-based representation and force the use of a third-party utility to migrate to the new version¹⁹.

In the donation we have daily rushes in FCP7 as well as FCP-X format. Neither of these can be read in an ordinary workstation in the library and the management of such files requires specific competences. Moreover, the edit decision list²⁰ contained in the central file makes direct references to other files (the raw audio or video parts) with absolute paths. The first manipulation that requires the use of the software and a well-equipped hardware is to recreate the links with the new installation. In order to try to figure out how we can manage this material, we first look for an expert (fortunately, there are knowledgeable people in the audiovisual department) and wait for a compatible hardware workstation. Having both of them provides us the ability to better understand the material (delimited all the files involved in a FCP

¹⁷ Apache Tika - a content analysis toolkit, <https://tika.apache.org/>.

¹⁸ Daily rushes are the raw, unedited footage shot during the making of a motion picture (definition taken from Wikipedia).

¹⁹ Refer to <https://support.apple.com/en-us/HT208054> and https://en.wikipedia.org/wiki/Final_Cut_Pro_X

²⁰ An edit decision list contains an ordered sequence of audiovisual material used in a film editing project.

project) and try to figure out the main piece of information.

Even though it was clear from the beginning that we would have to store the information as is and provide a basic bitstream preservation, we also intend to provide in an easy manner to our users some sort of substitute. Indeed, we don't view our preservation system as a dark archive but more like a repository of information that needs to be accessible as far as the legal restrictions permit us. In this case, because of the kind of material, a direct access through our digital library, Gallica or its version accessible only in its precinct, Gallica Intra Muros, is not envisioned but we intend to provide enough information so that the researchers know if the material is of interest to them.

In the case of film daily rushes, we are willing to provide a list of the material involved in the making (images, sound recordings, video footage) as well as the images of the timeline. Those advanced descriptions of the original material will be used as a surrogate for the original material. It allows us to give access to certain information in a simple way and, if necessary, to accept justified requests for communication that would involve the installation of specific equipment and the associated logistics.

For practical reasons dictated by our preservation system, we intend to ingest the original material and their surrogate in two different Information packages, probably at two very different times. From the preservation point of view, this is the first time we intend to ingest both an original and the result of a migration in two different packages. Usually the two representations are archived together and the relationship between what constitutes an original and a master is stated in the package. Moreover the migration itself can be described in the provenance metadata. This allows us to apply a strict policy for the master version (target of the migration) and a less strict one for the original (source). Here, we will need to ingest the FCP project as a master, even though we have no control on its format whatsoever. This implies lowering the bar of entry so much for this case that any kind of data could enter our systems afterwards, which we do not want to happen.

Therefore, once the decision of acceptance has been made, the original material is stored in a specific location and documented so that the intention for migration is clearly stated and the

reason and needs formalized as much as possible. A complete documentation of our level of knowledge is written and the risk associated with a possible loss of control is stated: proprietary format, hardware specificities, legal issues... A PREMIS Event [13] of type `migrationIntended`, informs about it:

```
<premis:event>
...
<premis:eventType>
migrationIntended
</premis:eventType>
...
<premis:eventOutcomeInformation>
  <premis:eventOutcome>
    type=transformationWithBackup,
    sourceUse=master,sourceFormat=fcf
  </premis:eventOutcome>
</premis:eventOutcomeInformation>
<premis:linkingAgentIdentifier>
  <premis:linkingAgentIdentifierType>
documentCode
  </premis:linkingAgentIdentifierType>
  <premis:linkingAgentIdentifierValue>
BnF-ADM-2021-012345
  </premis:linkingAgentIdentifierValue>
...
</premis:linkingAgentIdentifier>
</premis:event>
```

In our preservation system, we will implement a rule of a new kind stating that these files are allowed, but only if a `migrationIntended` event is attached. Therefore the existence of this case in our collections will be exposed.

In parallel, a surrogate is built that provides as much information as possible using only managed formats: it can be screenshots, a video of the representation of the material, part of it. This surrogate is directly linked to the original. Again, if the original is preserved at the bitstream level, the surrogate is meant to be enriched as we gain more information about the original material or find new ways to provide access to it.

```
<premis:event>
...
<premis:eventType>
migrationProcessed
</premis:eventType>
...
<premis:eventOutcomeInformation>
  <premis:eventOutcome>
    type=transformationWithBackup,
    sourceUse=master,
    sourceFormat=fcf,
    targetFormat=jpeg,
    satisfactionLevel=poor
  </premis:eventOutcome>
</premis:eventOutcomeInformation>
<premis:linkingAgentIdentifier>
  <premis:linkingAgentIdentifierType>
documentCode
  </premis:linkingAgentIdentifierType>
  <premis:linkingAgentIdentifierValue>
```



```

BnF-ADM-2021-012345
</premis:linkingAgentIdentifierValue>
<premis:linkingAgentRole>
performer
</premis:linkingAgentRole>
</premis:linkingAgentIdentifier>
<premis:linkingObjectIdentifier>
<premis:linkingObjectIdentifierType>
ark
</premis:linkingObjectIdentifierType>
<premis:linkingObjectIdentifierValue>
ark:/12148/m0n4rk
</premis:linkingObjectIdentifierValue>
<premis:linkingObjectRole>
source
</premis:linkingObjectRole>
</premis:linkingObjectIdentifier>
</premis:event>

```

As you may understand, even for complex or unreachable material, the preservation process starts at the beginning by capitalizing on the information available to us and seeking skills either inside or outside the library. Even if our grasp is weak, we do not intend to bury the material but on the contrary to make it visible by cataloging it, preserving it and providing access to a direct surrogate for it. In this way, we hope to be able to monitor it and possibly find innovative means of access. The mere fact that we record all this material could be an incentive to seek sponsorship or to consider a research program on it.

VII. WHICH REGULAR PROCESS EMERGED FROM THESE EXPERIMENTS?

Featuring Benjamin (functional analyst), Anne-Lise (collection manager)

A. Vocabulary

Benjamin: "In the triage and appraisal application we are currently developing, what should we call the operations that change the bitstream of objects we want to accession, prior to ingestion?"

Working together between people of different backgrounds implies agreeing on a common terminology. Thus the working group had to recommend a term corresponding to a "preservation operation carried out before ingesting into the preservation system and resulting in the modification of the bitstream". The candidates were the terms "migration", "conversion", "transformation" and "normalization".

The term "transformation" was preferred in the dialogue between different BnF entities. It corresponded indeed to a term defined by OAIS and was generic enough to be understood by all. In international writings, the term "normalization" is

also used, according to the generally adopted meaning.

On the other hand, were rejected:

- "Conversion", which was too restrictive because it suggested a change in container format, whereas the operation could affect the signal alone (a change in color model from CMYK to RGB, for example);
- "Migration", which evoked a migration of the system or the supports for the computer specialists;
- A variant of the previous one, "format migration", because it is not the format which is affected but the content.

B. Roles and missions

Although the decisions taken on the occasion of the various cases cited above are always questionable, they are indisputably better than those that the members of a single BnF department could have taken. But what are the profiles and skills of the agents involved in these decisions?

Four main profiles stand out today among the members of the Formats group, from the perspective of analyzing and processing natively digital objects before they enter the preservation system:

- The **collection manager** knows and understands the institution's documentary policy, the context of content creation, and maintains contact with the creator; they selects the content to be acquired by BnF, defines the intention of preservation, makes an informed decision on the acceptability of the content (appraisal) and on its technical and bibliographic treatment with the help of diagnostic tool(s), and justifies and documents these decisions, in agreement with the process manager.
Note: the collection managers were originally seen as relays for the working group's recommendations in their departments; it turned out that no decision could be taken without them!
- The **process manager** is responsible for the overall operation of the circuit, from the controls carried out by the QA services to the dissemination; they leads a community composed of the profiles mentioned above, ensures the coherence of the decisions taken by the collection managers, makes sure that the collections deposited are accessible, formalizes

and expresses the needs of these communities to the preservation experts.

- The **preservation expert** has skills on data and metadata formats (especially internal), on analysis tools, on the functioning of the preservation system; they analyzes the feedback from the diagnosis tools, they makes sure they are updated, they eventually makes them evolve, they defines the controls to be put in place in the preservation system or upstream, and helps documenting the transformation methods.
- The **signal specialist** has skills in editing and transforming signals - in OAIS vocabulary [6], this generally corresponds to "Content Information" -, on the uses and practices of the creators of these contents; they carries out complex transformations, evaluates the methods and results of a transformation, and helps documenting the transformation methods.

A gap in this organization remains: there is no profile that takes care of simple transformations. BnF "digital stacks managers" role is currently limited to the preservation system perimeter; as normalization takes place before ingestion, they are not engaged in this process yet.

Note that the organizational logic presented above is empirical and derived from the use cases described in the article. Eventually, a more thorough analysis of the missions and the skills required to carry them out, as well as the integration of these elements into job descriptions should be carried out. We could then rely on multiple works from the digital preservation community such as the DigCurV initiative [14].

C. Modeling a Regular Normalization Process

These cases forced BnF to reflect on the decision-making processes and the means of documenting them, in order to show how the documentary choices condition the technical decisions.

The normalization process was therefore defined as follow:

- 1) **Diagnose.** The diagnosis stage consists of determining whether the content as received by BnF can be deposited in the form of the file currently in its possession. It consists of comparing the properties of a file using analysis tools (characterization) with those of the preferred and accepted formats by BnF for a

given context and with the rules for constituting the package.

- 2) **Decide.** If the file is not in one of the formats acceptable to a given channel, decide what to do with the contents. It is necessary to make a choice between:
 - Rejection of the file, and therefore of its content (as described in section II);
 - Identification of another form of the digital representation or request for a new delivery after transformation by the Producer;
 - Acceptance of the file as it is (this option implies adapting the ingestion, preservation and access environments (as described in section III);
 - Transformation carried out by BnF to meet its own requirements (as described in sections IV and V).
- 3) **Study.** If the last option was chosen, determine whether an existing preservation strategy applies; if not, define a suitable transformation method: software tool, parameterization, implementation method.
- 4) **Perform.** Implement decisions taken in the previous step.
- 5) **Control.** Verify that the file produced complies with BnF's deposit and preservation requirements, and that the significant properties and functionalities of the content have been preserved during the transformation.
- 6) **Document.** Keep track of the transformation operation and, if a new study was needed, define BnF's policy in the form of a preservation strategy.

D. Documentation

Anne-Lise: "But how do we keep track of these decisions? We chose the other option one year ago... How can we improve consistency?"

Having noticed conflicting decisions for which the reasons were unclear, collection managers emphasized the need to document the transformations. The documentation process is linked to the transformation process described above, in the following way:

- 1) **Diagnosis and decision stages:** upon receipt of a homogeneous set of contents that do not comply with BnF's format policy, a **diagnosis and decision form** is created, documenting the nature of the contents, their production history,

their use by the Producer, the collection manager's preservation intention, the identification of their format, the analysis of the set according to the criteria grid in the policy document,²¹ and the appraisal decision. The form is filled out by the collection manager assisted by the process manager and possibly by preservation experts.

- 2) **Study stage:** If the decision concludes that the content needs to be transformed, the **list of transformations** is consulted to determine if one of them fits the case. In addition to recording the source format, the target format and the tool used, this local, non-automated "preservation action registry"²² emphasizes the justification for using such a transformation, its objectives and the above-mentioned criteria that were decisive in choosing the transformation.
- 3) If in the previous stage no existing transformation is applicable, the **study stage** results are recorded in a **report** listing the criteria for evaluating the transformation process and produced data (as described in part V). The document contains a detailed description of the implementation of each solution, the choice of a method and its justification. The **list of transformations** is also updated to include the new transformation.
- 4) **Documentation stage:** if the implementation method is manual, a **tutorial document** to reproduce it is produced in order to guide step by step the agent who will perform it in the future.
- 5) **Documentation stage:** in the METS manifest accompanying each Information Package, a **comment** describing the transformation operation is added to keep track of it and inform the reader.
- 6) **Documentation stage:** If the transformation appears to be sufficiently mastered and broadly applicable, it is considered a validated policy and will appear in the next version of the **policy document** [2].

EPILOGUE

In the last years, the 'Formats' working group appears to have gained maturity in both technical

and organizational domains. It has become clear that on preservation strategy issues no-one can take a decision alone, the right decision being the one that is both driven by librarians and informed and implemented by technicians.

Discussions happening in this working group made clear that expertise is not about developing a comprehensive knowledge on a specific domain, but rather about gathering insights from agents all around the institution and building a consensus by bringing together different points of view.

As it was recently recalled by William Killbride, "if you're doing digital preservation alone you're not doing it right" [16]!

Exeunt all softly

ACKNOWLEDGMENT

The authors would like to make a warm round of applause for the whole BnF "village". Particular gratitude goes to all members of the 'Formats Working Group' who are daily contributing to make this team a welcoming and inclusive place to learn and improve knowledge on digital preservation.

Authors owe special thanks to Chloé Perrot, Yannick Grandcolas, Rime Touil and Anne Paounov, who made particular contributions to the work described in this paper.

REFERENCES

- [1] B. Caron, "A Love Letter to Formats," *Open Preservation Foundation blogs*, November 2021. Available at <<https://openpreservation.org/blogs/a-love-letter-to-formats/>> (accessed on February 25th 2022).
- [2] *Formats de données pour la préservation à long terme : la politique de la BnF*, Bibliothèque nationale de France, October 2021. Available at <<https://hal-bnf.archives-ouvertes.fr/hal-03374030>> (accessed on February 25th 2022).
- [3] S. Hein & T. Steinke. *DRM and digital preservation: A use case at the German National Library*, in *Proceedings of the 11th International Conference on Digital Preservation, iPRES 2014*, Melbourne, Australia, October 6-10, 2014 Available at <<https://ipres-conference.org/ipres14/sites/default/files/upload/iPres-Proceedings-final.pdf>> (accessed on June 1st 2022).
- [4] S. Derrot, J.-P. Moreux, C. Oury & S. Reecht. *Preservation of ebooks: from digitized to born-digital*, in *Proceedings of the 11th International Conference on Digital Preservation, iPRES 2014*, Melbourne, Australia, October 6-10, 2014 Available at <<https://ipres-conference.org/ipres14/sites/default/files/upload/iPres-Proceedings-final.pdf>> (accessed on June 1st 2022).

²¹ See [2], p. 19.

²² In reference to the PAR international initiative [15], whose ambition is to register preservation actions across different repositories.

conference.org/ipres14/sites/default/files/upload/iPres-Proceedings-final.pdf> (accessed on June 1st 2022).

- [5] C. Web, D. Pearson, & P. Koerben, "Oh, you wanted us to preserve that?!" Statements of Preservation Intent for the National Library of Australia's Digital Collections. *D-Lib Magazine*, 2013, 19(1/2). Available at <<https://doi.org/10.1045/january2013-webb>> (accessed on June 23rd).
- [6] CCSDS, *Reference Model for an Open Archival Information System (OAIS)*, 2012. Available at <<https://public.ccsds.org/Pubs/650x0m2.pdf>> (accessed on February 25th 2022).
- [7] A. Bruys, B. Caron, Y. Grandcolas, T. Ledoux & A. Paounov, "If we want things to stay as they are, things will have to change," *Open Preservation Foundation blogs*, November 2021. Available at <<https://openpreservation.org/blogs/if-we-want-things-to-stay-as-they-are-things-will-have-to-change/>> (accessed on February 25th 2022).
- [8] T. Owens, *The Theory and Craft of Digital Preservation*, Baltimore: John Hopkins University Press, 2018.
- [9] L. Montague, A. Brown, G. Knight, S. Grace, *InSPECT Significant Properties Testing Report: Raster Images*, September 2018. Available at <https://figshare.com/articles/journal_contribution/InSPECT_Significant_Properties_Testing_Report_Raster_Images/7137803/1> (accessed on March 1st 2022).
- [10] F. Luan, M. Nygård, G. Sindre et al., "Using a multi-criteria decision making approach to evaluate format migration solutions", in *Proceedings of the International Conference on Management of Emergent Digital EcoSystems - MEDES '11*, presented at the International Conference, San Francisco, California, ACM Press, 2011. Available at <<http://dl.acm.org/citation.cfm?doid=2077489.2077498>>. (accessed on February 22nd 2022, p. 48).
- [11] J. van der Knijff, "PDF processing and analysis with open-source tools", *Bitsgalore*, September 6th 2021. Available at: <<https://www.bitsgalore.org/2021/09/06/pdf-processing-and-analysis-with-open-source-tools>>. (accessed September 6th 2021).
- [12] T. Ledoux, "Scanned vs native PDFs, how to differentiate them?", *OPF Blogs*, February 11th 2022. Available at: <<https://openpreservation.org/blogs/scanned-vs-native-pdfs-how-to-differentiate-them/?q=1>>. (accessed on March 3rd 2022).
- [13] B. Caron, A. Di Iorio, C. Blair, L. Bountouri, R. Guenther et al.: *PREMIS 3 OWL Ontology: Engaging sets of linked data*, in *Proceedings of the 15th International Conference on Digital Preservation, iPRES 2018*, Boston, MA, USA, September 24-28, 2018. Available at <<https://hdl.handle.net/11353/10.923631>>
- [14] DigCurV Curriculum Framework, a European Union funded project, 2013. Available at <<https://digcurv.gla.ac.uk/>> (accessed on March 7th 2022).
- [15] Preservation Actions Registry. Available at <<https://parcore.org/>> (accessed on March 7th 2022).
- [16] W. Killbride, "Why I iPres", *iPRES 2022 blogs*, [2021]. Available at <<https://ipres2022.scot/blog/>> (accessed on March 8th 2022).

USEABLE SOFTWARE FOREVER

The Emulation as a Service Infrastructure (EaaSI) Program of Work

Euan Cochrane

Yale University Library
United States
euan.cochrane@yale.edu
[0000-0001-9772-9743](tel:0000-0001-9772-9743)

Klaus Rechert

University of Applied
Sciences Kehl
Germany
rechert@hs-kehl.de
[0000-0002-2454-4374](tel:0000-0002-2454-4374)

Jurek Oberhauser

OpenSLX
Germany
jurek@openslx.com

Seth Anderson

Yale University Library
United States
seth.r.anderson@yale.edu

Claire Fox

Yale University Library
United States
claire.fox@yale.edu

Ethan Gates

Yale University Library
United States
ethan.gates@yale.edu
[0000-0002-9473-1394](tel:0000-0002-9473-1394)

Abstract – The Emulation as a Service Infrastructure (EaaSI) program of work is dedicated to ensuring all software is usable forever. In this paper we outline why we believe this vision is important, describe the technologies included in the EaaSI software, describe some of the challenges we face in realizing this vision, and finally outline some of the future developments we are working on to expand the impact of the EaaSI program of work.

Keywords – Emulation, software preservation, migration, file formats

Conference Topics – Community; Exchange; Innovation.

I. INTRODUCTION

“In short, software is eating the world.”

- Marc Andreessen, Wall Street Journal, August 20, 2011 [1]

Software is everywhere: software runs our societies’ businesses, industries, transportation systems, power grids, and healthcare systems. Software is essential to the provision of authentic digital evidence in our legal systems. Scientific research methods and their associated research outputs are increasingly software-dependent, and our economic and cultural heritage, including the records of our long-life economic and defense assets such as buildings, ships, and aircraft, is increasingly born-digital. Notable examples include computer Aided Design (CAD) files, websites, documents, slide

sets, emails, time-based media, spreadsheets, databases, design files, project management files, etc. Moreover, born digital objects frequently require specific older software applications in order to be accessed at all, or accessed without meaningful distortion. Without maintaining access to legacy (no-longer supported) software applications, aging software-dependent industrial control and healthcare systems will not be able to be troubleshooted or redeveloped, long-life economic and defense assets won’t be able to be maintained, digital evidence won’t be able to be authenticated, software-dependent research reproduced, or digital cultural heritage accessed.

With the importance of access to software in mind, the Emulation as a Service Infrastructure (EaaSI) program of work is endeavoring to ensure software is always accessible with a dedication to the vision “Usable software, forever”. In this paper we outline why emulation and software preservation are vital to all digital preservation endeavors and what EaaSI is doing to make emulation practical, scalable, and usable for all, indefinitely.

II. WHY SOFTWARE PRESERVATION AND EMULATION?

All digital files are functionally software. Born digital objects are fundamentally different in their nature compared to previous information capture technologies such as paper, film, or velum. While we

generally think of born digital objects as the files that software applications create, and which we share between systems, they are functionally much more complex, and these complexities have significant implications for our information management and long-term preservation approaches.

Born digital objects are distributed amongst multiple files, most of which are files provided by the software applications used in the creation of the objects and provision of the objects as ‘information experiences’ or ‘performances’. This characterization of born digital objects as information experiences or performances was first proposed in the National Archives of Australia’s green paper published in 2002 [2] in which they described a “performance model” for digital objects:

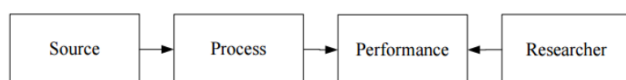


Figure 1 The Performance model

The implications of this model are significant, for example: if a user “opens” a Microsoft word file using a modern version Microsoft Word for Windows, this initiates a complex process in which hundreds of files on the computer provide instructions in response to the results of the processing of the information from the previous files in the process. A change to any one of those files has the potential to change the information presented to the users at the end of the process, change how the user can interact with the “information experience” they are presented with (for example it might make it impossible for the user to check the contents of metadata fields using the application’s User Interface (UI)), or completely terminate the process making the file “unopenable”.

The figures below provide an example in which the same spreadsheet “performance” is created using the combination of a primary data file (the traditional ‘record’) - a .xls file – and the files that represent two different applications. The result of the computer processing the instructions in the two different sets of files (both containing the primary data .xls file) is two different performances with different information being available for interaction in each. In one performance [figure 2], after executing the combination of instructions in the software files and data file a comment is made available in the User Interface (UI) providing important context. In the other performance [figure 3] the combination of

instructions in the set of files leads to no comment being visible or available for interaction.

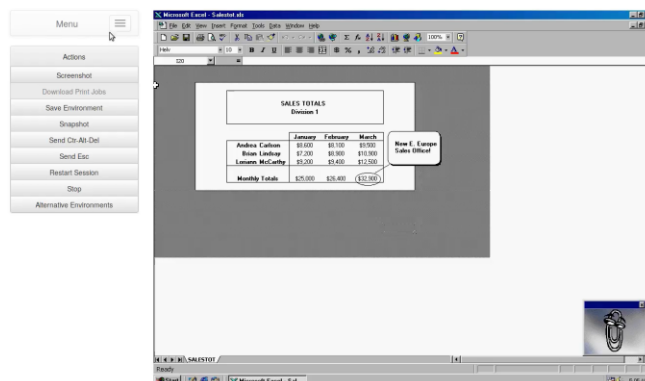


Figure 2 .xls workbook opened in Microsoft Excel for Windows 95

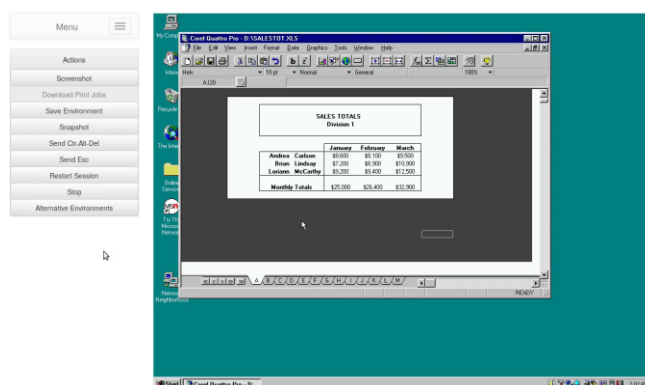


Figure 3 .xls workbook opened in Corel Quattro Pro for Windows 95

In some situations, these outcomes can be even more pernicious. Information that is not visible or even non-existent in the original performance can be added when the software files that made up the original performance are substituted for different files. For example, in the example below we see tags being added to the title of the document that indicate something about the document should be kept private. In the original these don’t exist, but when the primary data file is combined with a different set of software files, the resultant performance includes instructions that interpret parts of the primary data file as requiring the “private” tags to be presented as they are. This is deceptive and would likely not be useful as evidence of the security classification of the document.

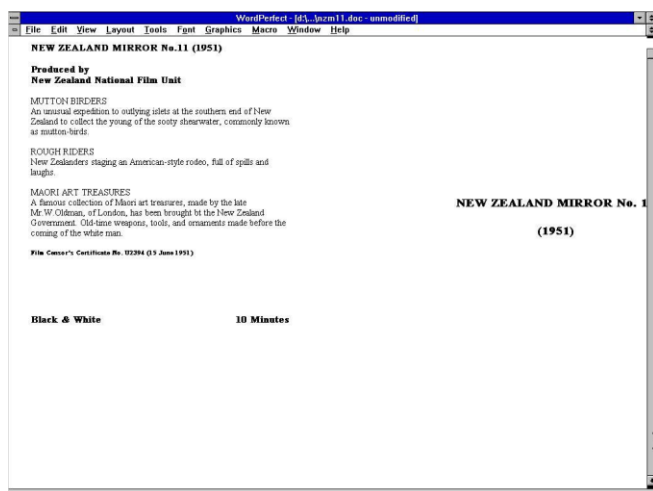


Figure 4 A WordPerfect file opened in WordPerfect 5.2 for Windows 3.11

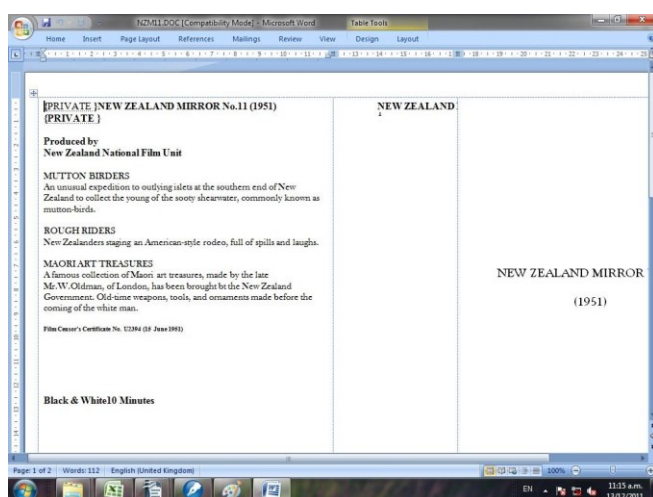


Figure 5 A WordPerfect file opened in LibreOffice 3.3.0

In both of these examples we see that digital objects are made up of many files in addition to the primary data files. All of the files involved provide sets of instructions that are executed by the computer to produce the performance or information experience which users can then interact with. It is trivially true that changing the primary data files is highly likely to result in a change to the resultant performance. This is because the computer will interpret the different instructions in the file(s) differently, producing a different performance. However, as we have seen in the examples above, changing the software files will also. All of the files involved in an object, whether those that are part of the software applications or the primary data file(s) that make up the object, have the same function as part of the object, the function as instructions interpreted by the computer. This

being the definition of software [3], it illustrates the general case that, functionally, all digital objects are software. So, to preserve any digital materials in a way that is meaningful and verifiable for users, we must be able to preserve all components of a software performance - the primary data files (e.g. the .doc, .xls, .wpd files), and the files that make up the software applications involved (e.g. the executable files, and many of the other files in the installation directory of each application).

To this end, software preservation and emulation proves valuable in a number of distinct use cases, including the following:

Providing Verifiable Digital Evidence Requires Software to be Preserved and Accessible

Formal interactions in the modern world are increasingly undertaken using digital technology. From property purchases, to banking, to communicating, to conducting and evaluating research, to legal proceedings, our formal interactions are increasingly 'born-digital' (originating in digital form [4]). With the COVID-19 pandemic this trend has only accelerated as we've collectively sought ways to continue operating society without having to interact in person. Given the digital nature of these formal interactions, the records of these interactions, and the records of government and society in general, are also now mostly born-digital.

Records are preserved primarily to provide evidence of activities that occurred [5]. To provide trustworthy evidence, records must be verifiable. As we saw in the earlier examples, if we only preserve the primary data files from digital objects, and not the original software application's files, then we will not be able to provide access to the performances. If we cannot provide access to the performances, it follows that we will not be able to provide access to the digital objects themselves (i.e. what results from the performances, the experience we can interact with). Without the software the objects can change in ways that can become not just incorrect, but potentially deceptive. In addition, the primary data files themselves (e.g. the .xls, .docx, .pdf, .psd files) should be considered functionally to be software. Since the primary data files depend on other software files, and are software themselves, in order to preserve access to trustworthy, verifiable digital evidence, we must preserve access to software.

Explaining (Selling) the Importance of Digital Preservation and Archiving by emphasizing the age of digital objects

When trying to explain why digital preservation is important it can be quite difficult to connect digital files with value, age, and importance. However, by presenting files in conjunction with legacy software, users much more quickly see and experience the age of the objects, which in turn makes them seem valuable as a consequence of their perceived age. The idea of needing to preserve things that seem “old” is already well-embedded in our cultural consciousness, so visibly emphasizing the passage of time makes it that much more “obvious” to users and our wider stakeholder community that digital preservation is important.

As the EaaS community has grown, we’ve experienced many instances when showing EaaS in action where users experience a strong reaction to seeing and hearing “old” legacy software running again. Since the microcomputer revolution of the 1970s, every generation has memories of their first experiences with computers. Sounds like a cassette tape loading a game into a Commodore 64, the Windows 95 start-up sound, the iconic “you’ve got mail” notification, the ubiquitous iPhone default ringtone, and the sound of a Skype call connecting, invoke visceral reactions in many of us. The look and feel of software from the 1980s and 1990s have themselves become so iconic as to provoke the creation a new genre of retro-software interfaces by modern-day fans (e.g. this recent game <https://www.kickstarter.com/projects/yachtclubgames/mina-the-hollower>). This emotive reaction coming from users of EaaS often helps the EaaS team and stakeholders when we are subsequently explaining why everything we do in digital preservation and digital archiving is important. Since things from the past are old, there is an assumption that they must need work to keep them accessible. This association between the emotive reaction and the assumption of work required to keep old things accessible, makes the task of explaining (or selling) the value of digital preservation much easier.

Emulation and Software Preservation are (together) the Only Option and/or the Only Economical Option

There are limited tools available for maintaining access to digital objects as technology changes. The primary candidates for maintaining access to born digital objects are:

a. Migration

Migration is a valuable method for ensuring content in digital objects can be reused in modern software. Migration is the process of replicating some content from a digital object performance using a different primary data file (or files) and a different set of software from the original. Normally the new software works on modern computers whereas the original software is considered functionally obsolete when the objects are migrated.

The challenges migration presents to digital archives are at least two-fold. Firstly, most archives retain the original primary data files from a digital object when they migrate content from the object. This means that normally their storage requirements roughly double after the migration has completed. This extra storage requirement has an economic and environmental impact which can be considerable, especially over time. Secondly, it is currently extremely costly to validate the results of a migration process. In research undertaken at Archives New Zealand [16], Cochrane found that it is very difficult to automatically test a migration process due to the difficulty in automatically identifying changes made to an object. This is partly due to the finding that most objects seemed to include at least one rarely used software feature, and so methods that use shortcuts that exclude rare features are not effective at scale. Another reason for this problem was identified as behavioral: some users used software in ways that meant the digital objects could only present their information using that software but did not use features that could be automatically looked for and automatically validated post-migration. This meant that there was no way to automatically/programmatically identify the features in the file to be migrated.

Cochrane also established that manually testing a migration-equivalent process took on average 9 minutes per object. Figure 6 shows a table from Cochrane’s “Rendering Matters” report [17] that extrapolates the time it would take to test the outcomes from migrating various percentages of various numbers of objects:

Table of time taken to test x percentage of objects (hours)										
Number of Objects	Percentage tested (%)									
	0.5	1	2	5	10	25	50	75	100	
100	0.075	0.15	0.3	0.75	1.5	3.75	7.5	11.25	15	
1,000	0.75	1.5	3	7.5	15	37.5	75	112.5	150	
5,000	3.75	7.5	15	37.5	75	187.5	375	562.5	750	
10,000	7.5	15	30	75	150	375	750	1125	1500	
25,000	18.75	37.5	75	187.5	375	937.5	1875	2812.5	3750	
50,000	37.5	75	150	375	750	1875	3750	5625	7500	
100,000	75	150	300	750	1500	3750	7500	11250	15000	
250,000	187.5	375	750	1875	3750	9375	18750	28125	37500	
500,000	375	750	1500	3750	7500	18750	37500	56250	75000	
1,000,000	750	1500	3000	7500	15000	37500	75000	112500	150000	
2,000,000	1500	3000	6000	15000	30000	75000	150000	225000	300000	
5,000,000	3750	7500	15000	37500	75000	187500	375000	562500	750000	
10,000,000	7500	15000	30000	75000	150000	375000	750000	1125000	1500000	

Figure 6 Table of an extrapolation of the time taken to test the migration of digital objects

Many archives now have over a million objects in them. Assuming this table is accurate, to test only 0.5% of 1,000,000 objects would take one person nearly 19 weeks working 40 hours a week with no breaks. Scaling manual testing of migration processes in this way would seem to be very expensive for most archives, and likely uneconomic. Additionally, many modern “documents” have no format to migrate to. Google Docs documents don’t have a “file format,” for example. Content from them can be exported into files with formats like .docx and .odt but natively they are made up of database entries and the other files that make up the server-based Google Docs software. The only way to preserve these performances (outside of rewriting the software) is to preserve the software that serves the “documents” in our web browsers. In the case of Google Docs, this would likely require preserving the software on multiple servers, including web servers, database servers, file servers and application servers, as these are required to provide the performance that we interact with when we “open” a document in Google Docs. We would also need to preserve access to a compatible web browser. Finally, cycles of both creation and retention of born-digital objects are unpredictable and archives may have to retain access to migration tools for a very long time if they intend to continue receiving old digital objects throughout that cycle. For instance, the donor of a digital collection may have continued using a particular piece of software long past its “official” support date or “end-of-life” (then donated even longer past). Given that, archives would likely have to use emulation to run migration tools as the tools themselves become functionally obsolete on current computers; it is not clear why an archive that was continually collecting would spend time migrating documents when they would always be able to migrate them just in time for when they are needed.

b. Normalization

Normalization is primarily described as undertaking migration on objects as soon as they are received in an archive. This is intended to ensure they are always available in a format that is compatible with modern software. This has all the same issues as migration, if not more so as archives immediately incur the costs of both migration and the preservation of at least two copies if they keep the originals.

c. Re-Writing of Software

Re-writing or recreating software so it is compatible with current computers is another option that has been proposed for ensuring access to digital objects over time. However, if we must re-write software to account for every change in file formats over time (as we assume *eventually* every format will become obsolete), this could rapidly become economically unfeasible. Given that there are currently at least 12575 file formats (as documented in Wikidata.org), that this number will only go up over time, and that testing to ensure each format + new-software combination would have to be extremely thorough to ensure the new software doesn’t change the information experience/performance, re-writing software is likely only economically practical when it is applied to re-writing emulators. Re-writing one emulator could ensure access to many emulated computers, which could ensure access to many legacy software applications which could ensure access to virtually unlimited digital objects.

d. Emulation

Emulation is the most economically feasible method for ensuring long term access to digital objects because in principle (discussed below) and in practice [18], it scales extremely effectively. As discussed, one emulator can ensure access to many emulated computers which can each ensure access to many legacy software applications which in turn can each ensure access to virtually unlimited digital objects. It’s also possible to preserve emulated computers at one organization permanently, and then share the emulated computers on demand-only when needed. In doing so, the work to create, document, and preserve emulated computers and the legacy software “environments” they provide access to can be distributed such that even small organizations can potentially afford to implement emulation and use preserved software to provide access to their content.

In addition, the work to verify the effectiveness of an emulator can be shared. Computer hardware compatibility verification is a well-established process that has always resulted in differences between specific hardware configurations (e.g., there are minor differences between all individual computers), changes that are accepted. The entire Personal Computer (PC) and software industry is predicated on the assumption that software can be installed on any PC that meets basic "compatibility" requirements. So fortunately, it is only those relatively lightweight requirements that ever need to be met to verify that an emulator is sufficient. This is in contrast with the work to verify the effectiveness of a migration process, which must be tested for every individual object as users don't expect their objects to change, especially when they are the thing that is meant to be being preserved.

The scalability of emulation and software preservation for digital preservation makes investments in emulation-based solutions worthwhile in most situations. In addition, emulation is a "just in time" method, i.e., it is only used just in time for when it is needed. This means that most of the time, once emulated computers have been created, they can be stored at a few organizations, then accessed by many more, only when they are needed. Furthermore, if future users want to reuse data from preserved objects, the data could still be migrated out into new files by using the original software to do so (if necessary, by chaining multiple applications together to move data between multiple formats). This would have the dual benefit of also enabling users to see what was lost during the migration (by enabling them to compare with the original in emulation) and enabling migration to occur whenever, and only when it was needed. In this scenario the work to verify the migrated data can be undertaken just in time also, further saving resources.

As well as being the most economic option, for a variety of objects software preservation and emulation are the only option for ensuring continued access to digital objects. Microsoft Chart files, for instance, no longer open in any modern application. In addition, games, disk images, and other types of complex digital objects often simply can't be migrated to new technologies.

Emulation For Appraising Digital Content

Until recently, objects that have been received by memory institutions as digital files (either 'born-

digital' or 'received-digital' (digitized elsewhere) have often been transferred to the institution on external media. This is beginning to change as organizations find better ways to support network/internet-based transfers [6], but for most memory organizations this is the primary way digital objects are transferred to their organizations and will be for quite some time. In addition, organizations will continue to receive digital storage drives or disk image files made from drives that come from the desktop computers of notable people and likewise for drives from servers from notable services/systems.

Disk images can be difficult to appraise, especially those representing the drives of entire computers where the context in which the data on them was used was as part of live systems. However, appraisal archivists/practitioners are usually unable to access that context and are limited to (at best) browsing the file systems on the drives to evaluate the value of any specific files or groups of files that they can find. This usually involves either:

- I. Opening a copy of the disk image (to ensure no changes are inadvertently made to the original) in a disk image review application (like FTK Imager), browsing the file system within the application, and exporting interesting files
- II. Mounting the disk image (likely in read-only mode) on their local file system and browsing it directly using whatever file explorer application is on their operating system

They then open the files with whatever "compatible" software is available on the modern computer they are using. Usually this means opening them in the software that the operating system has associated with a particular file extension, MIME type, or type code.

As we have established, opening files in non-original software can cause changes to be introduced to the information presented by the computer when opening the files. This can mean that when appraising files in the way described above, appraisal archivists may not see the value of something that did have real value. There are also cases where multiple files need to be interacted with together, along with software, in order to view the compound-object that they represent. Many office documents are tied to each other through Object Linking and Embedding (OLE) functionality [7] for example, and do not function or are missing content if all the

required files are not present and accessible in the same software application. Emulation allows appraisal archivists to instead open the disk image itself as a virtual computer if it contains an operating system or open the disk image attached to an emulated computer with software installed on it that is contemporary to the content on the disk image. In both cases, the ability for the archivist to understand the content and view it in a meaningful way, is greatly increased.

Computers and Software Applications are Historic Artifacts

Our focus should not be on digital objects alone, as current research has extended to address the historical notoriety of computer systems and software applications, representing a distinct field of study for which emulation is an essential component.

Computers are the artists', engineers', programmers', authors', and regular workers' toolkits of the current age. There exists a huge opportunity to archive, preserve, and make accessible these toolkits for future generations to not just view and interact with, but to reuse to create new outputs in the future in much the same way some artists use very old techniques to create their art. Without the ability to preserve software, or the ability to emulate old computers, preserving computer environments as artifacts will be impossible.

Many software applications are historic artifacts. From Microsoft Office's 'Clippy', to voting machine software, from minesweeper to Minecraft, software applications have had and continue to have a huge impact on society, and for this reason alone should be preserved for posterity.

III. WHY EaaS

For the average user, obtaining a legacy software application can be very difficult, and once obtained, legacy software can be challenging to install, authenticate, configure, and operate. Older applications are frequently unable to function on modern operating systems, and even when the requisite operating legacy systems can be found, they in turn, are unlikely to function, or function well, on contemporary computer hardware. The problem is only increasing as our computing hardware continues to advance.

Emulators solve many problems, by allowing users to easily run legacy software on modern computers.

Emulators are themselves software applications, applications that simulate one computer on another computer, allowing users to install and use software on the simulated computer. An emulated computer is a computer that is simulated or "emulated" using an emulator software application. Emulators are most often used to simulate computers that have a "hardware architecture" that is different from the computers that the emulators are being run on. This allows the user of the emulator to run software that is compatible with the emulated hardware but not compatible with the hardware that the emulator is running on.

Despite solving many problems, emulators also come with many challenges. Emulation technologies can be difficult to employ and particularly challenging to employ at scale. Emulators often require specialist expertise to use and are non-standard tools that information technology departments rarely support. These barriers to the large-scale use of emulators have been addressed by the "bwFLA Emulation as a Service (EaaS)" framework [8].

The EaaS framework enables emulated computers to be made more easily accessible and allows, via only a web browser, for seamless access to the software running within the emulated computer. With EaaS, users do not need to understand how to configure an emulator to ensure it runs a particular operating system. Instead, the EaaS framework provides templates for pre-configuring emulators to support a wide range of operating systems.

In addition, EaaS provides a way to save storage space when scaling the use of emulators for different applications that have similar dependencies. With EaaS users can create "derivative" computing environments (or "environments" – the term we use to refer to both the emulated computer and the software installed on its virtual drive) that are created by saving changes made to an existing environment in a separate file from that which stores the main environment. When the new environment is then re-run it uses both the derivative-environment file and the source environment file at run-time to provide the full environment experience. In doing so this saves the user from having to save copies of all the data that would be the same between two environments that are only different in a small way (e.g. one may have an additional application installed on it).

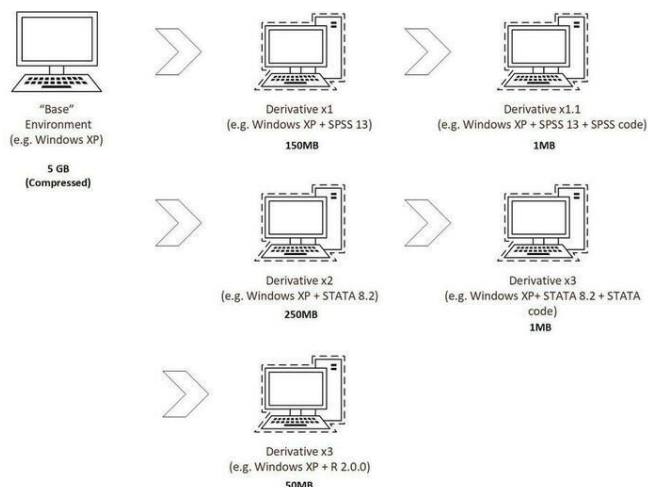


Figure 7: In EaaS, derivative environments can build on dependencies from base environments without the need to save redundant data.

IV. WHY EAAI

In starting the EaaSI program of work we recognized that the EaaS framework had the potential to enable the wide-spread use of hardware emulators and pre-configured “software environments”, at scale in cost-effective and efficient ways, and thus to become an invaluable resource for digital preservation efforts worldwide. However, EaaS does not alone solve the problem of finding software, or of the work required to configure and document all the software applications that we need to preserve. To do this the digital preservation community needed to make legacy software easy to find, and to share the work to configure and document emulated computers and the software applications they run. EaaSI was developed to fill this gap and to make the use of emulators and legacy software for access, easy.

EaaSI’s primary goal is to enable the scaling of access to both emulation technology and the use of legacy software for providing access to digital objects. To address the latter, we have established the open source EaaSI software and an EaaSI network in North America (with an additional nascent EaaSI network recently starting in Australia [9]). The North American network currently contains sixteen members, primarily at large research Universities. The EaaSI software is built on EaaS and makes use of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to enable users to request, share and synchronize metadata between nodes (installations of the EaaSI software) participating in

an EaaSI network. This allows one organization to find a software application, install it, configure it, and document it, then publish its metadata to the network. All other nodes in the network can regularly harvest published metadata so local staff-users can see what is available across the network. Staff-users then choose to replicate emulated computers (or “environments”) from other nodes and run them locally to accomplish a goal. For example, a staff-user may replicate an environment containing a Computer Aided Design (CAD) software application in order to provide access to a CAD object that a patron requested. The staff member does not need to configure the emulator (thanks to EaaS), or find, configure, or document the software (thanks to EaaSI). Instead, they may simply import their digital object, load it into an existing emulated computer, configure the file to open in compatible software, and save their work as a new derivative environment. This saves a significant amount of time and resources and makes using both emulation and legacy software to accomplish practical tasks, simple for the staff member.

EaaSI also provides a new interface for EaaS in which users can more easily discover, access and document software and computer environments. By making it easier to document software, we hope to ensure that it is easier for future users to use in the future when interface paradigms have changed.

Using these approaches of sharing and simplification, we are aiming to make the process of accessing content using original legacy software so seamless that users forget that it is complicated and expect legacy software to be available and usable whenever they need it. Users have the option to fully configure all aspects of the software and emulation tools themselves and thorough documentation is included with EaaSI [10]. However, our goal is that such configuration is always optional and normally unnecessary. To fulfill this vision we’re working to turn EaaSI into common infrastructure, shared and used by organizations everywhere.

V. BUILDING SERVICES ON THE EMULATION AS A SERVICE INFRASTRUCTURE

We’re working to enable legacy software to be seamlessly integrated into any long-term access and

re-use contexts, to become core-information management infrastructure. To do this we are working to build out EaaS's Application Programming Interfaces (APIs), to create tools using those APIs, and to facilitate a community-based development process. Focus on these strategies is intended to both solve problems and add to EaaS's practical use right now, but also inspire others to build on the EaaS infrastructure to solve problems in additional domains.

The Universal Virtual Interactor (UVI)

The UVI is a tool we have developed on top of the EaaS infrastructure to provide the ability for users to click on a link to a digital object (for example in a library's catalogue or an archival finding aid) and have it automatically open in a representative version of the "original" software, within their web browser, using an emulator. The name (and the use of "virtual" despite it also/primarily using emulation in addition to virtualization) is an homage to the Universal Virtual Computer (UVC) concept developed by IBM and the National Library of the Netherlands (Koninklijke Bibliotheek, KB). The UVI is intended to be "universal" and (theoretically) work with any files/digital objects. It's called an interactor not a "viewer" or "renderer", as it's not just for "rendering" or "viewing". Rendering and viewing are primarily passive activities, but digital object experiences are not passive, they're interactive. We want to be able to enable users to interact with their digital objects presented as an experience that is as close to the "original" as possible. That interaction might include such things as turning on and off "track changes" functionality in a document, viewing embedded metadata through standard application menus, browsing and submitting queries through database interfaces, interrogating and temporarily changing spreadsheet formulae or embedded scripts, etc.

With further development the UVI could be integrated with discovery and access platforms and configured to give users the choice of which software to use to complete their digital object performances. Further they could select a file from a catalog or finding aid, have it matched to multiple environments in EaaS by the UVI, then choose which environments to use to complete the information experiences/performances. This can all be completed on-demand, just in time for when the user requests it.

EaaS Virtual Reading Rooms

EaaS is designed to allow staff to add files to an emulated environment and securely share access to the environment with one or more users. EaaS can provide a dedicated access page for the environment, or an environment can be embedded on any arbitrary/custom page via HTML. This process could be used with legacy software environments that have many applications on them to provide access to multiple digital objects at once using a single environment. We are working to refine the process, particularly for providing access to secure materials, as a "Virtual Reading Room" service so that it could be seamlessly integrated with existing discovery and access platforms, services and workflows. In this increasingly remote-working context we hope this toolset will prove particularly attractive, not just for providing access to older objects using legacy software, but potentially for providing secure access to any restricted digital content.

EaaS Community

While design and desired functionality for various pieces of EaaS tooling may at this point be well-articulated (see above), converting emulation and EaaS into core services requires constant, open, critical feedback from its intended user community on implementation. The team regularly convenes and solicits input from current and prospective users, investing in and facilitating paths of communication (recurring calls with representatives of the North American EaaS network; an online Community Forum and issue tracker with registration open to all[11]) to ensure both that EaaS services address real-world needs and workflows, but also that the program remains tapped in to challenges potentially beyond the scope of EaaS tooling alone (see below).

VI. CHALLENGES

Legal/Copyright

Copyright law provides the biggest challenge to scaling EaaS globally. In the United States many of us are fortunate enough to be able to rely on the rights defined in the Copyright Act that are described in the Code of Best Practices in Fair Use for Software Preservation [12]. The code describes the legal grounds upon which the EaaS network participants are operating. However the situation in the United States is unfortunately not replicated globally, and

even what exists in the US, is not ideal. There continue to be challenges in the United States in a number of areas of copyright law including the need to circumvent DRM in order to maintain access to usable software (despite recent progress with Digital Millennium Copyright Act exemptions). Making progress in the area of copyright law will be one of the continuing challenges all EaaSI users will need to focus on in coming years.

Lack of Emulators

EaaSI is relying on existing open-source emulators developed primarily by volunteers. We're very grateful for their work and are able to provide access to many different types of emulated computers as a result. However, there are still gaps in our library of emulatable computers. Over time these gaps may grow unless additional emulators are created and integrated into the EaaSI framework.

Distributed Digital Objects

As discussed, more and more objects are becoming increasingly distributed across networks and the internet. Modern CAD/BIM designs are often dependent on files spread across multiple servers and computers with embedded references that are easily broken. Online web services often rely on databases and other files stored on multiple different computers ("servers") that together present a dynamic information experience to users on request. Modern video games are often sold solely as digital downloads, are constantly updated with new patches (which makes it hard to track their versions over time), and often have to be constantly connected to remote servers in order to be played. Mobile applications are often little more than a front end to various web-services and will become non-functional when those services go away. Some work has been done to address this by prototyping methods for preserving web services [13], and our work to add the ability to network devices in emulated networks will help to provide some infrastructure that could help address these issues.

Integration

To make EaaSI successful at scale it needs to be seamlessly compatible with many different types of systems. At the most immediate level, EaaSI needs to be able to integrate with digital preservation/digital repository systems so it can retrieve disk images and content from them for

management and access. It also needs to integrate with access and discovery platforms to enable environments provided by EaaSI to be made available directly to users via these existing platforms.

Integration with access and discovery platforms presents new challenges however. As we work to provide access to entire databases, web servers, and desktop environments from notable individuals using EaaSI, we will need to develop methods for indexing and documenting the content in them so they can be made discoverable through existing discovery tools and systems. Once discoverable there will be demand to enable direct linking into content that exists inside of environments made accessible by EaaSI, something that will be relatively challenging due to a lack of generic ways for achieving this across different computing platforms and legacy operating systems.

Finally, we will need to be able to easily get data in and out of legacy systems running in EaaSI, and to connect modern computers to services running in EaaSI via the legacy APIs. For example, users will likely want to be able to connect remotely to database servers running in EaaSI in order to be able to query them from modern applications. Fortunately, the querying protocols for many of these scenarios still have modern implementations that could enable that, but making secure connections to legacy computers running in EaaSI will require new features, and will need to be made as user-friendly as possible due to the potential for the knowledge for how to make this function fading over time.

Application Signatures and a Registry for Software

The National Archives of the United Kingdom (TNA) has done the world a great service with its creation and support of both the PRONOM file format registry and the DROID file format identification tool. These are widely used and acknowledged as some of the most important tools within the digital preservation community [14]. Being able to identify file formats is particularly important for undertaking migration as a strategy as it helps with matching migration tools to files. There is an equivalent approach that could be beneficial for an emulation strategy. By automatically identifying the interaction applications of objects using "application signatures" we could match primary data files to their interaction applications. Currently though, application

signatures to feed to a tool like DROID or Siegfried do not exist, and even more critically, there is no central registry for software applications that could provide us with identifiers to match to. PRONOM does include a very limited set of software metadata and Wikidata.org includes a great deal more. However, neither are likely appropriate as homes for signatures for interaction applications, nor for the information we might want to share publicly about software environments that include the applications. In other words, there is a gap here. A gap we will aim to fill in the future.

In the meantime, we are working to include information about software and environments that we're adding to EaaS into the Wikidata.org database. This data will then be able to be made available to digital preservation systems via the Preservation Action Registries API [15].

VII. THE FUTURE

EaaS is not static. There is still ongoing conceptual and technical development of EaaS, and likely will be forever. With a mission of "Useable software, forever" EaaS will need to evolve as the software landscape evolves. In the near term this will include needing to move past primarily supporting the emulation of single personal computers to other domains such as networks of computers, mobile devices, server machines, and whatever comes next. In this section we outline some of these initiatives and describe how we're moving the EaaS platform forwards to meet the needs of the changing digital landscape.

Mobile

Mobile computing has taken over the computing world over the last couple of decades. There are more mobile devices than personal computers and their cultural, scientific, and economic impact continues to be hugely historically important. For these reasons the EaaS team has recognized the need to be able to preserve and provide access to mobile operating systems, applications, and the files and data they support. We are in the process of adding an emulator for Android-based devices to the EaaS platform as a first step to exploring how best to address the challenges of preserving the mobile computing universe. Preserving mobile computing experiences provides many challenges, from replicating the experience of "app stores", to re-creating or simulating the various network services

that mobile applications rely on, to simulating output for the myriad of sensors that mobile devices have built into them, and finally to providing meaningful replicas of the experiences of interacting with the huge variety of physical devices that have made up our mobile universe since the early 2000s. By starting small we hope to provide a testing ground for trying different options for addressing this growing list of challenges.

Networks

Computers have been connected in networks since at least the late 1950s. Home computer users began networking their machines to others at scale with the wide availability of dial-up modems in the 1980s which led to services like Bulletin Board Systems (globally) and Minitel (in France), and finally to the internet that we all use today. Networks of computers and devices have become so prevalent and ubiquitous that we no longer think twice about the complexity involved in just connecting our phones to our local wireless networks when returning home. However, networks are complex. They involve multiple computers and services running on them that have to be orchestrated together to function appropriately. With EaaS we are working to enable users to create emulated "networks" made up of multiple emulated computers that are connected to each other and can share information between them over standard networking protocols. These may include complex research environments, email servers, database management systems, and more. We have a functional prototype of this software that we are currently in the process of adding to the core EaaS platform. This will be another transformative tool for our users. It will open a new set of potential use cases for the application of EaaS to ensure long term access to large scale systems, enterprise databases, functional interactive websites and more.

Automation

The task to configure and document legacy software is increasingly urgent as knowledge of how legacy software works and how to configure it is fading from institutional memories as those who used it leave the workforce. Increasing complexity and technological change means that in the future we'll have to have ways to enable users to interact with software that has interfaces that they've potentially never experienced. Even where software has been configured and basic documentation

created, it is often difficult for users to accomplish tasks, especially in complex enterprise systems such as those used in government and industry. To address this, we are developing methods to record interactions in EaaSI and play them back with specified parameters. This will allow future researchers to accomplish tasks using software running in EaaSI (such as finding all information about a user in a database) just by clicking a labeled button or calling an API function.

We have also developed a functioning prototype of a tool to automatically record metadata about compatible file formats using two different methods: 1. Optical Character Recognition (OCR) applied to the interfaces of graphical environments in EaaSI and 2. using the Windows Operating System programming interface to read the menu items and pass them out of the emulator. This can be used to automatically record metadata about the format's applications can open, save, export, and import, further reducing the manual effort needed when adding new applications to an EaaSI network.

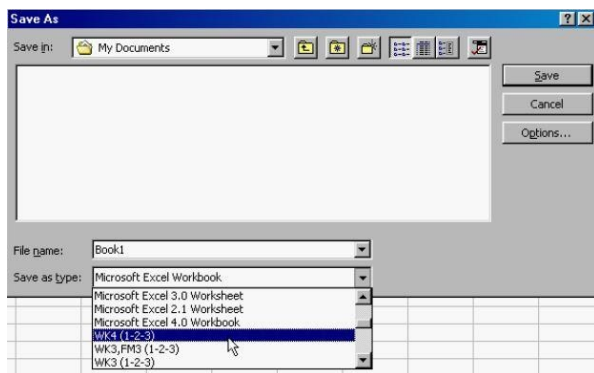


Figure 8 An example of the software application format compatibility metadata that we have developed a tool to automatically document.

We are further working to make use of this metadata to revive the idea prototyped in the PLANETS project of enabling automated migration of content between digital objects using emulation ("migration by emulation") [16]. This will allow on-demand migration of content from legacy digital objects into new objects that make use of modern software. It will make use of the original software to migrate content to a more open or newer format and this process could potentially be run multiple times in a "chain" process to create objects that are fully functional on modern computers.

VIII. CONCLUSION

Ensuring software is usable forever will be an ongoing challenge. However, it is essential for ensuring society's knowledge, culture, and history persist over time. While we're focused on using software to ensure content is unchanged over time, we also believe that regular reuse of content in archives helps keep both the materials and the organizations that steward them relevant. Enabling the use of emulation at scale will ensure that we can continue to migrate some content out of old digital objects and into new ones, at scale. This in turn will ensure the extractable content stays usable and therefore relevant to whomever can benefit from it in the future.

With the Emulation as a Service Infrastructure program of work we are building and freely sharing infrastructure; infrastructure that we hope is not just useful, but inspiring. We imagine a world in which digital content is always usable, regardless of the software that is part of it. To get there we are going to need others to build on what we have started. With the first EaaSI services: the UVI and the virtual reading room functionality, we hope that we are showing some of the potential for how EaaSI could be used at scale to solve long term preservation and information management issues. We don't know what else might be achievable with this infrastructure, but we're excited to find out.

REFERENCES

- [1] *Why Software Is Eating The World*. M. Andreessen. Wall Street Journal. August 20, 2011. Retrieved 8th of March 2022 <https://www.wsj.com/articles/SB10001424053111903480904576512250915629460>
- [2] *An Approach to the Preservation of Digital Records*. H. Heslop, S. Davis, A. Wilson. 2002. National Archives of Australia. <https://www.naa.gov.au/sites/default/files/2020-01/An-Approach-to-the-Preservation-of-Digital-Records.pdf>
- [3] *Software*. Wikipedia Contributors. 8th March 2022. <https://en.wikipedia.org/w/index.php?title=Software&oldid=1071678404> (last visited Mar. 8, 2022)
- [4] *Born-digital*. Wikipedia Contributors. 8th March 2022. <https://en.wikipedia.org/w/index.php?title=Born-digital&oldid=1064529162> (last visited Mar. 8, 2022)
- [5] *Record*. The Society of American Archivists. <https://dictionary.archivists.org/entry/record.html#:~:text=D%20earst%20evidence%20of%20events>. (Last visited Mar. 8, 2022)
- [6] *Exactly: A New Tool For Digital File Acquisitions*. B.Lyons. AVP. January 13, 2016. <https://blog.weareavp.com/exactly-a-new-tool-for-digital-file-acquisitions> (last visited Mar. 8 2022)

- [7] *Object Linking and Embedding*. Wikipedia Contributors. https://en.wikipedia.org/w/index.php?title=Object_Linking_and_Embedding&oldid=1049945261 (last visited Mar. 8, 2022).
- [8] *bwFLA - A Functional Approach to Digital Preservation*. PIK - Praxis der Informationsverarbeitung und Kommunikation 35(4):259-267. November 2012. DOI:10.1515/pik-2012-0044
- [9] *Australia will be getting its own Emulation as a Service Infrastructure (EaaS) network soon* [Online forum comment]. M. Swalwell. 2nd February 2022. https://groups.google.com/g/australasia-preserves/c/gaKLhFOTALE/m/7CkB-FH6BAAJ?utm_medium=email&utm_source=footer
- [10] *The EaaS Handbook*. The EaaS team. 2022. https://eaasi.gitlab.io/eaasi_user_handbook/ (last visited Mar. 8, 2022)
- [11] *EaaS Community Forum*. <https://forum.eaasi.cloud> (last visited June 23, 2022)
- [12] *Code of Best Practices in Fair Use for Software Preservation*. Software Preservation Network. Revised 2019, <https://www.softwarepreservationnetwork.org/wp-content/uploads/2020/02/2019.2.28-software-preservation-code-revised.pdf> captured at <https://perma.cc/FK5K-MWJR>.
- [13] *Preservation strategies for an internet-based artwork yesterday, today and tomorrow*. C. Roeck, R. Gieschke, K. Rechert, J. Noordegraaf. Proceedings of the 16th International Conference on Digital Preservation iPres 2019, p179. <https://ipres2019.org/static/proceedings/iPRES2019.pdf> DOI 10.17605/OSF.IO/GF2U9
- [14] *"And it's official! PRONOM wins the #DPWorldCup!"*. Tweet message. The Digital Preservation Coalition Team. Twitter. 25 February 2022. https://web.archive.org/web/20220225145245/https://twitter.com/dpc_chat/status/1497222900482457602
- [15] *PAR Overview*. PAR team. Accessed 8th March 2022. <https://parcore.org/>
- [16] *Demo: Migration-by-Emulation*. I Valizada, D. von Suchdoletz, K. Rechert. 2011. Proceedings of the 8th International Conference on Preservation of Digital Objects: iPRES 2011 - Singapore. P248. <https://phaidra.univie.ac.at/o:294293>
- [17] *Rendering Matters*. E. Cochrane. Archives New Zealand. January 31, 2012. <https://web.archive.org/web/20130218111126/http://archives.govt.nz/rendering-matters-report-results-research-digital-object-rendering>
- [18] *(Re-)publication of Preserved, Interactive Content - Theresa Duncan CD-ROMs: Visionary Videogames for Girls*. D. Espenschied, et al. Proceedings of the 12th International Conference on Digital Curation. Chapel Hill, NC: University of North Carolina, School of Information and Library Science, 2015. P233/241. <https://phaidra.univie.ac.at/view/o:429524>

HOW DO USERS DISCOVER DIGITAL PRESERVATION TOOLS?

Report on a survey of professionals

Amber L. Cushing

University College Dublin
Ireland

Amber.cushing@ucd.ie

[0000-0003-0186-0689](tel:0000-0003-0186-0689)

Tess Burchmore

University College Dublin
Ireland

Sarah Conroy

University College Dublin
Ireland

Phoebe Doyle

University College Dublin
Ireland

Niamh Hegarty

University College Dublin
Ireland

Rebecca Kelly

University College Dublin
Ireland

Payton Kufeldt

University College Dublin
Ireland

Morgan McGann

University College Dublin
Ireland

Cian Ormond

University College Dublin
Ireland

Gerard Quine

University College Dublin
Ireland

Martina Reba

University College Dublin
Ireland

Ronan Woods

University College Dublin
Ireland

Abstract – While information is available that charts the development of digital preservation tools and their use via case studies, less is known about *how* users discover and adopt digital preservation tools in different contexts. This study reports on a short survey of 68 professionals who utilize digital preservation tools, and how they discover and adopt those tools. Findings suggest that the role of community is important when discovering and adopting tools. Findings were then applied to digital preservation education to inform the ways in which tools are taught in formal digital preservation education programs.

Keywords – Digital preservation tools, surveys.

Conference Topics – Community, innovation.

As registries demonstrate, there are a variety of proprietary and open source tools available for use. Previous research has explored digital curation competencies [1], [2] educating digital curators [3], [4] case studies of use of certain tools and services for example [5], [6], and the consideration of tools in digital preservation workflows [7].

Yet, there is limited understanding as to *how* the professionals who engage in digital preservation discover and select these tools for use. For example, if someone was just starting out, where would they start? How would they find and select the most useful, relevant tools for their situation? While this question can help address the how and why of tool selection, it also has implications for digital preservation education.

Recently, Yoon et al. [8] surveyed digital preservation syllabi in the US and Canada and found the role of teaching digital preservation tools via class activities

I. INTRODUCTION

Part and parcel of digital preservation is the use of tools to complete to complete preservation activities.

to be an important part of formal digital preservation education. However, how can educators best go about teaching use of these tools? File formats will change over time and digital objects may become increasingly complex, so how can educators provide students with lasting skills, beyond just providing practice with the tools that are currently favored?

This paper reports on a short survey of digital preservation professionals and how they select tools for use. The survey was completed by students enrolled in the Digital Curation: Core Concepts module at University College Dublin between September – December 2021. While the findings of the survey are limited based on the sample size, they can elaborate an understanding of the context of digital preservation work practices, by shedding some light on how professionals seek out resources to support tool discovery and selection. The survey results can also be viewed in the context of digital preservation education. While the survey has the potential to inform how tools emerge and may grow a user base, it can also inform the ways in which educators approach tool use and education in the context of professional digital preservation education.

A. Selecting and using digital preservation tools

Overall, most advice related to selecting tools for digital preservation is organized via the two most well-known models governing digital preservation practice: the DCC Digital Curation Lifecycle and the OAIS Reference model. While these models organize information about tools and assist the user in discovery of tools information, little is known about how individual users discover newly created digital preservation tools and software solutions. Peer reviewed journals often struggle to keep up with the ever-changing open-source technology landscape. Conferences, blogs, websites, message boards, and social media platforms often function as the most up to date information available about digital preservation tools, advising on practical aspects of various tools and systems associated with digital preservation [9]. The Digital Preservation Coalition [10] exists “to secure digital legacy” worldwide. Their website provides a comprehensive guide to the use of open source and proprietary tools as part of a digital curation workflow, making clear that the user must consider the institutional setting when working with combinations of tools to reduce costs.

In her review of the state of the art of digital preservation in 2018, Rieger [11] listed “the

availability of new preservation systems and tools” as part of “what’s working well” in the field. Yet, availability of tools does not always translate to a general understanding of how to discover which tool is going to best fit specific work contexts. The *Digital Preservation Handbook*’s [10] “Technical Solutions and Tools” section advises that “before selecting digital preservation tools it is important to consider carefully the technical workflow and institutional setting in which they are embedded.” The *Handbook* references the Northumberland Estates Case Study, which recommends evaluating new tools using a “product analysis scorecard” which helps the user map the product/tool to the OAIS reference model compliance requirements. The guide also recommends assessing how the tool will deliver preservation actions discussed in an existing preservation plan.

The relationship between tools and workflows has enjoyed recent attention in the OSSArcFlow project. According to the project’s *Guide to Documenting Born-Digital Archival Workflows*, the purpose of the project was to document born digital archiving activity and to offer advice in the selection and implementation of workflows. According to the project guide, institutions report having “to manage significant gaps and overlaps between different tools and environments” [12]. The guide lists tools that are commonly utilized at different steps of a digital preservation workflow and emphasizes the importance of assessing how tools may be compatible within a current or ideal workflow. The guide focuses on how practitioners utilize tools on the context of a workflow, but tool selection is not discussed specifically in detail.

The *Handbook* [10] also highlights the role of Digital Preservation Registries, particularly the Community Owned Digital Preservation Registry (COPTR), a wiki based registry of digital preservation tools [13]. According to the webpage, COPTR’s “main aim is to help practitioners discover preservation tools that will help them tackle particular preservation challenges.” The Registry allows the user to search for tools according to DCC Digital Curation Lifecycle stages which was developed to assist in research data management, function, content type, and file format. COPTR’s strengths lie in it’s ability to act as a clearing house for digital preservation tools information, but providing links to further, more detailed information about digital preservation tools. However, while COPTR is well positioned to provide access to information about tools and software

solutions, there is less information available about how to select a tool to fit a specific context.

One of the most common formats that offers evaluation of digital preservation tools is the practitioner case study. In this format, the user describes their use of a tool/software solution in the context of their own individual case. These case studies can be found in formal venues such as the *International Journal of Digital Curation*, as well as conference series, such as *IPRES*, and informal venues, such as blogs. While practitioner case studies can provide insight into how a tool might work in a specific situation, the cases typically focus on implementation and review, with little detail about how one learned of the tool/software in the first place. For example, Trujillo et al. [5] describe how the five college library team came together to assess a need to work as a team to address digital preservation, including readiness for such a program. Consultation of the POWRR Registry of tools is mentioned, but there is less detail about *how* the team came to adopt Archivematica.

Another venue for the discussion of digital preservation tools is the communities that develop around the use of these tools. Tools typically fall into the categories of proprietary or open source, and both categories are accompanied by user communities that share advice and reflection about use of the digital preservation tool. Informal discussion occurs via social media, message boards, and email lists. An active community can be central to the development of an open source tool in particular, as it allows for up-to-date discussion on the which to complete tasks and to understand what needs to be done for these tools to be improved upon. These communities are worldwide, but some are targeted to a smaller, local community, such as the Dutch Digital Heritage Network “Erfgoedkit,” which provides Dutch language support for archivists selecting digital preservation tools to enhance digital heritage efforts [14].

B. Tools and digital preservation education

In the last two decades significant work has been published about digital preservation education, providing snapshots of how competencies may remain stable or change over time. This work has explored digital preservation competencies, the teaching of technology skills in digital preservation, and the content of digital preservation courses and modules [1], [2], [3], [8], [15], [16]. Much of this work cements the need to provide students with the ability

to understand technology but does not offer substantial detail about how best to teach students to find, discover, and use technical tools.

Starting in 2007, Lee et al. [1] list “understanding technology” as a necessary competency. In their discussion of digital stewardship pedagogy, Bastian et al. [3] describe technology as a necessary competency because technology is necessary in the oversight of collections, but stress that this understanding of technology must be grounded in context. The authors also discuss their development of a Digital Curriculum Laboratory, where “users of the laboratory can experiment with and evaluate tools and standards for their relevance to the kinds of content” via teaching scenarios designed by educators (p. 617). Like Bastian et al. [3], Feng and Richards [2] utilized literature review analysis and found that “hands on” technical practice in digital curation education is vital.

Using literature analysis, Kim et al. [16] list “understanding software” as a necessary knowledge and skill for digital curators. Yet in their analysis of job postings, the authors found that knowledge, skills, and abilities associated with specific tools and applications was requested in 45% of job postings analyzed in 2013. “Working in an information technology-intensive environment” was listed in 50% of the job postings. Kim et al. [16] summarized these findings as a “curation technologies competency” that included “competency required to identify, use, and develop tools and applications to support digital curation activities” (p. 79).

Yoon [8] analyzed 59 digital preservation syllabi (US and Canada only) to develop their findings and found “a need to integrate digital preservation tools and technologies into course content through class activities” (p. 1). In their case study, Cushing and Shankar [15] found that practitioners desired continuing professional development (CPD) education about how to use digital curation tools.

Existing research clearly demonstrates the need for digital preservation education to include how to select and use digital preservation tools, but this research lacks specific detail about how knowledge related to how to select tools can be delivered to students effectively.

II. METHODS AND PROCEDURE

The current study was designed by the primary author to answer the question, “how do digital preservation practitioners discover and select digital preservation tools?” In addition to responding to the

research question, a secondary goal of the project was to improve student knowledge about digital preservation practice, and also to develop student research skills in relation to digital preservation work and work practices.

The study was completed by eleven students enrolled in the 2021-2022 Term 1 Digital Curation: Core Concepts graduate level module at the UCD School of Information Studies. The students worked in teams of 2-3 students. After one team of students completing a preparatory literature review, another team designed the data collection methods for the study. A pair of students then developed the questionnaire and used it to collect the data, and a final team of three students conducted data analysis. The primary author completed the final analysis and discussion.

The online survey was targeted towards members of the digital curation field. As such, no specific group was pre-selected to participate in answering the survey questions. To reach working digital preservation specialists, the questionnaire was created and then subsequently shared by the University College Dublin School of Information and Communications Twitter channel because this platform allows for fast communication to a large group of people. To further our reach, people already working in the field were also asked to share a link to the online survey with their contacts through word-of-mouth recruitment. A well-connected digital preservation manager also distributed the call for participants via several listservs and on their own Twitter account.

SurveyMonkey was used to administer the 13-question survey which contained questions relating to the use and implementation of digital preservation tools. The survey was open for a two-week period in late Autumn, 2021. Permission for researchers to collect and analyse the answers given by participants was obtained using an Information Packet. Any free text that identified places, names, or other personal details was deidentified during data cleaning and analysis.

The survey included a mixture of question types to allow for numerical data to be extracted as well as to allow participants to express their opinions on current digital preservation practices. Questions were created following the consultation of several sources on the state of current digital preservation practices, such as the Community Owned digital Preservation Tool Registry (COPTR) which provided a list of current tools used in the field. Question types

included yes/no questions, open-ended multiple-choice questions, and a singular short answer question which asked participants to express any outstanding opinions on the implementation and use of digital preservation tools that were not covered by previous survey questions.

III. RESULTS

In total, 68 participants completed the survey. Numerical data from the yes/no and multiple-choice questions was analysed using descriptive statistics produced in Excel to highlight any trends in the data. The participant who responded to the questionnaire were frequent users of digital preservation tools. In figure 1, most reported using tools daily.

Table 1 provides multiple choice answers to the question: how do you find information about useful tools? Among the participants, 70.59% answered that it depends on a project by other institutions/organizations. The most popular responses were conferences and online communities. Next, 48.53% chose word-of-mouth and 45.59% chose blogs as their source of information. Finally, 30.88% of those who responded to the survey used social media, while 27.94% used other sources.

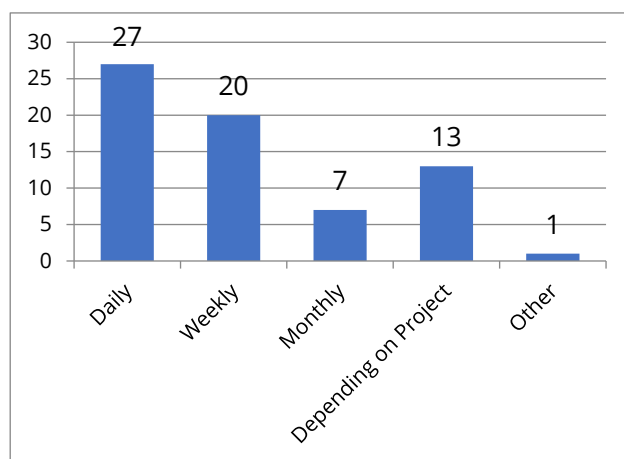


Figure 1 How often do you use digital preservation tools?

Table 1 Responses to the question "how do you find information about useful tools?"

Format	Quantity	Percentage
Following projects	48	70.59
Conferences	46	67.65
Online forums	45	66.18
Word of mouth	33	48.53
Blogs	31	45.59
Social media	21	30.88
Other	19	27.94

When asked to select the sources for information on useful tools, 70.59% of participants reported that they looked to another institution facing a similar situation for advice. After that, 48.54% of participants looked to conferences and online communities for information, followed by 30% of participants who relied on social media channels for this information. Lastly, 27% of participants selected “other sources.” These “other” sources included the Digital Preservation Coalition (DPC), Google/Web search, COPTR, Open Preservation Foundation (OPF), blogs, International Internet Preservation Consortium (IIPC), literature, peer recommendation, or sector training. The results of this question are not mutually exclusive, as many sources listed under “other” are considered to be digital preservation communities. Next we asked which digital preservation tool features were most valued (see table 2). Open source was considered to be the most valued feature of a tool, with 76.47% of participants selecting this choice. Graphical User Interface was the second most valued feature of participants (60.29%), followed by API (30.88%), and Command-line (26.47%). Only 14.71% of participants’ most valued paid commercial support. Responses to “other” (13.17%) included an active online community, good documentation and sufficient functionality.

Next, participants were asked to give their views on the following statement: “It is extremely difficult to discover and choose suitable digital preservation tools” using a Likert scale. Figure 2 illustrates this, with 27 participants agreeing, 23 being neutral, and only 16 participants disagreeing and believing that it is easy to discover and choose digital preservation tool.

Table 2 The most valued features of a digital preservation tool

Feature	Quantity	Percentage
Open source	52	76.47%
GUI	41	60.29%
API	21	30.88%
Command line	18	26.47%
Other	13	19.12%
Paid commercial support	10	14.71%

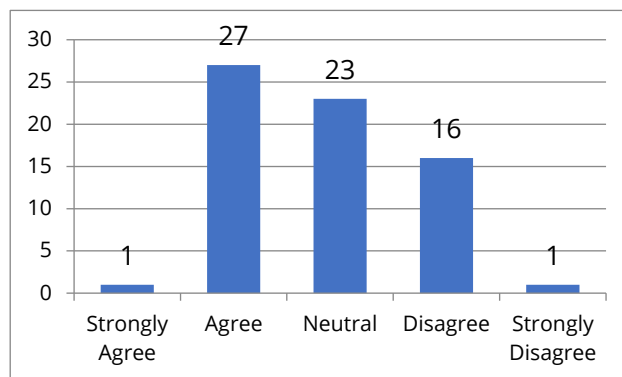


Figure 2 Responses to “It is extremely difficult to discover and choose suitable digital preservation tools”

When asked how much consideration is placed on file formats and standards when making the decision to use a digital preservation tool and/or software, most participants (51.47%) selected “a lot” followed by “some” (30.88%), “not applicable” (11.76%), “not much” (5.88%), and “none” (1.47%). In addition, when asked how much consideration is placed on information security when searching for a digital preservation tool, nearly half of participants (48.52%) selected “a lot”, followed by “some” (33.82%) “not applicable” (11.76%), “not much” (4.41%), and “none” (1.47%).

A Likert scale was used to gather responses to the statement “I place an extremely high priority on finding digital preservation tools that are sustainable with active community participation” (Figure 3).

We asked participants if they are frequent users of an online community for digital preservation tools. Slightly over half of participants (52%) stated that they were not part of an online community. Of those who were part of an online Community, the most used groups were the DPC community, Bitcurator community, and Preservica community.

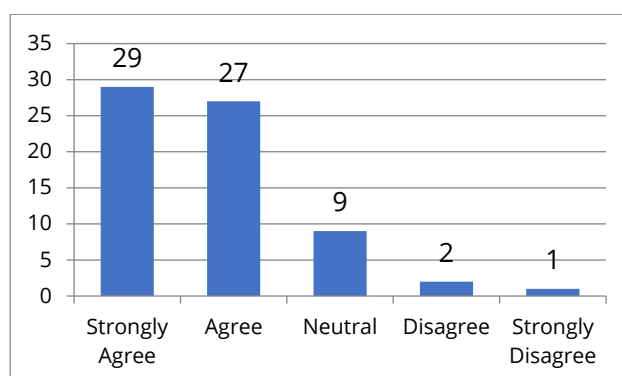


Figure 3 Responses to the statement “I place an extremely high priority on finding digital preservation tools that are sustainable with active community participation.”

There was little overlap when asked about the use of specific tools: participants listed 48 different tools,

with 34/48 tools listed being used by one person. The most common tools listed were JHOVE (15 participants), Bitcurator (7 participants) Archivematica (5 participants), and Exiftool (5 participants). Participants reported less autonomy in making decisions to utilize proprietary tools and software, with 48/68 (70.58%) reporting that the decision to adopt a proprietary tool often requires administrative, IT department, and budget approvals.

Lastly, we asked participants about the importance of Registries such as COPTR when finding digital preservation tools (Figure 4).

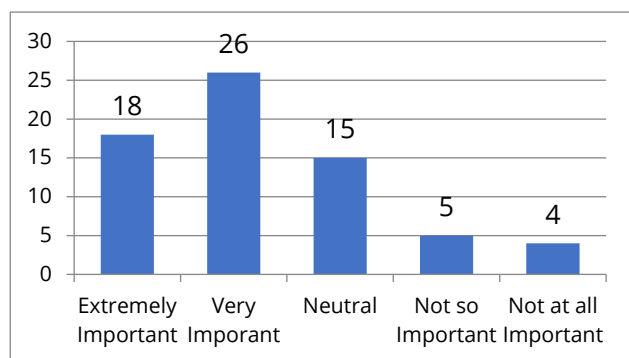


Figure 4 How important are registries such as COPTR and PRONOM in finding the digital preservation tools/software you may need?

IV. DISCUSSION

While the survey data collected provides results concerning how digital preservation tools are discovered and selected, the survey results can also contribute to a discussion of how to approach the teaching of digital preservation tools in formal digital preservation education programs. The survey results will be discussed in both of these contexts.

The survey sample size was small (68 participants), and as such, results are not generalizable. In addition, future research that involves qualitative data collection on the issue, such as a focus group, could provide context to survey results.

A. Digital preservation tool discovery and selection

Overall, the results of the survey highlight different factors that influence tool discovery and selection. Participants seemed to prefer open source to proprietary tools, with most participants citing open source as an important feature of a tool. The greater digital preservation community also plays a role, which is not surprising. However, as demonstrated in the question: “how do you find information about useful tools?” and potentially the question about frequency of use of communities, participants may

subscribe to different definitions of the term “community” in this context. For example, is a professional organization also a community? Must a community be aptly named, such as a space named an “online community forum”? Is social media a “community”? Results suggest that a “community” holds value for the discovery of digital preservation tools, but the definition of what constitutes a “community” lacks agreement. Whether this term needs defining in this context, and what benefits that might bring is unclear from the results. However, understandings of the term community in the context of digital preservation may prove a fruitful avenue for future research.

The role of community is exemplified via upkeep of the community owned COPTR registry, which acts as a clearinghouse of information about digital preservation tools and how they can be utilized. This is supported by the finding that most participants selected “very important” in response to the question that asked about the importance of tools registries. The role of a community supporting a tool was also considered to be important, with the majority of participants’ placing a high priority on tools that are accompanied by an activity community of users. This is not surprising, as tools with an active community can lend support to someone just starting out with the tool and also provide guidance on situations where the tool may be most useful. Future research might dig further into these user communities, to discern the characteristics of the communities that are considered most successful and the relationship between community and tool success and longevity, especially since many open-source tools are developed and launched using temporary financial support. As such, understanding user communities (however they may be defined) may help to understand how to make digital preservation tools more efficient and sustainable and better support novice users.

While the role of communities and community owned resources like COPTR are important, it is concerning that even though these resources exist and many find them useful, most participants agreed with the statement that it was extremely difficult to discover new digital preservation tools. A natural follow up query when viewing this data is to ponder “why?”—what makes it so difficult to discover new digital preservation tools? Rieger’s [11] challenges offer some context to position this finding. Rieger [11] cites “lack of assessment metrics” as a challenge for digital preservation. Rieger’s statement is most

likely meant to apply to the digital preservation field in general, as she points out the lack of “collaborative approaches to explore what constitutes success and how we identify it, and measure outcomes associated with digital preservation” (p. 12).

However, the same could be said for digital preservation tools: while COPTR provides information about tools and how they can be used, there is less information available that provides an assessment of tools, beyond case studies. These case studies can provide insight into how a tool was used, but the case study format is not easily compared and assessed. Further case studies can be difficult to discover, especially if they were not formally published in a conference proceeding or journal this is indexed for discoverability. The ability to easily compare and assess tools may be one of the vital missing pieces that allows a user to adequately discover *and then* compare different digital preservation tools. Perhaps, this can be addressed with the addition of an index of case studies of specific tools linked to the tools listed in the COPTR registry. This could ease discoverability of assessment information associated with tools and work toward empirical, generalizable assessment practices for digital preservation tools.

In addition, tool developers that host user communities may also create space to host case studies of tool use to ease discoverability. They could also go one step further and encourage and then highlight case studies of their tools. Finally, to build on Rieger’s [11] point, perhaps we should start to ask how best to measure success with a digital preservation tool and then formalize that measurement so that it can be accessible to the wider digital preservation community.

B. Digital preservation education and digital preservation tools

Previous research has made it clear that digital preservation education needs to address technology and the use of digital preservation tools, by providing students in-class experiences to engage with different digital preservation tools [8]. However, this in and of itself can be a challenge, as many academics do not use a variety of digital preservation tools on a regular basis and may need to rely on guest lectures from practitioners as the best resources to explain and demonstrate implementation to students.

Placed in the context of digital preservation education, the survey results suggest that in teaching

digital preservation tools, educators may contextualize hands-on activities with tools with a greater discussion of how to go about assessing differences between digital preservation tools, when resources to perform this assessment may be lacking.

This is where the role of the digital preservation community in assisting with tool discovery, implementation, and use could be introduced to students. Use of the community could also be taught via different scenarios, as Batian et al. [3] suggests. These scenarios and teaching cases could allow students to assess a landscape and propose a new tool, which may include a requirement to explain how the tool may be integrated into the wider digital preservation community, as well as how the tool can be assessed. This aligns with the call for computational thinking in archival education, which includes the use and understanding of preservation tools [17].

Finally, it is worth noting that survey participants strongly preferred open source versus proprietary tools for digital preservation. This has implications for education, as students may need to be able to assess, compare, and contrast open source and proprietary tool models. The best-case scenario would provide students experience with both formats, and the ability to make informed decisions of why to choose one model versus another model.

V. CONCLUSION

This paper reports on a survey of 68 participants queried about the discovery and selection of digital preservation tools. Findings suggest that participants have difficulty discovering new tools and rely on tool user communities for support in selecting tools to meet their needs. However, there may be different understandings of what constitutes a community. Tool registries also play an important role in tool discovery, but information about how to assess and measure different tools is lacking and may aid future tool discovery and selection. In the context of education, digital preservation educators may be well placed to position hands on experience with digital preservation tools with learning about the vital role of digital preservation communities, as well as providing students with the skills to assess tools for use.

VI. REFERENCES

- [1] C.A. Lee, H. Tibbo & J. Schaffer, Lee, Christopher A., Helen R. Tibbo, and John C. Schaefer. "DigCCurr: Building an International Digital Curation Curriculum & the Carolina Digital Curation Fellowship Program," in *Proceedings of Archiving* May 21 - May 24, 2007, Arlington, VA, USA, Society for Imaging Science and Technology, 2007.
- [2] Y. Feng and L Richards, "A review of digital curation professional competencies: theory and current practices," *Records Management Journal*, vol. 28, no. 1, pp. 62-78, 2018.
- [3] J. Bastian, M. Cloonan and R. Harvey, "From teacher to learner to user: developing a digital stewardship pedagogy," *Library Trends* vol. 59, no. 4, pp. 607-622, 2011.
- [4] P. Botticelli, B. Fulton, R. Pearce-Moses, and C. Szuter, "Educating digital curators: challenges and opportunities," *International Journal of Digital Curation* vol. 6, no. 2, 2011.
- [5] S. Trujillo, M. Bergin, M. Jessup, J. Radding, S.W. and McGowan, "Archivematica outside the box: piloting a common approach to digital preservation at the five college libraries," *Digital Library Perspectives*, vol. 33, no. 2, pp. 117-127, 2017.
- [6] E. Iglesias and W. Meesangnil, "Using Amazon S3 in digital preservation in a mid sized academic library: a case study of CCSU ERIS digital archive system," *Code4Lib Journal*, vol. 12, 2010.
- [7] C. Post, A. Chassanoff, C. Lee, A. Rablin, Y. Zhang, K.. Skinner and S. Meister, "Digital curation at work: modeling workflows for digital archival materials," *ACM/IEEE Joint Conference on Digital Libraries (JCDL)* 2019. Available: ACM Digital Library, [Accessed 7 March 2022].
- [8] A. Yoon, A. Murillo, and P.A. McNally, "Digital preservation in LIS education: a content analysis of course syllabi," *Journal of Education for Library and Information Science*, v. 62, no. 1, pp. 61-86, 2021.
- [9] "Digital preservation tools and systems," [no date]. Available: <https://www.nationalarchives.gov.uk/archives-sector/advice-and-guidance/managing-your-collection/preserving-digital-collections/digital-preservation-tools-systems>. [Accessed 7 March 2022].
- [10] Digital Preservation Coalition, *Digital preservation coalition handbook*, 2nd Ed. Available: <https://www.dpconline.org/handbook>. [Accessed: 7 March 2022].
- [11] O. Rieger, "The state of digital preservation in 2018: a snapshot of challenges and gaps," *Ithaka S+R*, [Online], 2018.
- [12] A. Chassanoff, and C. Post, *OSSArcFlow guide to documenting born digital archival workflows*, Available: <https://educopia.org/ossarcflow/>, 2020. [Online]. [Accessed 7 March 2022].
- [13] Community owned digital preservation tool registry (COPTR). [no date]. Available: https://coptr.digipres.org/index.php/Main_Page.
- [14] "Toolbox for digital heritage" [no date]. Available: <https://erfgoedkit.nl/>. [Accessed: 7 March 2022].
- [15] A. Cushing and K. Shankar, "Digital curation on a small island," *Archives and Records* vol. 40, no. 2, pp. 146-163, 2019.
- [16] J. Kim, E. Warga, and W. Moen. "Competencies required for digital curation: an analysis of job advertisements." *International Journal of Digital Curation* vol. 8, 1 [Online]. Available: <http://www.ijdc.net/article/view/8.1.66>, 2013. [Accessed 7 March 2022].
- [17] R. Marciano, "Toward a computational framework for library and archival education: report on a preliminary literature and curriculum review" [1 April 2019]. [Online]. Available: <https://ai-collaboratory.net/wp-content/uploads/2020/04/imlsCASresearchGroupReport.pdf>. [Accessed: 7 March 2022].

APPRAISAL AND SELECTION ON A LONG-TERM PRESERVATION REPOSITORY?

Can you repeat that, please?

Luís Faria

KEEP SOLUTIONS
Portugal
lfaria@keep.pt

Miguel Guimarães

KEEP SOLUTIONS
Portugal
mguimaraes@keep.pt

Miguel Ferreira

KEEP SOLUTIONS
Portugal
mferreira@keep.pt

Abstract – In this article we describe the way MoReq 2010 has been used to guide the development of appraisal and selection workflows in an open-source long-term digital repository system. Even though the destruction of records is an irreversible process that is deemed contrary to the mission of a long-term digital preservation repository, we argue that there are enough real-world use case scenarios that justify the need to bring these features to digital repositories as well further demystifying the idea that digital archives are places where records come to die, and instead reinforce the notion that these can be used as vessels to revitalize information and support operational systems in the day-to-day business operations.

Keywords – Appraisal, Selection, Preservation, Repository, RODA.

Conference Topics – Innovation.

I. INTRODUCTION

Long-term digital preservation repositories have long been viewed as the final destination for digital objects (or records) that are at the end of their lifecycle (also known as “inactive records” in certain spheres) and are defined for permanent retention. But this concept, inherited from paper-based records management, is found to be inadequate for electronically-generated records [1], where the mutability, versatility and complexity of the records (data and metadata) and their dependency on intermediate systems (software and hardware) blur the division between active and inactive records and the requirements of the activities necessary to preserve them for the time-span they are defined to be retained.

This new reality also affects how and when records are transferred from production systems (i.e. the system where the records were originally

created) to archival systems and blurs the division between both systems, which ends up causing an overlap between the requirements of both systems and the information that is maintained by each of these systems in any given time. One of the issues found in this context is related to the moment in the information life-cycle when a record should entail additional considerations regarding digital preservation, (defined in form of a preservation plan, which might require activities to be performed that may not be supported by the production system) and the moment after which a record could, should or must be eliminated from the original system where it was produced. Some institutions, such as Portuguese National Archives, determine that as a general rule-of-thumb additional digital preservation activities should be done to records that are older than 7-years. Although the adequateness of such a general rule is debatable, it is easy to find situations where the retention period of those records ends after the record has been transferred to the long-term preservation archive. Transfer of electronic records is also very different from paper-based records, as records can be maintained in both systems, or partially destroyed or completely eliminated from the production system. But, independently of how transfer is performed, the archival system needs to support the records disposal and retention workflows.

This was the challenge encountered by the Swedish Customs (Tullverket), i.e., to add disposal

and retention features to RODA¹, an open-source long-term digital preservation repository used by archives and other large institutions to safekeep digital records and auxiliate in the implementation of digital preservation plans.

RODA is a digital repository solution that delivers functionality for all the main units of the OAIS reference model. RODA is capable of ingesting, managing and providing access to the various types of digital objects produced by large corporations or public bodies. RODA is based on open-source technologies and is supported by existing standards such as the Open Archival Information System (OAIS), Metadata Encoding and Transmission Standard (METS), Encoded Archival Description (EAD), Dublin Core (DC) and PREMIS (Preservation Metadata).

This paper documents the standards chosen to guide the requirements for the development of disposal and retention workflows, the additional requirements brought by case-studies of archives in different countries, and the technical details of how these features were included in RODA version 4 released in March 2021.

II. SELECTION AND APPRAISAL, RETENTION AND DESTRUCTION

Traditionally, in an archival context, appraisal is the process of determining if records and other materials have permanent (archival) value, i.e., which are to be kept for specified periods of time and which are to be destroyed. Only the records selected for permanent retention are expected to be transferred to a more permanent archival facility. This decision process typically takes place where the records were originally created (e.g. in the production system).

The evaluation process determines based on legal requirements and on current and potential usefulness, which records should be retained and for how long. This process is in itself a multiple decision-making process inside whatever structure is relevant, that being a small group, a larger department, a whole organization or the more traditional archival approach involving retention rules that go beyond organizational boundaries.

Currently, records may be selected for transfer to an archival system even if they are not for permanent retention. This happens when the retention period is longer than the “preservation period” of the production system. The “preservation

period” refers to the amount of time after which a record should be subject to preservation actions and other assessment activities, defined in the preservation plan of the institution and that are not feasible to be done in the system that currently holds the records.

III. APPRAISAL STRATEGIES

Appraisal and the definition of the disposal schedule and retention period of records, may be done at many levels, such as fonds or collections, series, files, or even at item level [2]. Studied use cases from the Swedish Customs and the Portuguese National Archives showed two different strategies to define intra- and inter-institutional disposal schedules and retention periods for records.

The Swedish Customs define their disposal schedules and retention periods using the archival hierarchy, which is based on ISAD(g), structuring their records by their retention periods. In this use case, records are organized in series in the archival hierarchy, which determine the retention period for all records within the context of that series (i.e. by inheritance).

The Portuguese National Archives, on the other hand, define the disposal schedules and retention periods using a centralized hierarchical classification system that is process-oriented. Records are classified using classes from a common taxonomy. Records can be classified using multiple classes. National “classification schemes” such as these define the possible record classes and set the relationship between the classes and the records disposal action and retention period [3]. This allows the definition of a disposal classification hierarchy that is orthogonal to the local archival classification hierarchy, as they may have incompatible purposes and objectives.

IV. DISPOSAL SCHEDULE

Taking into account the requirements and approaches from both the use cases analyzed, the most promising standard that establishes a disposal and retention process was MoReq 2010:

A record, once it has been created in a MoReq Compatible Record System (MCRS), can never be deleted in full, as if it had never existed. This concept of accountability is important to good records management: although the complete record and its content no longer exist, there remains a residual record

¹ <https://www.roda-community.org>

to show that it was once held by the MCRS. The residual record, which remains with the MCRS for the life of the system, proves not only that a record was once active but also, and possibly more importantly, that the record was properly disposed of under an appropriate disposal schedule.

It is the record's disposal schedule that determines how long a record is retained and how it is subsequently disposed of at the end of its retention period. [4]

Disposal schedules are categorized by the following attributes:

Table I

Disposal schedule attributes and description categorization

Field	Description	Mandatory
Title	The identifying name or title of the disposal schedule	Yes
Description	Description of the disposal schedule	No
Mandate	Textual reference to a legal or other type of instrument that provides the authority for a disposal schedule	No
Scope Notes	Guidance to authorized users indicating how best to apply a particular entity and stating any organizational policies or constraints on its use	No
Disposal Action	Code describing the action to be taken on disposal of the record (Possible values: Retain permanently, Review, Destroy)	Yes
Retention Trigger Element Identifier	The descriptive metadata field used to calculate the retention period	Yes (if Disposal Action is different from Retain permanently)
Retention Period	Number of days, weeks, months or years specified for retaining a record after the retention period is triggered	Yes (if Disposal Action is different from Retain permanently)

The MoReq 2010 standard was developed for Electronic Record Management Systems, which is not an exact fit for a long-term digital preservation repository. Not all disposal actions defined in the MoReq2010 are supported. Transfer workflow was found to be outside of scope for this development project because, for the analyzed cases (Swedish Customs and Portuguese National Archives) the archive is still currently seen as the final stop of the record holding-systems journey. Therefore, it was decided that RODA would only support three types of disposal actions: Retain permanently, review, and destroy. Also, due to budget constraints and a limited foreseen use by the sponsoring institution, the review lifecycle is not fully supported to the extent that is defined in the MoReq2010 standard.

Records marked to be retained permanently do not define a retention period. Records marked for review or destroy actions have an associated retention period which needs to be configured during the disposal schedule creation. A descriptive metadata field of type "date" is used as an input to calculate the retention period, based on the associated disposal schedule, which may have different granularities: days, weeks, months, or years. There is also an option to have no retention period meaning the record is ready to continue the review or destruction life-cycle depending on the disposal action.

Records with disposal schedules associated with them will have certain operations disabled, even by users with administration permissions, such as "Remove record". In order to delete records that are under a disposal schedule they need to be disassociated first and then deleted.

The list of the disposal schedules available for records in the repository is available in the "Disposal policies" top-page in RODA (Fig. 1). A disposal schedule can be associated with a record manually or automatically (more details below). Once a record is associated with a disposal schedule, the disposal schedule can no longer be completely eliminated from the system. Instead, it becomes disabled, serving as evidence that it was once associated with a record and that it may have affected its disposal.

RODA offers at the disposal schedule level three different roles. A role to list and view disposal schedule information, a role to manage disposal schedule information and a role to associate or disassociate disposal schedule from records.

Disposal policies

In this page you can consult the different disposal policies that are associated with this repository. Information about the disposal schedules, disposal holds and disposal rules created for the purpose of manage the life cycles of intellectual entities.

Disposal schedules

Disposal schedules set the minimum requirements for the maintenance, retention or destruction actions to be taken in the existing or future intellectual entities in this repository. A intellectual entity may only be destroyed as part of a disposal process governed by the disposal schedule assigned to that entity. It is the intellectual entity's disposal schedule that determines how long a record is retained and how it is subsequently disposed of at the end of its retention period.

Title	Mandate	Period	Action	State
Disposal schedule inactive			Retain permanently	Inactive
Disposal schedule 1		1 year	Destroy	Active

Disposal rules

Disposal rules are a set of requirements that determine the disposal schedule for each intellectual entity in this repository. The disposal rules can be applied at any time in order to maintain the repository consistency. Disposal rules can also be applied during the ingest process. Disposal rules have a priority property in which they are executed. If a record is not covered by any of the rules, it will not be associated to a disposal schedule.

#	Title	Selection method	Condition	Schedule
1	Rule 1	Child of	Specifications	Disposal schedule 1
2	Rule 2	Metadata field	title is test	Disposal schedule 1

Disposal holds

Disposal holds are legal or other administrative orders that interrupts the normal disposal process and prevents the destruction of an intellectual entity while the disposal hold is in place. Where the disposal hold is associated with an individual record, it prevents the destruction of that record while the disposal hold remains active. Once the disposal hold is lifted, the record disposal process continues.

Title	Mandate	State
Disposal hold inactive		Lifted
Disposal hold 1	This is an example of a mandate	Active

Figure 1 RODA: Disposal policies.

V. DISPOSAL ACTIONS

While all disposal schedules must conform to the MoReq2010 disposal process, they may specify different behaviors [4]. This behavior is defined by a different disposal action, which must specify one of four different possible outcomes:

- Retain permanently;
- Review at the end of the retention period;
- Transfer at the end of the retention period;
- Destroy at the end of the retention period.

As explained before, the transfer disposal action is not to be supported in this development iteration. The other disposal actions are detailed below, further detailing technical details of how RODA implements them.

A. Retain Permanently

An important aspect of records management is the preservation of important records for very long periods of time, including the ability to designate some records that are never to be discarded. This is done by applying a disposal schedule with a retention trigger that specifies permanent retention. [4]

Although RODA's main drive is long-term preservation of records elected for permanent conservation, additional controls were added to ensure that records with a "retain permanently" disposal schedule may not be deleted by operations, even by users with administration roles. To perform such operations, a user must first change or remove the disposal schedule associated with the record.

B. Review (Partial Support)

There are some occasions when the importance of a record and the length of time it should be retained are not known at the time the record is created, and cannot be calculated simply from subsequent events (such as transfer to the archive). It may also be that, in some jurisdictions, the retention period is so long that it is felt that the guidance for their retention may change in the intervening period. Under these circumstances, where there is reasonable doubt about their final destiny, records can be scheduled for later review, rather than for permanent retention, transfer or destruction.

The outcome of the review must include the application of a disposal schedule to the record based on the review decision. The new disposal record will replace the previous schedule associated with the record and will then specify the ultimate fate of the record, or it may be used to schedule another late review, or to retain the record permanently. [4]

The acceptable review periods are defined in the list of disposal schedules, which allow to define a policy on how reviews are done and for how long the definition of the disposal action can be postponed for a record.

MoReq further defines that the disposal schedule can set strict limits for how long a review confirmation can take. In RODA, records can be marked to be reviewed after a retention period, but the workflow for review confirmation is not supported. Records simply must be assigned with a new disposal schedule to get out of the review list, be it a destruction, a retain permanently schedule, or a (predefined) review schedule with a larger retention period.

RODA provides a dashboard where one can search through the list of records that are overdue for review, allowing users to inspect them and to set, for each or in batch, a new disposal schedule to retain permanently, destroy or review with a different retention period.

C. Destroy

When an active record is destroyed, its metadata and event history are pruned and its content is deleted. The remaining metadata of the record, along with their remaining event history, make up the residual record [4], which serves as evidence that the record once existed.

Pruning is an important process in ensuring the proper destruction of the content of records, especially in sensitive environments where these events and metadata may reveal information about the original content of the record and may be able to be used to partially (or fully) reconstruct the destroyed content. [4]

In RODA, pruning of metadata is configured by an XSLT per supported metadata schema (e.g. EAD, Dublin Core, etc.). This allows the customization of how metadata is pruned for each type of record metadata schema. This is a static configuration for all records that must be in place before destroying the records, but it can iteratively be improved and extended via configuration, although it cannot affect existing residual records.

RODA presents a list of records that are overdue for destruction, as shown in Fig. 2, which can be searched and filtered in several ways. A disposal confirmation can be initiated by selecting from records that are overdue the ones the user wishes to destroy. Then, a formal destruction workflow takes place, where a printable report is finally produced.

Create disposal confirmation

In this page you can see the intellectual entities that are overdue for destruction or review. If the records are ready for destruction it is possible to integrate them into a new disposal confirmation or change their disposal schedule. If the records are ready for review, you can only change their disposal schedule. It is also possible to view the records which their retention period calculation failed.

Buttons: Show records to destroy, Show records to review, Retention period errors

Search... [dropdown] [search icon] [info icon]

<input type="checkbox"/> Level	Title	Disposal schedule	Retention start date	Overdue on
<input type="checkbox"/> Fonds	test	Disposal schedule 1	2016-03-31	2017-03-31

EXPORT 1-1 of 1

Figure 2 RODA: Create disposal confirmation.

When the records destruction finally happens, the descriptive metadata will be pruned and all its associated files will be removed, leaving the record in a RESIDUAL state.

VI. DISPOSAL RULES

Disposal rules are a set of requirements that determine the disposal schedule for each record in the repository via a selection method.

This is an extension to the MoReq standard that allows users to select a batch of records and associate a disposal schedule to them. This functionality offers a powerful mechanism to automate the definition of retention periods for several records at once. During the ingest workflow it is possible to associate a disposal schedule to the record being ingested simply by defining a set of rules.

Table II
Disposal rule attributes and description categorization

Field	Description	Mandatory
Order	Order by which the rules will be applied to records	Yes
Title	The identifying name or title of the disposal rule	Yes
Description	Description of the disposal rule	No
Schedule	Disposal schedule that will be associated with the record	Yes
Selection Method	Condition that will trigger the disposal rule.	Yes

The disposal rules must have an order, a title, a disposal schedule and a selection method. The description attribute is optional. The order attribute allows for rules to be prioritized. The title identifies the name of the disposal rule. The description gives more information about the rule itself. The schedule attribute refers to the disposal schedule that will be associated with the record. The selection method is the condition that will trigger the disposal rule (see Table II).

Regarding the selection method there are two possible values that are currently supported: "child

of", or "metadata field". "Child of" means that all descendants of the selected record in its hierarchical organization within the catalog will have the disposal schedule associated. "Metadata field" method is related to the record's own descriptive metadata. Currently it only supports exact matches. The metadata fields can be configured and therefore tailored to comply with the institution's retention policy.

Disposal rules can be applied during the ingest workflow via a plugin that acts as an additional ingest step, or to the whole repository when run manually by a user. When applying to the whole repository there are two options available: 1) override previous disposal schedule associations; or 2) preserve disposal schedules that have been manually associated to a record while overriding automatic associations.

RODA ships with two user roles that are dedicated to the management of disposal rules. A role aims to list and view disposal rules information and a role to manage disposal rules information.

VII. DISPOSAL HOLDS

Disposal holds are legal or other administrative orders that interrupt the normal disposal process and prevent the destruction of an intellectual entity while the disposal hold is in place. [4]

When lifting a disposal hold all intellectual entities that were on hold can resume the normal disposal process. After the lift the disposal hold remains as an historical reference and it cannot be reused. Disposal holds can only be deleted if they were never associated with an intellectual entity.

When a record is being held by a disposal hold, even if it is not associated with a disposal schedule, RODA disallows certain operations from being performed, even to users with administration permissions. If a record is associated with a disposal hold, operations such as remove record, move record (in the archival hierarchy); create, edit or delete descendants; create, edit or delete representations; and edit descriptive metadata for this record or any of its descendants are disabled. These constraints are applied to keep the record and its descendants safe and unaltered until the legal or administrative mandate that caused the disposal hold is lifted.

Disposal holds are categorized by the following attributes:

Table III
Disposal hold attributes and description categorization

Field	Description	Mandatory
Title	The name or title of the disposal hold	Yes
Description	Description of the disposal hold	No
Mandate	Textual reference to a legal or other instrument that provides the authority for the disposal hold	No
Scope Notes	Guidance to authorized users indicating how best to apply the hold and stating any organizational policies or constraints that may affect its application	No

RODA offers three user roles to manage disposal holds. A role that is able to list and view disposal holds information, a role to manage disposal holds information and a role to associate or disassociate disposal holds from records.

VIII. DISPOSAL CONFIRMATION

In MoReq 2010, disposal (or destruction) confirmation is the period up to when the destruction of the record is to be executed, but the standard does not provide any more guidance on how this process should be done. Based on the analyzed use cases, specially from the Portuguese National Archives, we were able to verify that the actual destruction of records might need to be approved by a managerial authority in the organization or even a third-party outside the organization, for example, the Portuguese National Archives is required to explicit consent for the deletion of records in any Portuguese government agency.

To allow the support for this use case, a formal disposal confirmation workflow was added to RODA. This enables an administrative (signed) confirmation to be pursued prior to destruction of records. This confirmation is done by producing a report that aggregates all necessary metadata from the records to be destroyed in printable format (paper or PDF). This report should be formally accepted by the respective authoritative body, after which the destruction must be explicitly requested by the operator (which might not be the same person).

Disposal confirmations

In this page you can consult the disposal confirmations that were created for this repository. A disposal confirmation consists of a report that aggregates the intellectual entities and extra metadata information. In order to destroy the intellectual entities associated to a disposal confirmation you need to explicit execute the destroy action. After destroyed the intellectual entities within the disposal confirmation can be restored or permanently deleted.

Search...					
Title	Creation date	Creator	Status	# AIP	Storage size
disposal confirmation	2022-03-08 09:49:52	admin	Approved	10	125 GB
disposal confirmation 1	2022-03-08 12:02:15	admin	Pending	6	9.9 MB

Figure 3 RODA: Disposal confirmations.

RODA presents a dashboard (Fig. 3) which displays pending, approved, restored or permanently deleted disposal confirmations.

The list contains metadata about the number of records affected by the disposal confirmation ("# AIP" column) and storage size that was or will be reclaimed by the permanent destruction of the records affected by the ruling.

Disposal report nrº 273c60cb-91fc-4f72-87d8-b0fe47da9a0d

This disposal agreement created by admin on 2022-03-08 13:28:00, was approved on 2022-03-08 on the date defined below, allowing the destruction of 3 records.

List of records to be destroyed

Level	Name	Fonds/Collection	Nr Files	Size	Schedule id	Holds id
Fonds	Test		10	5 GB	backdf60-2976-4331-b27a-26cfa79f849	
Fonds	Test		5	3.2 GB	backdf60-2976-4331-b27a-26cfa79f849	
Fonds	Test		4	40 MB	backdf60-2976-4331-b27a-26cfa79f849	

Figure 4 RODA: Disposal confirmation report.

Disposal confirmations with a pending state are still waiting for a confirmation to initiate the destruction process. After the destruction is confirmed, the batch of records identified in the disposal confirmation will be either restored or permanently destroyed (Fig. 4). Automatic permanent destruction after a period can also be configured. But after the initial destruction operation the records are already removed from the repository. This safety net feature is an extension to the MoReq2010 procedure and is further explained in the "Disposal Bin" section.

The disposal confirmation report can be customized to meet the institution branding requirements and informational needs. There are two levels of customization, one referring to the report displayed on RODA's interface and the other related to the report that will be printed-out. This customization is done via a templating system to tailor the report to the institution's branding and bureaucratic needs and procedures for destroying records.

Once a record is assigned to a disposal confirmation, the record itself and its descendants can no longer be associated with another disposal schedule or disposal hold. The operations of remove record, move record (in the archival hierarchy);

create, edit or delete descendants; create, edit or delete representations; and edit descriptive metadata for this record or any of its descendants are disallowed.

Digital Information LifeCycle Interoperability Standards Board

Assigned to a disposal confirmation

Disposal confirmation

Disposal confirmation

disposal confirmation 1

Creation date

2022-03-08

Status

Pending

Figure 5 RODA: Record assigned to a disposal confirmation

As depicted in Fig. 5, each record provides detailed information of the disposal schedules, holds and confirmations associated with it, and also presents the calculated retention period as well as the destruction operation details.

At the disposal confirmation level, RODA implements five user roles. A role to list and view disposal confirmation information, a role to manage disposal confirmation information, a role to destroy records according to the disposal confirmation, a role to restore destroyed records according to the disposal confirmation and a role to permanently delete destroyed records according to the disposal confirmation.

IX. DESTROYED RECORDS

Active records and residual records are logically separated since their meaning and use is completely different. To list and search through the destroyed records there is a special page, only available to authorized user roles. This page allows users to search through the pruned metadata and inspect the events and metadata of destruction operations, including the disposal schedule, disposal holds that affected the record retention, the parties involved in the disposal confirmation and the authorization for destruction. All this information is accessible via a single-entry point (Fig. 6).

Destroyed records

In this page you can see the intellectual entities that are in a destroyed state. The destroyed record, which remains for the life of the system, proves not only that a record was once active but also, and possibly more importantly, that the record was properly disposed by an appropriate disposal schedule.

Level	Title	Retention start date	Disposal schedule	Destroyed on	Destroyed by
Fonds	test	2016-03-31	Disposal schedule 1	2022-03-08	admin

Figure 6 RODA: destroyed records.

RODA also provides visual cues to better identify records that are destroyed and when they were destroyed (Fig. 7).

Destroyed

test

Destroyed on 2022-03-08

0 risk incidences, 14 preservation events and 80 log entries

Created by admin on 2022-03-08 and last updated by admin on 2022-03-08

Encoded Archival Description 2002

Identity

Reference code

1249678e-a631-4419-8046-0b9cb5e6a83d

Figure 7 RODA: Destroyed record with visual cues.

X. DISPOSAL BIN

Destruction of records is an irreversible process that is contrary to the main drivers of a long-term digital preservation repository. Due to that fact, an additional safety net feature was included to allow users to recover records that were improperly destroyed.

During the destruction of the records, a copy of each record is created in a logically separated storage, inaccessible for any RODA process except for the "restore" or "permanent destruction" actions. After destruction, a whole disposal confirmation can be either permanently deleted or restored to the previous state.

The restore process will recover all AIPs to their state previous to destruction and mark the disposal confirmation as restored. When this action is performed, the whole batch of records will be restored as the confirmation authorization (now revoked) was done for the entire set of records and not just a few of these. Records will be again overdue for destruction and the restored disposal confirmation can no longer be changed. A new disposal confirmation will need to be created to destroy all or part of the previously restored records.

Permanent deletion will remove the backup and make the destruction irreversible. The disposal confirmation will be marked as "deleted" and it can no longer be changed. The permanent destruction can also be set up to be automatic after a period of time, for example permanently deleting records after one month of the destruction being confirmed.

Restore and permanent deletion operations require special user roles to be executed.

XI. PRESERVATION METADATA

All disposal related activities over records are fully documented in preservation metadata using

PREMIS events². These events can be related to specific records (object-level events) or to the whole archive (repository-level events). The preservation metadata document the provenance of records and also their final destination, ensure all relevant actions made in the records are properly recorded and follow a well-defined procedure, supporting the case for the authenticity of the digital objects and their proper destruction.

Every disposal related operation creates a preservation event which can be listed and inspected in the Preservation Event page (Fig. 8). Preservation event types were selected based on the provided controlled vocabularies³. A summary of each preservation event per disposal operation can be consulted in Table IV.

Table IV
Preservation events created by disposal operations

Disposal operation	Preservation event type (and level)
Associate or disassociate a disposal schedule	Policy assignment (object-level)
Associate or disassociate a disposal holds	Policy assignment (object-level)
Lift a disposal hold	Policy assignment (object-level)
Assign or withdraw a record to a disposal confirmation	Update (object-level)
Destroy the record via disposal confirmation action	Destruction (object-level)
Restore record from disposal bin	Recovery (object-level)
Create a disposal confirmation report	Creation (repository-level)
Remove a disposal confirmation report	Deletion (repository-level)
Permanently delete records from disposal confirmation report	Deletion (repository-level)

Preservation events

A preservation event aggregates metadata about actions, specifically documenting which objects it affects and which human or software agents intervened. Documentation of actions that modify an object is critical to maintaining digital provenance, a key element of authenticity. Actions that create new relationships or alter existing relationships are important in explaining those relationships. Even actions that alter nothing, such as validity and integrity checks on objects, can be important to record for management purposes.

Date	Type	Detail	Outcome
2022-03-08 14:11:48	recovery	AIP restored from disposal bin	Success
2022-03-08 14:11:39	destruction	AIP destroyed by disposal confirmation	Success
2022-03-08 14:07:54	creation	The process of creating an object of the repository.	Success
2022-03-08 14:06:49	policy assignment	Associate disposal schedule to AIP	Success
2022-03-08 13:28:00	creation	Create disposal confirmation report	Success
2022-03-08 13:28:00	update	Disposal confirmation assign	Success
2022-03-08 12:02:15	update	Disposal confirmation assign	Success
2022-03-08 12:02:15	creation	Create disposal confirmation report	Success
2022-03-08 12:02:15	update	Disposal confirmation assign	Success
2022-03-08 12:02:15	update	Disposal confirmation assign	Success
2022-03-08 12:02:15	update	Disposal confirmation assign	Success
2022-03-08 12:02:15	update	Disposal confirmation assign	Success
2022-03-08 12:02:15	update	Disposal confirmation assign	Success
2022-03-08 12:00:33	policy assignment	Associate disposal schedule to AIP	Success
2022-03-08 12:00:03	policy assignment	Disassociate disposal hold from AIP	Success
2022-03-08 11:48:55	policy assignment	Apply disposal hold to AIP	Success
2022-03-08 11:47:44	policy assignment	Associate disposal schedule to AIP	Success
2022-03-08 10:37:43	deletion	Permanently delete records from disposal confirmation report	Success
2022-03-08 09:50:36	destruction	AIP destroyed by disposal confirmation	Success
2022-03-08 09:49:52	creation	Create disposal confirmation report	Success

Figure 8 RODA: Preservation events

XII. CONCLUSION

Long-term digital preservation repositories have been presented with new challenges as they are exposed to more complex use case scenarios and used by different types of institutions.

That is the case of the Swedish Customs, in which internal policies, and national and European legislation (e.g. GDPR⁴) bring data retention requirements to records that are in scope for long-term preservation activities.

The policies and workflows that govern the disposal of records are very different from country to country and institution to institution. In some situations, the destruction may be authorized by the same person operating the archive, in others the top-management of the institution must be involved in the authorization process, while in other cases destruction policies are centralized and applicable to multiple institutions, thus requiring the involvement of external parties in the authorization process.

Although strong guidance was provided by process definitions included in MoReq 2010, some extensions were required to ensure the use cases identified in the case studies were supported.

Furthermore, special attention was given to additional controls and recording of evidence, to ensure disposal procedures were correctly followed and to provide evidence that records were deleted following proper procedures. Preservation metadata is produced in every disposal-related operation to ensure these processes are documented.

² <https://www.loc.gov/standards/premis/>

³ <https://id.loc.gov/vocabulary/preservation/eventType.html>

⁴ GDPR: General Data Protection Regulation

The capacity to fit different jurisdictions, policies, disposal approaches and organizational structures was ensured by designing a customizable workflow that includes ability to tailor disposal schedules, how they are associated with records (manually or automatically), how the retention period is calculated, how record metadata schemes fit into the retention period calculation, how record metadata schemes are pruned upon destruction, what information is available in the disposal confirmation reports and their design, and which users are able to use each of the available operations.

A great importance is given to be clear and evident on which disposal policies are installed and effective and how they were set up in the past. A global view of the disposal policies in effect is available in the "Disposal Policies" dashboard, and every record provides clear information of which disposal policies affect them and what is the calculated retention period.

Although destruction seems to be antagonistic to the core objectives of a long-term digital preservation repository, it is a necessary process to comply with policies and legal requirements to which institutions are subject to. It then becomes essential to ensure that destruction is done following proper procedures and guarantee that no record was destroyed when it should not be.

Historically, records in (national) archives have been exempt from the legal requirements to destroy information as records were expected to be preserved forever, but long-term preservation is no longer limited to (national) archives and thus must expand their capabilities to support digital preservation in every type of institution.

The disposal features described in this paper have been released in March 2021 on RODA 4. These features are available on GitHub⁵ and can be inspected on the product demonstration site⁶.

ACKNOWLEDGMENT

We would like to thank the Swedish Customs by sponsoring the development of the disposal features presented in this paper, and ensuring that the resulting developments did not only fit their own use case but more general use cases from the user community. A special thanks to Magnus Wåhlberg for taking the time to study and re-design the Swedish Customs requirements to follow the

relevant standards and ensuring the result would serve not just his institution but also the entire world.

REFERENCES

- [1] Z. M. Yusof and R. W. Chell, "The Records Life Cycle: an inadequate concept for technology-generated records," *Inf. Dev.*, vol. 16, no. 3, pp. 135–141, Sep. 2000, doi: 10.1177/02666666004240413.
- [2] Internationaler Archivrat and International Council of Archives, Eds., *ISAD(G): General international standard archival description; adopted by the Committee on Descriptive Standards, Stockholm, Sweden, 19-22 Sept. 1999*, 2. Ed. Ottawa: International Council of Archives, 2000.
- [3] A. Lourenço, J. C. Ramalho, M. R. Gago, and P. Penteado, "Plataforma CLAV: contributo para a disponibilização de dados abertos da Administração Pública em Portugal," *Cad. BAD*, no. 2, Art. no. 2, 2019, doi: 10.48798/cadernosbad.2047.
- [4] DLM Forum Foundation, *Modular requirements for records systems : core services & plug-in modules (version 1.1), volume 1*. LU: Publications Office, 2011. Accessed: Mar. 07, 2022. [Online]. Available: <https://data.europa.eu/doi/10.2792/2045>

⁵ <https://github.com/keeps/roda>

⁶ <https://demo.roda-community.org/#welcome>

A GENERIC EMULATOR INTERFACE FOR DIGITAL PRESERVATION

Towards a Collaborative Distributed Emulator Registry

Rafael Gieschke

University of Freiburg
Germany
rafael.gieschke@rz.uni-freiburg.de
[0000-0002-2778-4218](https://orcid.org/0000-0002-2778-4218)

Klaus Rechert

University of Applied Sciences Kehl
Germany
rechert@hs-kehl.de
[0000-0002-2454-4374](https://orcid.org/0000-0002-2454-4374)

Abstract – Emulation frameworks as well as emulation as an access strategy have matured. With more simplified access to emulation and improved workflows, there is still a gap to be closed, primarily the availability of emulators and especially for smaller niches like arcade games or pre-PC computers. In this article we propose methods to simplify emulator preparation for framework integration as well as describing an emulator's technical capabilities. Both, the technical design and the technical description of emulators will provide a foundation for cooperative work to identify, list, describe and integrate emulators of interest for the digital preservation community.

Keywords – Emulation, Metadata, Registry
Conference Topics – Innovation, Exchange

I. INTRODUCTION

With maturing emulation frameworks as well as wider acceptance and usage of emulation as a strategy within the digital preservation community, the availability of emulators as a crucial precondition and a future risk factor has become a focus of attention. Implementing emulators is a highly technical and complex task which requires significant development resources as well as expertise. Even though considered in the past [1], implementing emulators within the preservation community is currently – and for the foreseeing future – not a realistic options. The IT research and development budgets of memory institutions are already strained due to a variety of challenges posed by the ongoing digital transformation of public administrations and businesses. Furthermore, it is not sufficient to implement just a single emulator but a wide range of platforms, computer systems and use-case are to be

supported to cope with the breadth of born digital content from different decades and types.

Fortunately, there is already a very lively emulator ecosystem outside of the preservation community. High quality emulators have been implemented for various purposes, e.g. by enthusiasts and hobby programmers for nostalgic reasons or with support of the (commercial) hardware and software industry to support development and testing. One of the main foundations and success factors of the bwFLA emulation framework [2] was to tap into that pool of emulators and make them accessible and usable for digital preservation workflows through a unified control interface. Continuously integrating further emulators for digital preservation is necessary to address a number of conceptual and technical challenges. Most obviously, increasing the variety of emulators is necessary to cover smaller niches from arcade machines, game consoles to the pre-IBM PC era of mainframes and early home computers such as the BBC Micro or the ZX Spectrum. While plenty of emulators are available as (open source) standalone software, the integration into emulation frameworks currently requires a so-called "wrapper" within the framework to adapt to the emulator's control interface, i.e., how the emulator is configured and started, to allow integration into generic preservation workflows, e.g., open and render a document in an emulated software setup or simply play a computer game without additional knowledge about the emulator specifics. This additional adaptation work remains a significant hurdle. The already quite limited development capacity has been focused on high volume and high demand workflows

and emulators. To improve this, broadening the developer base is of high importance, e.g., by simplifying the adaptation and integration tasks and decoupling the resulting emulator packages technically from the emulation framework, i.e., maintenance of the emulation framework and maintenance of emulator packages not only become independent tasks, but also preparation of emulator packages could be done without knowledge of internals of the emulation framework (e.g. EaaS).

A further problem regarding implementation, usage and maintenance of emulators is that there is currently no overview about emulators currently in use by the preservation community or what kind of emulators do exist and are of potential interest for the preservation community. Additionally, it is necessary to describe their technical capabilities and ideally verify these, e.g., based on relevant use cases. For instance, if there are multiple options for a given artifact, which emulator could be used or is recommended. Ideally, this information is not only shared within the preservation community but is also machine actionable such that tool support can be implemented to support search and automated workflows.

Finally, emulators are contemporary software and, like any other contemporary software, today's emulators will eventually become obsolete. Preparations are to be made for that event, first and foremost, by knowing if an emulator requires a more recent substitute, which emulated hardware components precisely this substitute should provide, if the substitute is then likely to be compatible or which one is the best match even if not 100% compatible.

In this paper, we have identified, defined and implemented a generic emulator API, such that – ideally – any emulation framework is able to integrate and reuse this work. Furthermore, for emulators to be described technically, we propose extensible metadata using semantic web technologies, to build a shared corpus of machine actionable metadata for emulators.

II. RELATED WORK

Emulators for digital preservation, e.g., Dioscuri[1] or the more abstract Universal Virtual Computer

(UVC)[3] have been proposed and developed with the idea to produce durable implementations of machines or machine code interpreters. In practice, this approach has failed due to the technical complexity of implementing emulators of rather complex computer systems. To achieve compatibility of software or operating systems, not only the architecture's CPU (supporting its instruction set architecture) needs to be implemented but also a wide range of additional hardware (and peripherals) like graphic cards, sound cards, input devices, and more. Due to the fast technical progress of computer platforms, not only the emulators need to be constantly adapted to a new technical ecosystem but there are also new concepts and components to be integrated.

Packaging freely available emulators to become usable as tool for preservation purposes and to some degree portable has been implemented within the KEEP Emulation Framework (KEEP EF) [4], [5]. Emulators became available through a simple "emulator archive". The packaged emulators, however, had been setup to run locally on the user's computer and thus, suffered from compatibility problems and reduced portability. Furthermore, the abstraction of the individual emulator control was rather limited. Abstraction of emulator controls, e.g., as a unified API is however a crucial precondition for the creation of complex workflows, automation, and interoperability. The rise of cloud computing led to a quite similar problem set. The lack of standards, e.g., every cloud provider using its own hypervisor (implementation or configuration) and, thus, hardware abstraction, made moving virtual machines from one provider to another difficult [6]. The DMTF Open Virtualization Format (OVF) defines and standardizes, amongst others, technical metadata describing the virtual machine's hardware configuration [7]. libvirt¹ is an example for a generic control API implementation for virtualization systems [8]. The main goal is to provide a stable interface for arbitrary hypervisors, implementing wrapper code (drivers) and an independent API front-end. We aim for a similar but more lightweight architecture abstracting specific control components (similar to the libvirt driver code) for disk, network, video, etc., components with the API endpoints being integrated within the emulator packages. For this, these approaches from the cloud context provide

¹ libvirt Virtualization API, <https://libvirt.org/>

valuable guidance on an abstract level, however, these projects are targeted to complex, well documented contemporary systems and quite difficult to extend to a technically diverse emulation landscape. Additionally, preservation workflows are less focused on performance aspects, i.e., require less fine granular controls but a higher degree of abstraction to cover a wider variety of hardware components.

For the purposes of framework integration, developing an emulator package reaches beyond a container-based portable software setup (e.g. Docker image, a Snap package or similar). Technically, the emulator package must not only be self-contained but also be self-describing, such that emulators can be found and selected based on their technical capabilities, i.e., the specific emulated hardware that is available and the potential hardware-related configuration options. With the KEEP metadata proposal [9] and later TOTEM registry and metadata schema [10] an initial tool-set for describing hardware and software became available. The proposed schema modeled hardware dependencies in a very detailed way but the schema has not yet found wider adoption yet mostly due to its static design. Entity fields were descriptive strings or user provided identifiers. This made an automated, tool-based generation, refinement, search and comparison of entities difficult. Digital preservation and in particular emulation-based workflows need to be highly automated to scale with breadth and amount of digital objects. There are now tools and concepts to identify file formats [11], software dependencies [12] and suggest software setups [13], but there is little to no support to identify the relevant technical platform, e.g., based on binary code analysis or operating system information, which technical platform (e.g. x86 PC, Apple PowerPC, Macintosh m68k, etc.) is required. The main reason so far is that there is no registry yet with documented available emulator hardware as well as documented connections of these hardware components to software, such as operating systems, libraries, drivers, or other digital artifacts.

With PREMIS v3.0 [14] the concept of environments has been introduced [15]. While environments can be described now in a rather flexible way, e.g. as an *Intellectual Entity*, the focus remains on descriptive elements, since hardware is considered as a physical object. Therefore, it is further necessary to resolve

the ambiguity of emulators representing one or many abstract physical concepts as well as being a software object with versatile configuration options and potentially multiple hardware platforms. Relating for instance a software stack (operating system, rendering software) and an digital object to an abstract hardware object additionally requires the relation to an emulator and more specifically information which of the emulator's configured hardware features are required. Describing the required hardware configuration and especially a "representation" in the form of a configured emulator remains difficult.

III. DESCRIBING EMULATORS

In order to list, maintain and (re-)use emulators for digital preservation purposes, technical description of their capabilities are required. A practical and rather general description of an emulator might be sufficient to describe the fact that an emulator emulates one or multiple guest platforms, e.g., computer systems or a combination of computer and operating system. These emulated systems are composed of multiple (default) hardware components with additional configurable or optional hardware elements. For instance, a typical computer setup always contains a specific CPU model but the CPU model is not sufficient to define or describe a specific computer system. A specific CPU might be used by multiple platforms, e.g., the MOS Technology 6502 CPU has been used in such diverse platforms as, among many more, the Apple II, the Commodore VIC-20, as well as the original Nintendo Entertainment System. A description at a computer system or platform level implies a specific set of components, e.g., a game for the Commodore VIC-20 will expect both the platform's MOS Technology 6502 CPU as well as its VIC graphics processor to be present and will not run on a Apple II using the same CPU.

For practical purposes, configurable or optional hardware components are of greater interest, e.g., the presence of a cassette/tape drive or a floppy drive of the VIC-20. A game might run from one or the other. This is even more apparent when using PC sound cards like the Sound Blaster 16 or the AdLib Music Synthesizer Card as games might very well only be compatible with one or the other. For an emulation setup, it is thus far more important – and easier, since these options are explicitly exposed by

the emulator software as configurable elements – to describe configurable or optional components than to describe implicit or default components of a given platform. Even if in some cases an optional component might be required, e.g., an emulated PC will generally have to have at least one (optional) storage device to be able to load any software, it can still be regarded as optional in the sense of not being implied by the emulated platform.

We therefore propose technical emulator metadata able to describe the emulator's supported platforms (e.g., QEMU is able to emulate the IBM PC platform, certain types of Apple Power Mac systems, and many more) and for each of these platforms the configurable components the emulator is able to emulate. The platform's detailed description, i.e., all required hardware components, and thus, elements an emulator has to implement to support a given platform, can be outsourced to public knowledge bases like Wikidata, DBpedia, or similar. Configurable and optional hardware components can be of different types. While emulated platforms differ greatly from each other, all of them share similar concepts like storage devices, input devices, video devices, audio devices, other output devices and possibly network devices. Some of these devices will have associated properties, e.g., a storage device has an inserted medium whereas an (Ethernet) network card has a MAC address.

A. Describing Hardware Components

As a first step towards a machine actionable emulator description, the (optional) components of an emulated platform have to be assigned an identifier so that they can be referenced from and included into emulation environments. While from a technical perspective, it would be enough to enumerate "component 1", "component 2", etc., additional descriptive and technical metadata about the components is crucial not only for users to be able to make an informed choice about which optional components are necessary in an emulation environment but also for automated or assisted selection of a suitable emulator, comparing emulators, finding substitutes for emulates and others.

For some hardware components, the community already has collected machine-readable linked open data describing their properties, e.g., the NE1000 network card², which is a member (a subclass) of the more general concept "network card"³ and of which Wikidata provides an image. Hence, it might be tempting to describe hardware emulated by the emulator directly using, e.g., entity URLs from Wikidata. However, this would not really be accurate as, e.g., QEMU does not provide a NE1000 network card directly but a specific implementation thereof. Other emulators might implement the NE1000 network card slightly differently such that there are cases in which it is not compatible with the QEMU implementation, e.g., QEMU might emulate a NE1000-compatible card connected to the ISA bus, while other emulators might connect it to the PCI bus. Furthermore, it might later be discovered that the emulated network card is not really a NE1000-compatible but actually a NE2000-compatible card.

Keeping existing environment metadata stable in this case while adding additional knowledge makes it necessary to describe the emulation environments using a two-step approach: The only guaranteed (and verified) information at creation time of emulation environments is the exact chosen configuration of the emulator package. While the developer of an emulator package might not understand initially all consequences of options passed to the emulator will have to the final hardware configuration, e.g., will the QEMU option `-hda` create an IDE or AHCI drive, will its storage controller be attached via PCI, the developer can assign a (unique) identifier to every device type corresponding to a configuration option of the emulator it supports. It is only important at this step to assign (at least one) unique identifier to every component the developer of the emulator packages wishes to expose. Identifiers should then stay stable over future versions of the same emulator package, i.e., the same identifier should reference the same emulated component. We expect this condition to be feasible as the source code of the emulator package is available and only updated for new emulator versions. For emulators themselves, developers or maintainers of emulator packages might have to scan the changelog or release notes for changes regarding, e.g., default options. We expect that this

² Novel Network Card NE1000,
<http://www.wikidata.org/entity/Q1959993>

³ "Network Card" Wikidata entry,
<http://www.wikidata.org/entity/Q165233>

can be supported by, potentially automatically, trying to run existing emulation environments with the new version of the emulator package. For example, if the QEMU option `-hda` were to change from emulating an IDE drive to emulating an AHCI drive, existing emulation environments containing Microsoft Windows would stop to run completely, which could easily be spotted. The existing component with its existing identifier would then have to pass further command-line options to QEMU and a new identifier could be introduced for the new default behavior. Linking these identifiers to, e.g., real-world devices can then be done as a second step, possibly at a later time, without modifying existing environments. In the same way, wrong assumptions about emulated devices (does QEMU really emulate a 80486 CPU for the IBM PC platform by default) can be corrected retroactively.

B. Example

Figure 1 shows the possible description of the QEMU emulator in an emulator package. It is able to emulate both the IBM PC as well as the Power Mac platform. For the IBM PC platform, the emulator package can emulate a hard disk drive and a network interface card. For the Power Mac platform, it can only emulate a hard disk drive. In both cases, this does not necessarily mean that QEMU itself is not able to emulate any other devices or platforms but that the emulator package only exposes (yet) the described devices using the generic API. The metadata is further enhanced with knowledge not directly used by the proposed API, e.g., a human-readable title for the platforms and devices, the information that the emulated hard disk drive uses (Wikidata property P2283) the PATA interface (Wikidata entity Q230360) in case of the emulated IBM PC and the SCSI interface (Q220868) in case of the Power Mac, or the fact that QEMU emulates a PowerPC G4 CPU (Q430856) as implicit (non-optional) component for the platform by default. The non-optional, implicit information is not necessary for starting and interacting (technically) with the emulator but might be very useful to determine if an emulation environment using this emulator package could potentially also run on another emulator, i.e., an emulator also emulating a PowerMac with a PowerPC G4 CPU and a SCSI hard disk drive.

```
{
  "dc:title": "QEMU 2.0.0",
  "emulator":
    "http://www.wikidata.org/entity/Q624699",
  "supportedMachines": [
    {
      "@id": "#x86_64",
      "@type": "eaas:machine",
      "dc:title": "IBM PC (x86_64/AMD64)",
      "supportedHardwareComponents": [
        {
          "@id": "#x86_64-disk",
          "@type": "eaas:drive",
          "medium": "eaas:disk",
          "dc:title": "Default Harddrive",
          "wdt:P2283":
            "http://www.wikidata.org/entity/Q230360"
        },
        {
          "@id": "#x86_64-nic",
          "@type": "eaas:nic",
          "emulates":
            "http://www.wikidata.org/entity/Q502465"
        }
      ]
    }, {
      "@id": "#ppc",
      "@type": "eaas:machine",
      "dc:title": "Power Mac (PowerPC, 32 bit)",
      "emulates":
        "http://www.wikidata.org/entity/Q209860",
      "defaultHardwareComponents": [
        {
          "emulates":
            "http://www.wikidata.org/entity/Q430856"
        }
      ]
    }
  ]
}
```

Figure 1 Emulator description

Of equal importance, is that the metadata could be used in much broader applications, e.g., querying the list of all emulator supported computer systems produced by Apple or, the other way round, determining if any relevant computer systems or optional components are still missing from being supported. It is important to stress that, as with all semantic data, providing the information in a machine-readable format facilitates all kinds of potential uses of the data, many of which were not yet anticipated. Allowing to enhance the emulator description with custom metadata is made possible by use of JSON-LD as an established metadata format as well as assigning (referenceable) identifiers (i.e., URIs in the form of URLs) to each distinct component an emulator package supports.

C. Emulator Configuration

While the emulator metadata describes which components the emulator is able to emulate, an emulator configuration describes which components have been selected and should be emulated to represent and instantiate a specific *environment*. We re-use the term environment with similar semantics as in PREMIS to describe fully configured computer setup. For example, a user might want to combine a QEMU IBM PC configured with a Sound Blaster 16 and a NE2000 network with a hard disk drive containing a software setup including Windows 95 as its operating system together with additional software installed as well as an (emulated) optical drive with an ISO image attached. The configuration can only reference concrete platforms and components defined in the respective emulator description and, generally, not arbitrary emulator options, intentionally reducing the number and granularity of supported options. This two-step approach allows to both decouple the emulator configuration from the emulator and, thus, improve its maintainability but also to describe emulation environments in a generic format, allowing the emulation framework to concentrate on this generic format instead of having to include specific knowledge about each emulator.

Figure 2 shows a further example of a configured QEMU emulator emulating an Apple Power Mac with a hard disk drive and a CD-ROM drive. The chosen JSON-LD format again allows the surrounding emulation framework to record additional information about the configured components, e.g., the data source of the hard disk drive. For the CD-ROM drive, no data source is present and it could be assumed to only be present without an ISO image attached. The "index" property is proposed as a generic way to signal in which order an emulator package should add the respective emulated devices to the emulated computer. This order might have an effect, e.g., on the Microsoft Windows operating system family for the assignment of "drive letter" to hard disk and optical drives. It is hoped to be more versatile than, e.g., inventing platform- and interface-specific ways to assign drives to individual ports as these ways would have to be developed for each emulator individually and most probably not be interoperable anyway.

The properties "path", "nativeConfig", and the "frameworkComponents" are technical properties that form part of the interface between the emulation framework and the emulator package and are described in more detail in the next section. "path" allows to provide an associated path, e.g., the path at which the emulation framework will provide the emulator the data, e.g., a disk image. "frameworkComponents" are components that are technically required to access the output of the emulator (i.e., the emulation software) but are not emulated components of the target emulation environment. "nativeConfig" is a simplification for users to provide (non-semantic) unstructured additional configuration options to an emulator package in order to support experiments with the emulator package. When using any "nativeConfig" properties, the user, however, cannot expect the emulation environment to be compatible with any future versions of the emulator package or be interoperable with any other emulators.

```
{
  "machine": "https://purl.org/emulation-
archive/qemu#ppc",
  "hardwareComponents": [
    {
      "index": 1,
      "component": "https://purl.org/emulation-
archive/qemu#ppc-harddisk",
      "path": "/mnt/disks/disk1",
      "binding": "urn:uuid:32af5a69-6033-485e-
a881-e36ee0d67cc5"
    },
    {
      "index": 2,
      "@id": "#cdrom-1",
      "component": "https://purl.org/emulation-
archive/qemu#ppc-cdrom"
    }
  ],
  "frameworkComponents": [
```

Figure 2 Emulator configuration

To simplify configuration work, emulator configuration templates, i.e., preconfigured and tested combinations of concrete components for a given emulator and platform are useful. These are identical to emulator configurations but, e.g., in case of hard disk drive, leave out configured data sources. They can support users in choosing a sensible start configuration, e.g., for a given target operating system, like including a Sound Blaster 16 and a

NE2000-compatible network card in an emulator configuration template targeting Windows 95. This can also be target operating-system specific, as, e.g., while the Linux's kernels default x86 32-bit configuration has support for NE2000-compatible network cards (a popular hardware from the early x86 era), its default x86(-64) 64-bit configuration only has support for Intel E1000-compatible network cards (a popular hardware from the x86-64 era). Templates could be developed by users independently of the emulator package and also provide the equivalent of default components for a given context, e.g., a default sound card for Windows 95. Templates further help to minimize the number of different configurations of a given emulator package and thereby easing future migrations to another emulator, where only very few configurations would have to be tested to have confidence that all emulation environments created from the same emulator configuration template remain working. On the contrary, it is much easier and more elaborate to collect such "default device" knowledge outside of the emulator package, e.g., also taking into account specific operating systems.

IV. ENCAPSULATING EMULATORS

It is not sufficient for emulator packages to be self-describing, they also need to be self-contained and self-executing, i.e., the package has to include all necessary parts to execute the emulator and be able to translate the generic API to the concrete API of the emulator so that emulation frameworks only need to speak one generic API.

To be able to preserve emulators and be able to run them independently of the host operating system, an emulator and all of its dependencies have to be encapsulated. The general goal is to have a very high forward compatibility: an existing emulator package should (in the existing version) continue to work for as long as possible. If changes are necessary nonetheless, it is preferable to require the same changes for all emulator packages and not require to maintain and adapt every emulator package individually.

One could ensure this, e.g., by archiving the emulator's source code and compiling it on the host platform as soon as the emulator is used. While this might seem to be useful for guaranteeing independence from the host platform, in practice, the approach is not really feasible, due to both time needed for compilation as well as fragility of the compilation setup in regard to external libraries, used compiler versions and host platform. An emulator which was originally implemented for the x86 architecture, when ported e.g., to the ARM architecture, will usually not be usable without code adaptations. Thus, the most feasible approach is to package and archive binaries of the emulator and all of its dependencies, so that you will be able to at least run the emulator in exactly the same version again in future. For future maintenance, it generally is advisable to also archive the emulator's source code, for which there are already existing initiatives, which can be relied upon.⁴ A generally accepted way to package application in a self-contained package are Dockerfiles⁵ and Docker images. The latter are being standardized as OCI Image Format⁶ and only include compiled binary files, while a Dockerfile is able to describe the packaging process in a human readable as well as machine actionable format. Even though emulators should be replaced with equivalent contemporary implementations for performance, security and user-convenience reasons, packaging and preserving emulators as containers allows to resurrect these for reference purposes [16].

Besides the emulator and its dependencies, the emulator package also has to include a component implementing the generic API and speaking to the emulator. While the generic API is intentionally designed as simple as possible, so that this component can be implemented in a broad variety of programming languages and its implementation can also evolve over time without breaking existing emulator packages, a template implementation in JavaScript is provided as a basis for new emulator packages.⁷

⁴ e.g., Software Heritage source code archive, <https://www.softwareheritage.org/>

⁵ Docker builder reference documentation, <https://docs.docker.com/engine/reference/builder/>

⁶ OCI Image Format Specification, <https://github.com/opencontainers/image-spec/blob/main/spec.md>

⁷ For a preliminary example, see <https://gitlab.com/emulation-as-a-service/experiments/qemu-ld/-/blob/main/emulator.js>

A. Control API

The used Linux containers as defined by the Open Container Initiative (OCI) Runtime Specification⁸ provide a solid encapsulation of executed programs. This ensures that executed emulators can only access resources, e.g., files, inside their own container and not communicate with outside system services provided by the host. In the case of emulators, this especially includes graphical input and output, sound, and network. The main primitives provided by the OCI Runtime Specification to communicate with the world outside of the container are the initial processes's standard input/output streams and files or directories explicitly shared between the container and the host system (technically through so-called "bind mounts"). Every intended way of interaction with the emulator thus has to be implemented using these primitives. Incidentally, this also helps with forward compatibility as it reduces the used interfaces which have to be maintained, preserved and maybe re-implemented drastically.

The component translating from the generic API will take the configuration format described in the previous section and, generally, pass them to and start the emulator using command-line arguments or by generating emulator-specific configuration files. While, e.g., a sound card can run independently, for some categories of devices external data sources might have to be provided, e.g., a hard disk drive might need a disk image. The emulation framework has to prepare and provide this data to the container, e.g., by using the path property for hard disk drives to specify where the emulator package can find the respective disk image in the form of a raw file and, using the OCI Runtime Specification, share this file with the container at the configured path.

During execution of the emulator, further interaction with the emulator inside the container is needed. The most basic ways of interacting with emulators that have been identified for this work are keyboard and mouse input, video and sound output, and network access. We have evaluated each of these

areas individually and tried to find a suitable and stable (standardized) interface.

1. Video Output

For video output, the most basic interface would be a frame buffer, i.e., a serialization of the pixels shown on an emulated computer screen, starting at the upper left pixel of the first displayed line and ending at the lower right pixel of the last displayed line, using a fixed number of bytes, e.g., 3 bytes with one byte for each of the pixel's red, green, and blue component. We argue that a frame buffer offers a higher forward compatibility than more pre-processed serializations of video like, e.g., an H.264 video stream.

At the same time, it typically puts no big computational burden on the emulator package to produce a frame buffer, which will typically be the first step in producing any video output anyway, and not have to further process it. Additionally, due to encapsulation, this post-processing would have to be done in generic software without any acceleration possibly provided by individual host systems. It is thus much preferable to have access to the most basic video output (i.e., a frame buffer) from the emulator. This video output can then be further post-processed by the individual host system with its individual host-specific acceleration capabilities, e.g., encoding it to a compressed H.264 video stream to be sent to a user's web browser.

A frame buffer alone, though, is more like a generic abstraction and not already a technical realization. While it could be realized by constantly updating a file shared between host and container with the current content of the emulated display or sent as byte-stream through a socket or network, the technical characteristics need to also be communicated. For instance, each pixel could be represented by 2, 3 or 4 bytes per pixel. Furthermore, the dimensions of an emulated computer screen might be fixed and known in advance, for many platforms it can also change during execution, causing the size of the frame buffer to change as well.

⁸ Open Container Initiative (OCI) Runtime Specification, <https://github.com/opencontainers/runtime-spec/blob/main/spec.md>

2. Input Events

With forward compatibility in mind, we propose that the video protocol should be a protocol that is common, well understood, and ideally has a wide range of actively maintained implementations as well as tool support. Potential options include X11⁹, Microsoft's Remote Desktop Protocol (RDP)¹⁰, the Remote Framebuffer Protocol (RFB)¹¹ used by various VNC programs, and Wayland¹². Evaluating these protocols, RDP does not enjoy wider Open Source tool support and, as a proprietary protocol, it is likely to be more difficult to maintain and support over time. We currently consider Wayland as not fully mature, such that forward compatibility of the current state is not ensured. We still expect that Wayland matures quickly and will become a viable candidate in the future. The RFB protocol is a high-level protocol also offering different video-codecs. Its "Raw Encoding" would be equivalent to the aforementioned framebuffer but few emulators offer direct RFB support, requiring an additional active component inside the emulator container capturing the emulator's output and producing a RFB stream. X11, a very well established and very mature (though somewhat complex) protocol, is already supported by almost all software with graphic output within the Linux/Open Source ecosystem including emulators. It supports access to the frame buffer of an application from any other application, and optionally, by directly using shared memory. While this is nowadays sometimes considered as a problem regarding security, for our purposes it is helpful to access any application output from outside the container. Security is not an issue in this specific case since a separate X11 server is deployed in each container and the container is already separated from all other containers as well as the host system. Additionally, X11 offers native support for accessing the X11 server using UNIX domain sockets represented as (special) files in the file system. These can easily be shared and accessed by the host system.

In order to allow users to interact with an emulated guest, emulated input devices need to be connected to the user's contemporary input devices. All aforementioned protocols not only offer transport options for the emulator's video output but also the ability for input events. In case of X11, the XTEST extension¹³ allows sending relative and absolute pointer coordinates and keyboard input to the emulator. For keyboard input, sending "keycodes" based on a key's physical location as well as (indirectly) sending "keysyms" based on a key's meaning is supported. This is relevant as an operating system in an emulator will typically allow users to configure keyboard layouts itself and thus passing through the location of a key as opposed to its (irrelevant) meaning on the host system is desirable, e.g., the letter "Z" typed on a "QWERTY" keyboard would be interpreted as a "Y" on a "QWERTZ" keyboard and a "W" on an "AZERTY" keyboard. For RFB, allowing to send keys based on their location is only possible using an extension¹⁴, making X11 a more desirable protocol for keyboard input independently of video output as well.

3. Audio Output

For sound output, popular options used by emulators within the Linux ecosystem include the Advanced Linux Sound Architecture (ALSA)¹⁵, PulseAudio¹⁶ and PipeWire.¹⁷ While ALSA is a rather low-level interface mainly used as interface to the Linux kernel, PulseAudio shares many of the useful characteristics of X11, in particular, exposure via a UNIX domain socket as a file in the file system. We consider PipeWire to be in a similar state as Wayland, quite promising but still not matured enough yet. Thus, we chose to use PulseAudio for sound output.

4. Network

For network access, one has to differentiate between access to the Internet used by the emulator itself

⁹ Currently maintained by the X.Org project (<https://gitlab.freedesktop.org/xorg/protocol/>)

¹⁰ https://docs.microsoft.com/en-us/openspecs/windows_protocols/ms-rdpbcgr/5073f4ed-1e93-45e1-b039-6e30c385867c

¹¹ Specified in RFC 6143 (<https://www.rfc-editor.org/info/rfc6143>) and currently maintained by the community (<https://github.com/rfbproto/rfbproto/blob/master/rfbproto.rst>)

¹² <https://gitlab.freedesktop.org/wayland/wayland/-/tree/main/protocol> and

<https://gitlab.freedesktop.org/wayland/wayland-protocols>

¹³ <https://gitlab.freedesktop.org/xorg/protocol/xorgproto/-/blob/master/specs/xextproto/xtest.xml>

¹⁴ QEMU Extended Key Event Message

¹⁵ <https://alsa-project.org/>

¹⁶ <https://www.freedesktop.org/wiki/Software/PulseAudio/>

¹⁷ <https://pipewire.org/>

(e.g., an emulator might support downloading content from the live Internet) and the emulator emulating a network interface card to the guest platform. The former is not necessarily desirable as the goal of encapsulating the emulator is to eliminate any dependencies on external sources. By enabling network namespaces, the OCI Runtime Specification already covers this by forbidding access to the Internet. Managing network traffic from an emulated network interface card is handled separately. Currently, the only relevant network type supported by any emulators – if there is support for network – is Ethernet, which has to be accepted from and sent to the emulator, preferably via shared (special) files exposed within the local file system. We have chosen the Virtual Distributed Ethernet (VDEv2)¹⁸ [17] project to provide such an abstraction of Ethernet. Particularly, its vdeplug library defines a simple format, i.e., prefixes each Ethernet frame with its length as 16-bit big-endian integer, to turn Ethernet frames into a stream, which can then easily be tunneled over many types of transport. [18]

5. Emulator Control

During execution of an emulator, users might want to change its configuration, e.g., changing the media of an emulated CD-ROM drive. Like the initial configuration, this request has to be translated into specific actions passed to the emulator, e.g., QEMU's QMP monitor. The emulator package thus has to accept generic requests and turn them into emulator-specific requests during runtime.

As easily in a wide range of languages implementable control protocol, we propose a JSON-RPC¹⁹ based protocol sent over the container's initial process's standard input/output streams. An example to change the media in an emulated CD-ROM drive can be seen in Figure 3.

Conveniently, the same protocol can be used to transfer both the initial configuration with the request to start the emulator as well as any events originating from the emulator, e.g., a notification that an emulator has exited, errored, or a virtual power button was pressed, back to the emulation framework.

```
{
  "id": 100,
  "method": "changePath",
  "params": {
    "id": "#cdrom-1",
    "path": "/mnt/disks/new-cd-rom.raw"
  }
}
```

Figure 3 Control protocol

B. Re-usable Implementation

The proposed architecture deliberately is as independent from a surrounding framework as possible. It only requires to run Linux containers according to the OCI Runtime Specification, to share relevant files and directories with the host system. It does not predefine any special control files but relies on sending requests to the container's initial process via its standard input. For the emulator's input and output, existing and widely used protocols are used. We hope that the provided emulator packages can thus be useful and re-usable for new and future implementations of emulation frameworks.

The external emulation framework has to provide access to the exposed input and output, e.g., keyboard input and video output, to the user. It can use, e.g., Xpra²⁰ to allow users access from their web browser.

We propose to use GitHub as a collaborative platform for collecting information about emulators, turning them into emulator packages, and maintaining them. A common "emulation-archive" project facilitates discovery of existing emulator packages, which can exist as one repository per emulator. A template repository can be provided, which can be forked as a basis for new emulator implementations.

V. CONCLUSION

In this article we laid out the technical foundations to improve portability and re-use of emulators. Machine actionable technical metadata will support further automation, interoperability and eventually, a more sustainable emulation infrastructure. But most importantly, the technical design as well as the

¹⁸ <https://github.com/virtualsquare/vde-2>

¹⁹ <https://www.jsonrpc.org/>

²⁰ <https://xpra.org/>

extensible technical descriptions of emulators provide a framework for cooperative work to identify, list, describe and integrate emulators of interest for the digital preservation community. We have deliberately chosen an open, collaborative way as the emulator landscape is quite scattered as is the detailed knowledge about less common computer systems.

REFERENCES

- [1] J. R. van der Hoeven, R. Verdegem, and B. Lohman, "Emulation for digital preservation in practice: The results," *Int. J. Digital Curation*, vol. 2, pp. 123–132, 2007.
- [2] K. Rechert, I. Valizada, D. von Suchodoletz, and J. Latocha, "BwFLA - A functional approach to digital preservation," *Prax. Inf.verarb. Kommun. (PIK)*, vol. 35, no. 4, pp. 259–267, 2012.
- [3] J. R. van der Hoeven, R. J. van Diessen, and K. van der MEER, "Development of a universal virtual computer (uvc) for long-term preservation of digital objects," *Journal of Information Science*, vol. 31, no. 3, pp. 196–208, 2005.
- [4] B. Lohman, B. Kiers, D. Michel, and J. van der Hoeven, "Emulation as a business solution: The emulation framework," in *Proceedings of the 8th international conference on preservation of digital objects (iPRES 2011)*, 2011, pp. 167–170.
- [5] W. Bergmeyer, "The keep emulation framework." in *Proceedings of the 1st international workshop on semantic digital archives (sda 2011)*, 2011, pp. 8–22. Available: <http://ceur-ws.org/Vol-801/paper1.pdf>
- [6] S. Ortiz Jr, "The problem with cloud-computing standardization," *Computer*, vol. 44, no. 7, pp. 13–16, 2011.
- [7] DMTF System Virtualization, Partitioning, and Clustering Working Group, "Open virtualization format specification." Document DSP0243, 2010.
- [8] M. Bolte, M. Sievers, G. Birkenheuer, O. Niehörster, and A. Brinkmann, "Non-intrusive virtualization management using libvirt," in *2010 design, automation & test in europe conference & exhibition (date 2010)*, 2010, pp. 574–579.
- [9] D. Anderson, J. Delve, and D. Pinchbeck, "Toward a workable emulation-based preservation strategy: Rationale and technical metadata," *New review of information networking*, vol. 15, no. 2, pp. 110–131, 2010.
- [10] J. Delve, L. Konstantelos, A. Ciuffreda, and D. Anderson, "Documenting technical environments for posterity: The totem registry and metadata schema," *PIK-Praxis der Informationsverarbeitung und Kommunikation*, vol. 35, no. 4, pp. 227–233, 2012.
- [11] A. Brown, "Automatic format identification using PRONOM and DROID," *Digital Preservation Technical Paper*, vol. 1, 2006.
- [12] K. Thornton, E. Cochrane, T. Ledoux, B. Caron, and C. Wilson, "Modeling the domain of digital preservation in wikidata." 2017.
- [13] J. Giessl, R. Gieschke, K. Rechert, and E. Cochrane, "Automating the selection of emulated rendering environments for born-digital data-sets," in *Linking theory and practice of digital libraries - 25th international conference on theory and practice of digital libraries*, 2021, vol. 12866, pp. 106–111.
- [14] PREMIS Editorial Committee, "PREMIS data dictionary for preservation metadata, version 3 (library of congress, november 2015)."
- [15] A. Dappert and A. Farquhar, "Digital preservation metadata practice for computing environments," in *Digital preservation metadata for practitioners: Implementing premis*, A. Dappert, R. S. Guenther, and S. Peyrard, Eds. Cham: Springer International Publishing, 2016, pp. 129–138.
- [16] K. Rechert, T. Liebetraut, D. Wehrle, and E. Cochrane, "Preserving containers - requirements and a todo-list," in *Digital libraries: Knowledge, information, and data in an open access society - 18th international conference on asia-pacific digital libraries, ICADL 2016*, 2016, vol. 10075, pp. 225–230.
- [17] R. Davoli, "Vde: Virtual distributed ethernet," in *First international conference on testbeds and research infrastructures for the development of networks and communities*, 2005, pp. 213–220.
- [18] R. Geischke, K. Rechert, and S. Mocken, "Preserving access to web servers. A case study preserving output of collaborative research centers," 2021.

PRESERVATION STRATEGIES FOR NEW FORMS OF SCHOLARSHIP

Deb Verhoff

NYU Libraries
United States
deb.verhoff@nyu.edu
[0000-0003-1700-7199](tel:0000-0003-1700-7199)

Karen Hanson

Portico
United States
karen.hanson@ithaka.org
[0000-0002-9354-8328](tel:0000-0002-9354-8328)

Jonathan Greenberg

NYU Libraries
United States
jonathan.greenberg@nyu.edu
[0000-0002-3429-4428](tel:0000-0002-3429-4428)

Abstract – The advance in technologies for publishing digital scholarship has outpaced the development of technologies for reliably preserving it. Authors and publishers are creating increasingly sophisticated products without realizing that some of their enhancement choices might put preservability--and valuable scholarship--at risk. In a project funded by Andrew W. Mellon Foundation and led by NYU Libraries, a group of digital preservation institutions, libraries, and university presses collaborated to study examples of these dynamic forms of scholarship to determine they could be preserved in their current form and whether it would be possible to do this at scale. This paper will provide a summary of this project and key themes that could impact preservation of enhanced scholarly works.

Keywords – scholarly publishing, web archiving, emulation, dynamic content, preservation strategy

Conference Topics – Community; Innovation.

I. INTRODUCTION

Scholars are making extensive use of new digital technologies to express their research. Publishers, in turn, are working to support increasingly complex publications that are not easily represented in print. These enhanced digital products introduce new complexities in content and user experience. Examples include publications with embedded audio and video content, high-resolution images, data, maps, and visualizations; non-linear paths of engagement; and complex interactive features that depend on third party platforms or APIs, such as YouTube or Google Maps. Each of these innovations presents preservation challenges; their combination creates an even greater challenge: the need to maintain multiple formats and the connections among them, all within workflows designed for simpler objects.

To study this challenge, a group of digital preservation institutions, libraries, and university presses worked together on an Andrew W. Mellon Foundation funded project, Enhancing Services to Preserve New Forms of Scholarship, led by New York University Libraries. Preservation service providers, such as Portico and CLOCKSS, rely on economies of scale with replicable processes, and as such, they must determine what aspects of new scholarly communication can be preserved at scale. Authors and publishers, for their part, must provide sufficient contextual information for publications in order for essential features to be preserved. Together, a team of publishers, librarians, and preservation specialists examined a variety of enhanced digital publications in order to identify what can be effectively preserved at scale with existing technologies. This analysis was used to produce a recommended set of practices to help authors and publishers prioritize and plan their enhanced digital products for maximum preservability. A full report [1] on the project and the resulting guidelines [2] for authors, publishers, and publishing platform developers have been published. A summary of the project and reflections on key themes that could impact preservation of enhanced scholarly works are described in this paper.

II. METHODS

Project participants represented scholarly publishers, preservation services organizations, and libraries that may provide publishing services, preservation services, or both. Publishers included NYU Press, Michigan Publishing, the University of Minnesota Press, UBC Press and Stanford University Press. Four out of five of the participating publishers also participated as platform developers: NYU Press

for Open Square, Michigan Publishing for Fulcrum, the University of Minnesota Press for Manifold, and RavenSpace at UBC Press. Preservation service organizations included CLOCKSS, Portico, and the libraries of the University of Michigan and NYU.

The 18-month-long project was divided into three sprints, with publications grouped by their technical features and in order of what was perceived to be the least to most complex. During the first sprint, the team worked with EPUB-based publications that include a variety of multimedia and supplementary material either within the EPUB itself or as a platform-level resource. During the second sprint, the team modeled solutions for preserving web publications with a linear, text-based structure and a broader range of added digital resources. Though these publications are interactive, users engage with them through a predictable set of interactions. Many of the publications in both the first and second sprints support enhanced features such as annotations, embedded multimedia, and data visualizations. The third sprint covered the most complex, media rich, and nonlinear publications for which an interactive experience is at the forefront. In this sprint, the team worked with more dynamic publications in which third party dependencies are an integral component. The workflow within each of the sprints was designed to capture data from the participants during each phase of submission and evaluation for a publication.

During an initial evaluation phase, the assigned publishers and preservation partners collaborated to perform a detailed review of each publication. Together they defined the core intellectual components of the publication — those that must be preserved for future audiences to fully understand the work's substance and arguments. Publishers provided detailed instructions for the playback or reading experience of the material submitted. They described what an intended audience should be able to do when the archived content is made available in the future. These core intellectual components served as acceptance criteria for the success of the work done in subsequent phases. In addition, description and documentation of these components gave preservation providers a more complete understanding of the context and dependencies for a work.

In the preservation action phase, each publication was analyzed by one or two preservation

services. A series of tools and techniques was applied, including normalization of export packages, web archiving (LOCKSS, Heritrix, Brozzler, Squidwarc, Memento Tracer, and Browsertrix), and emulation (EaaS). Preservation specialists determined which of the publication's required core components could be preserved and to what degree the approach might be scalable. Works that progressed through the preservation actions were moved forward for assessment.

The Portico and CLOCKSS model is to provide access to ("trigger") a scholarly work if it is no longer available through any publisher. The services register their triggered copy with CrossRef so that researchers will be redirected to the preserved copy if using the DOI. This makes access an important consideration for both services, and so evaluating the rendition copy for fidelity of the core intellectual components was one component of this analysis. Though there are risks that occur over time as technologies change, if the sample rendition is not close to matching the publisher requirements, then the preservation is challenged from the outset. Publishers received a mixture of mockups and actual preservation packages for what the items could look like if triggered for access. They tested them to determine whether the required and preferred features were captured appropriately, and they answered questions related to the playback experience of the preservation copy of a work. This process captured the degree to which the archived content matched the preservation goals and expectations about what would be preserved. The preservation services documented what was preservable using current tools. They recorded any constraints such as technical limitations, scalability of the approach, or limits on what was feasible in the time frame provided. 20 complex works were analyzed to determine their preservability at scale. Among them were 17 works from six different publishing software platforms, plus three websites that were constructed to present a single work. Though these works represent a diverse sampling, some cross-cutting themes emerged, each with implications for preservation strategy. The key themes are described here.

Together, the project team recorded lessons learned from each work. They made note of patterns that supported preservation and modifications that a publisher could have made during the creation of the original work to improve the preservability of the

material while maintaining the essential aspects of the content. This formed the basis of the guidelines [2] for improving the preservability of these works. In turn, the preservation specialists documented their boundaries. They identified the effort required to create each new preservation workflow, as well as the likelihood that the approaches could be replicated at scale. The team also noted improvements to both the publishers' and preservation services' existing workflows that could help accommodate future requests and improve efficiency.

III. RESULTS AND OBSERVATIONS

Resources Not Supplements

Both CLOCKSS and Portico preserve supplemental files that are provided with a publication, but the majority of traditional publications do not have any. Where they do have supplements, they are typically few in quantity, rarely have comprehensive metadata, and are sometimes not included in export packages sent for preservation. Most of the publications analyzed not only include additional resources but have an unusually large quantity and diversity of them. Of the 20 publications analyzed, 17 have files in addition to the main text, 11 have over 100 files, and five have over 400. Resource types include text, image, audio, video, software, and a wide variety of data files. *Developing Writers in Higher Education* [3], for example, includes 283 PDFs, 31 videos, 22 audio files, and three images in addition to the EPUB for the text, totaling 5.9GB.

The text plus these resources are considered to be the work. In four platforms analyzed, structured descriptive metadata is applied to these resources. Each has a dedicated landing page within the platform, and in some cases, a persistent identifier is assigned to support independent citation. When looking at how these resources relate to the main text, they are either: visually embedded in the text; linked directly from the text using the landing page URL; or unlinked supplements available with the main text to provide context. Two platforms, Fulcrum and Manifold, refer to these additional files as "resources" and the others call them "files," which implies a more ambiguous relationship to the text than supplements. Conversations with the publishers confirmed that this distinction is intentional.

Increasingly, funded research requirements prescribe sharing supporting evidence for a publication. This project showed that the traditional lines between text, figures, and supplements continue to blur with "figures" being independently citable artifacts and "supplements" being a vital part of the work. For preservation purposes, the inclusion of these resources in the publishers' exports, the addition of structured metadata, and use of persistent identifiers is helpful. While working with the publishers, the preservation services highlighted the advantages of using non-proprietary, broadly adopted file formats where possible, but recognize that the innovative nature of the works means there will likely always be unexpected formats in the archive. The addition of descriptive metadata is especially helpful in these instances. Also challenging is to ensure that these works, which are internally a map of linked resources, are captured appropriately with all components and the relationships between them intact.

Preservation Strategy Considerations

For the preservation services, the complexity, volume, and variety of formats within a single work presents a challenge for managing and eventually supporting access to the work. First, the diversity of file types highlights the importance of collective efforts such as PRONOM to ensure a high proportion of the files can be identified and matched to an appropriate rendition approach in the future. Second, preservation services will need to consider how to arrange these complex composite works in the archive to ensure they are manageable, discoverable, and eventually accessible. In some cases, it may be practical to keep the entire work in a single Archival Information Package (AIP) and focus on extracting and indexing metadata to reveal the component resources. Alternatively, it may be more elegant to atomize a complex work so that each component resource has its own AIP with links and relationships between the resources recorded in structural metadata. This atomization would allow for flexibility in package management (for versioning individual resources, migration etc.) and more closely reflects how they are managed on the publisher platforms. Finally, the treatment of resources as citable artifacts adds complexity to rights management. Traditional publishing workflows manage the rights for embedded figure graphics in the context of the work, but if managing hundreds of resources that can be viewed independently or as one, the rights status must be

defined through the structured metadata or constrained through the publishing workflows in order for preservation decisions to be possible at scale.

Resources Embedded via Iframes

One of the most frequent challenges found within the publications analyzed was the use of HTML iframes to visually embed the content of a webpage into a work. Iframes are present in the majority of the works reviewed during this project. They are primarily used for media players, user contributed content, or data visualizations such as maps. It is technically simple to embed web content in web content, and generally acceptable to use iframes on the live web without procuring rights for the embedded content. Attempting to copy and archive these features, however, presents a variety of challenges. This content may be lost without coordination between the publisher and the preservation service.

Three key factors related to iframes affect the options for publishers and preservation services. The first factor is the format of the publication. The research focused on EPUBs and web-based publications¹. For EPUBs, the technical challenges are more complex than for web-based publications. The EPUB specification [4] allows iframes but requires that a fallback reference be defined since a reader may not support them. Iframes were found in five of the 10 EPUBs evaluated and were used by two of the three publishers that produced EPUBs. None of the iframes in the examples had fallbacks. This dependency needs careful management if the publication is to be preserved.

A second factor is whether the iframe resource is on the publisher's platform or a third-party platform e.g. YouTube. If using a third-party platform, the long-term viability for the content improves if it is uploaded and managed by the publisher, and original files and metadata are retained in case the third-party version becomes unavailable. If a work is a composite of webpages on multiple platforms all managed by the publisher and all original files are intact, it becomes plausible to craft processes that pull content together for preservation. Using iframes to include third party platforms not managed by the publisher is challenging both technically and legally.

In several examples with YouTube videos that weren't managed by the publisher, the content became unavailable after publication. In this respect, use of third-party platforms to embed things is not just a preservation challenge but one of sustainability for the publisher since the content can disappear before a preservation service is involved.

The final factor is how dynamic the iframe content is. All iframe resources are referenced using a URL. In some cases, all relevant data is loaded when the URL first loads. In others, a limited and predictable set of interactions may load all necessary data (e.g. click play). Either of these may be possible to archive with a web crawler if the original files are unavailable or not sufficient to represent the functionality of the iframe. When iframe content is highly dynamic, that is, when user interaction depends on perpetual communication with the server, it can be difficult to preserve. In these cases, resources are composed of an open-ended number of possible URLs that vary by user interaction. Typical examples of features that are dynamic in this way are map visualizations, IIIF viewers, and search features for which each user interaction loads a new response from the server. The more dynamic a resource is, the less likely it can be preserved at scale in website form. The only option may be for the publisher to provide the underlying data and/or software for the resource if available. Website preservation will be discussed further in The Experience of the Work section below.

One of the challenges in articulating guidelines for handling iframes was identifying their characteristics and mapping them to the methods for mitigating loss. If using a web crawler to preserve a web-based publication with an iframe featuring a simple static HTML page hosted on the publisher's platform, the iframe is likely inconsequential to the preservation approach. The same static HTML page in an iframe within an EPUB presents a more complicated challenge to harvest and then associate the page with the EPUB file. If the iframe contains dynamic data-driven content or exists on a third-party platform not managed by the publisher, the challenges are multiplied.

¹ Some EPUBs were both downloadable and presented on the website using an EPUB reader; the online version is considered web-based.

Preservation Strategy Considerations

For services that aim to preserve these forms of scholarship, building a strategy for iframes depends on the combination of the factors described. Ideally publishers would keep track of the use of iframes in publications or label them so that domain names or URLs that are in scope for crawls can be easily identified by preservation services. Where the preservation copy must cross boundaries of formats in order to cover the content (for example, where iframes are embedded in an EPUB, or data files are supplied for a visualization that cannot be copied), preservation services will need to consider the appropriate strategy for each format and ensure the metadata tracks the links and relationships between the original iframe URL and archived resources. Considering how to present these parts in a way that is useful for future scholars helps focus this work on ensuring all data necessary to do this is collected.

Living Documents

Managing and connecting versions of content over time is a common digital preservation challenge. In traditional academic publishing, a DOI or ISBN is assigned to a particular version of record. While imperfect [5][6], this rigidity has been useful for those who preserve scholarship in supporting review of content for duplication and completeness. Discussions about versioning of scholarly contributions that fall outside of traditional workflows have been developing for a number of years in communities like Force11. Similarly, this research highlighted the need to record new versions of scholarly works outside of the traditionally controlled correction and retraction workflows.

Perpetual Drafts

Two of the 20 publications evaluated were in draft state during the assessment. On Revaluation of Value on the Manifold platform is in a perpetual draft state and may remain that way indefinitely with occasional updates. The publisher indicated that even though publications on the Manifold platform were in draft state, it was important not to wait until they were officially “published” to preserve them since the draft state and iterative approach to the work may be intentional.

User Contributed Content

Seven of the works had user contributed comments or annotations. Annotations and highlighting are built into the Manifold platform, and

the landing page for each book integrates Tweets that have referenced the publication. *Rhizcomics* [7] from Michigan Publishing features both a Disqus comments integration and a Hypothesis annotation toolbar. While this content was considered nice-to-have for preservation in most examples, some publishers explained that for certain cases this was an important piece of the work. Some annotations were added by the authors after publication and others held useful context.

Preservation of user contribution features has technical, legal, and ethical challenges. When managed within the platform software, there is more flexibility since publishers can incorporate language to support preservation into the Terms of Service. It also allows for data export and migration of user contributed content to new platforms. Many third-party integrations for comments and annotations are tied to the URL and may be at risk of loss if the URL changes. When a third-party service is used, their platform Terms may hinder preservation. Even if the content is legal to preserve - Hypothesis users, for example, implicitly agree to make public content CC0 licensed [8] by using the platform - unless moderated, there is nothing to prevent users from posting copyrighted content. In the case of integrations such as Twitter feeds, copying an account handle, photo, and Tweet content without the permission of the author prompts ethical and legal concerns. For these reasons, inclusion of user-contributed content for the purpose of preservation must be weighed against the risk factors. Where this content is considered vital and is covered by Terms, it instead becomes a versioning challenge within which parts of the content might change while its identifier remains the same.

Preservation Strategy Considerations

For works with non-traditional requirements for versioning, preservation services and publishers should discuss what parts of the publication might change and over what period, then establish criteria for determining when to preserve a new copy. Many of the works in this research were large with numerous component parts. Efficient versioning criteria combined with workflows that only update the files that have changed can avoid unnecessary redundancy and overuse of storage. If versioned content is eventually triggered by the preservation service, there will need to be a mutual understanding about which version(s) should be made available for access.

The Experience of the Work

Traditional digital publications primarily simulate print publications; they consist of static linear text broken up by sections and images. Many of the works analyzed for this project present users with a carefully crafted dynamic experience. The publisher's impression of how much of this experience should be preserved varied for each work. Conversations to understand the scope of the experience that should be preserved were critical to determining the most efficient approach for preservation. With publications on Fulcrum, for example, the specific experience offered by the platform was viewed as less important than preserving the component parts and connections between them so that they could be reassembled on a future platform. Three other works, whose platform was designed as part of the publication, offer a unique experience that is fundamental to understanding the creator's intent. RavenSpace also has a number of important interactive features that are difficult to separate from the platform (the popup agreement asking that visitors are respectful guests, the ability to search the site using the First Nations keyboard, and the non-linear style of navigation).

For these works, if it can be performed with reasonable accuracy and at scale, a web harvested version can be the most efficient way to copy the work and then quickly re-render it using a WARC player to maintain elements of the original experience. A useful aspect of the CLOCKSS and Portico service model is the option to spend time customizing a solution to match a platform's unique features. In each of the platforms analyzed for web harvesting (Manifold, Fulcrum, Scalar, and RavenSpace), a fully automated crawl without any site-specific configuration did not record all of the features that were considered vital to the experience of the publication. None of these platforms include sitemaps, and so, a mixed strategy was applied to ensure the crawlers visited all of the URLs that made up the publication's vital functionality. For Manifold and Scalar, the open API was used to create a sitemap and additional configuration was added to ensure URLs that result from key user interactions (e.g. opening out the menu levels on Scalar) were retrieved. CLOCKSS utilized the LOCKSS technology for the crawls, while Portico tested a selection of browser-based crawlers, with Brozzler used most frequently. Ultimately the biggest challenges were

the same across all crawler tools - archiving highly dynamic features in which the combination of URLs that make up the feature cannot be reasonably predicted using a script. Data driven search interfaces, IIIF viewers, and map visualizations, for example, were consistently missed from web crawls since these load new URLs based on specific user interactions.

A final experiment to test options for preserving the experience involved recreating two of the most dynamic publications on virtual machines so that their websites could be emulated in the future. This was attempted for *As I Remember It* [9] and *Filming Revolution* [10], since web harvesting attempts fell short for these two. The publications, both built on LAMP stacks, had to be adapted for encapsulation. This took several days for each [11] and involved copying dependencies (multimedia, fonts, etc) to a local directory on the machine and then updating the code to point to those directories. Once encapsulated, the machines were loaded into the EaaSI platform and tested with the Internet connection disabled. For both publications, the playback via EaaSI was at a very high quality that met all of the publisher's requirements. While a preservation service is unlikely to apply significant code edits as part of their usual services, our purpose here was to understand the effort of encapsulation and confirm that this approach might be feasible during initial development of the project with little to no extra work if the developer is aware of the preservation and sustainability implications of external dependencies. In one illustrative example, a site's load function was called when its Google font loaded successfully. If Google stopped supporting that font, the site would stop working and a developer would have to determine why. If the publisher did not have a developer available to analyze the issue, the publication might be taken offline. Using a local non-proprietary font would have eliminated this risk. When the project is preserved, these challenges are transferred to the preservation service, and repairing websites does not scale well across hundreds or thousands of projects.

Preservation Strategy Considerations

There are a diverse set of tools for website archiving, and many support extensive customization at the platform level. It is clear that customization can go a long way to improving the quality of web crawling, and for services working with

specific publishers, knowing the platform is an important advantage. The challenge then becomes monitoring the quality of crawls over time to ensure the tools maintain an accurate crawl, and that platform changes are detected and remain preservable.

In some cases, platforms are too dynamic to be harvested using a web crawler, and the only option for preserving the experience is server-side preservation. While creating a virtual machine to replicate a one-off project like *Filming Revolution* seems appropriate, it is more complicated to envision how to do this efficiently across thousands of works from the same platform since preserving thousands of virtual machines would be very costly. In theory, a virtual machine containing a pre-installed publisher platform could be prepared, along with a short script to bootstrap a work into it. The theory is untested, and scalability is contingent on highly consistent packages from the publisher. The packages seen during this research did not meet this requirement but had potential. If successful, this may be the most efficient approach to preserving the experience of some of the most complex works from publisher platforms.

IV. CONCLUSION

Enhancing Services to Preserve New Forms of Scholarship set out to determine what aspects of enhanced dynamic scholarship could be preserved at scale. In the majority of examined works, with preservation services giving individual attention to each, it was possible to identify an approach that would be acceptable for the publisher. The exceptions were those in which a significant portion of the work was dependent on a third-party service and there was no way (legally, ethically, and within the timeframe for the analysis) to copy that content or represent it locally in a more preservable form.

While preservation approaches could be applied to navigate challenges within individual works, it was the scalability aspect that introduced the biggest constraints. As workflows were retested on different projects from the same platform, some patterns around what scaled were revealed. The overall structure and text of a work can be captured consistently if (a) it follows a predictable template or conforms to format standards and best practices and (b) it is possible to spend time configuring preservation workflows that align with that template. For example, if standard HTML conventions are

followed for hyperlinks and multimedia, these may be easily crawled using a standard web crawler without additional configuration. In most cases, however, the features that caused the work to meet the criteria for inclusion in this project were the ones whose implementation varied widely, making them challenging to preserve at scale and at the highest risk of loss. The novelty of these features in a publishing context means there are few standards or best practices for how to integrate them into the work in a form that makes it easy to design scalable workflows for preservation. When configuring a workflow for this kind of content, the preservation services must therefore depend on patterns established in examples provided. If a single feature strays from the patterns established during the configurations, the workflow could miss important components and possibly do so without detection. In many instances, the features that tended to introduce unpredictability in the quality of preservation were inside iframes. These often hold content that makes the work unique and so cannot be broadly excluded, but also represent the biggest challenge to managing the scalability of the preservation process.

As is often the case with digital preservation, technical challenges were also sometimes surpassed by legal or even ethical questions (in the case of user-contributed content) around whether the content should be preserved. With no automated way to make the distinction, an excess of caution around undefined license status can lead to significant and unnecessary loss.

The level of effort for building a scalable approach for preservation was also a challenge. Capturing the core features of each work in a multi-publication platform took weeks instead of days due to the complexity of the works. Spending weeks to develop a unique configuration might be an acceptable level of effort for broadly adopted platforms, but is much less scalable or affordable if there are many different platforms with a small number of works on each or a lot of inconsistency between each work. Add to this challenge building in quality control to detect minor variations between templates, and the effort required for high quality preservation at scale may become insurmountable.

A remedy to these scalability challenges is for publishers, authors, and platform developers to introduce some uniformity and emphasize approaches that will support automation in

preserving the works. The guidelines that resulted from this project were conceived to facilitate a conversation between preservation services and those that create complex enhanced scholarly works to enable the creators and curators of the works to play a role in planning for preservation.

We recognize that these guidelines will likely be difficult for the most under-resourced publishers to implement, which may compound the existing challenge of preserving works from smaller publishers. Moving forward, the project team will continue to partner with those involved in developing commonly used open source platforms so that changes made for preservation at the platform level can be felt by all users of the platform. If the preservation and publishing communities can coalesce around some standard approaches and continue this conversation as innovations progress, the preservation services can make changes to their services to improve support for new forms of scholarship that will scale.

REFERENCES

- [1] J. Greenberg, K. Hanson, and D. Verhoff, "Guidelines for Preserving New Forms of Scholarship," 2021 [Online]. Available: <https://doi.org/10.33682/221c-b2xi>
- [2] J. Greenberg, K. Hanson, and D. Verhoff, "Report on Enhancing Services to Preserve New Forms of Scholarship," 2021 [Online]. Available: <https://doi.org/10.33682/0dvh-dvr2>
- [3] A. R. Gere, *Developing Writers in Higher Education: A Longitudinal Study*. Ann Arbor, MI: University of Michigan Press, 2019. [E-book] Available: <https://doi.org/10.3998/mpub.10079890>
- [4] The World Wide Web Consortium, "EPUB Packages 3.2," May 8, 2019 [Online]. Available: <https://www.w3.org/publishing/epub3/epub-packages.html>
- [5] L. J. Hinchliffe, "The State of the Version of Record," The Scholarly Kitchen, blog, February 14, 2022 [Online]. Available: <https://scholarlykitchen.sspnet.org/2022/02/14/the-state-of-the-version-of-record>
- [6] M. Klein and L. Balakireva, "On the Persistence of Persistent Identifiers of the Scholarly Web," arXiv, April 6, 2020 [Online]. Available: <https://arxiv.org/abs/2004.03011>
- [7] J. Helms, *Rhizcomics*. Ann Arbor, MI: Michigan Publishing, n.d. [E-book] Available: <https://www.digitalrhetoriccollaborative.org/rhizcomics>
- [8] Hypothesis, "What is the license on annotations?," Accessed March 3, 2022 [Online]. Available: <https://web.hypothes.is/help/what-is-the-license-on-annotations>
- [9] E. Paul, *As I Remember It: Teachings (?ams ta?aw) from the Life of a Sliammon Elder*. Vancouver, BC: RavenSpace Publishing, 2019. [E-book] Available: <https://doi.org/10.14288/SNS9-9159>
- [10] A. Lebow, *Filming Revolution: A meta-documentary about filmmaking in Egypt since the Revolution*. Redwood City, CA: Stanford University Press, 2018. [E-book] Available: <https://doi.org/10.21627/2018fr>
- [11] J. Mulliken, "Emulation progress through collaboration," SUPdigital, blog, March 10, 2021 [Online]. Available: <http://blog.supdigital.org/emulation-progress-through-collaboration>

GREEN GOES WITH ANYTHING

Decreasing Environmental Impact of Digital Libraries at Virginia Tech

Alex Kinnaman

Virginia Tech
United States
alexk93@vt.edu
[0000-0001-8943-8946](tel:0000-0001-8943-8946)

Alan Munshower

Virginia Tech
United States
alanmun@vt.edu
[0000-0002-2878-5896](tel:0000-0002-2878-5896)

Abstract – This paper examines existing digital library practices at Virginia Tech University Libraries, and explores changes in documentation and practice that will foster a more environmentally sustainable collections platform.

Keywords – sustainability, digital libraries, digital preservation, archives, appraisal

Conference Topics – Environment

I. INTRODUCTION

As digital library practitioners, we are investigating ways to guide digital curation practices more broadly across Virginia Tech University Libraries (VTUL), while prioritizing considerations for environmental sustainability. In doing so, we explore university and professional standards and ethics, using the 2019 article “Towards Environmentally Sustainable Digital Preservation” [1] as a guide to focus on immediate areas that we can address in our digital library workflows. We investigate our workflows for appraisal, digitization, fixity checking, and storage choices to identify areas of improvement that find balance between best practices and environmental sustainability. This topic aligns with the conference theme Environment, and seeks to understand the environmental impact of VTUL’s digital preservation choices on the community in which we live.

II. LITERATURE REVIEW

While federal action in the United States specifically addressing climate change has only emerged since the early 1990’s, libraries and archives have been attentive to growing concern decades prior [2]. C. Durham writes, “all cultural institutions

are vulnerable to other aspects of the Climate Emergency...[and] need to prepare and adapt for the world humanity has created for itself, and they need to prepare quickly” [3]. There is an evident impact of digital preservation activities on the environment.

Beginning in the 1980’s, innovative concepts such as natural air-conditioning of paper materials underlied the environmentally friendly mission-specific work of lending and efficient management of physical materials [4]. Similar practices spread internationally to address conservation by using structural, rather than artificial means, to control the environment [5]. The digital age, and the accelerating proliferation of technology, has removed digital content managers from a similar physical awareness of their environmental impact in day-to-day work. Large datasets and complex digital objects are primary responsibilities of cultural heritage institutions, often with many parties involved in the accessioning, processing, and management. However, the effects of not triaging these processes through audit or inventory can be compounding. These necessary actions may be in conflict with an environmentally-sustainable approach to collection management.

Missions and Statements of Shared Value from professional organizations are valuable resources, as we look to others for guidance on a charge towards more environmentally sustainable digital curation practices at VTUL. The Society of American Archivists (SAA) makes a clear case for green-focused practice, charging members to “Devise environmentally sustainable techniques for preserving collections and serving communities” [6]. There is an

understanding of the balance of the ever-present dialogue with environmental considerations: “[D]eveloping acquisition, processing, storage, and service models—must necessarily involve an ongoing awareness of the impact of archival work on the environment” [6].

While not addressing Environmental Stewardship directly in its 2018 Declaration of Shared Values, the Digital Preservation Services Collaborative (DPSC) has listed sustainability as a core value. Partnering sustainability and affordability in the list core values, DPSC is presenting sustainability as a general duty in providing services, though the key value of accountability may also serve to guide decision making on climate policies and renewable energy options [7].

The National Archives and Records Administration of the United States (NARA), has created a climate action plan, specifically aimed at addressing “one of the most significant issues impacting...long term continuity” [8]. Among the plan’s five action items is the “strengthen[ing] of NARA’s climate resilience by leveraging cloud-based solutions.” Benefits outlined include the safeguarding against weather events, a more secure data supply chain, and notably that a move to cloud systems “may ultimately reduce GHG [greenhouse gas] emissions due to consolidated cooling and controlling of the data centers” [8]. This is presented mainly as a hypothetical in the plan, not offering evidence for greenhouse gas reductions, other than demonstrating that a shift from in-person to virtual reading room practices will generally contribute to less emissions.

There is a growing corpus of scholars interested in further exploring the challenges of environmental stewardship and digital preservation. Most fundamental for the purposes of this article, is the work by K. Pendergras et al. “Toward Environmentally Sustainable Digital Preservation [1].” Critically, the authors parse out different types of sustainability efforts in the field, focusing their scholarship on environmental sustainability and digital preservation practices.

This comprehensive look at current practices provides a framework for organizations to shift towards environmentally sustainable goals.

“[Cultural Heritage Organizations] need to reduce the amount of digital content that they preserve while reducing the resource-intensity of its storage and delivery. To do so, cultural heritage professionals must reevaluate their basic assumptions of appraisal, permanence, and availability of digital content” [1].

Recommended approaches for this paradigm shift include addressing appraisal, permanence, determination of acceptable loss, fixity check methods and frequency, choice of storage technologies, file format migration policies, and the number of redundant copies.

While K. Pendergrass et al. [1] offer a number of avenues to explore in their paradigm shift, much of the existing additional literature has an emphasis on storage and the raw energy consumption of large data sets. This concern frames the immediacy of the need to create sustainable practices.

“Every decision to acquire, preserve, or replicate a byte of data is, essentially, a commitment to put some amount more carbon into the earth’s atmosphere. This reality should prompt a meaningful though difficult conversation about whether the survival of knowledge into the distant future will be primarily dependent on deliberately preserving less of it at lower quality” [9].

Virginia Tech itself is located in Montgomery County, Virginia. Virginia has a long history of coal mining as a major economic backbone. Beginning in the late-18th century coal has been mined in portions of Montgomery and Pulaski counties [10] after which production of coal ebbed and flowed until it climaxed in 1943-44 and continued well into the 1960’s. The worst but not only disaster on record occurred in April 1946 when a mine in McCoy, Virginia exploded from a methane leak and killed 12 miners, orphaning 51 children [11]. The relationship between the economy and industrial energy extraction in Virginia and in Montgomery County has lasted 250 years, and continues to be a primary source of income for the state¹ and a major cultural hub of the community.

As a cultural institution in the middle of the primary location for coal mining in southwest Virginia, Virginia Tech plays a role in tracking energy consumption for the University. The Virginia Tech

¹ Virginia Coal: <https://vept.energy.vt.edu/coal.html>

mission statement is as follows: "Inspired by our land-grant identity and guided by our motto, Ut Prosim (That I May Serve), Virginia Tech is an inclusive community of knowledge, discovery, and creativity dedicated to improving the quality of life and the human condition within the Commonwealth of Virginia and throughout the world." Improving the quality of life and the human condition applies to many facets of the University, including environmental sustainability. The Energy Patterns and Trends Electronic Database provides an authoritative resource on Virginia energy consumption. This supports the Virginia Department of Mines, Minerals and Energy and the Virginia Center for Coal and Energy Research in "responding to information requests from the general public and legislative bodies."²

Virginia Tech's Division of Campus Planning, Infrastructure, and Facilities' Office of Energy Management has established energy efficiency design guidelines to reduce electric and water usage during facility construction on campus.³ They have also developed a 5 Year Energy Action Plan that ended in 2020 and supported the current iteration of the Virginia Tech Climate Action Commitment, which aims to set the university on a path to carbon neutrality by 2030. Virginia Tech releases Sustainability Annual Reports⁴ to track progress on various sustainability projects. Progress is measured using the The Sustainability Tracking, Assessment & Rating System⁵ from the Association for the Advancement of Sustainability in Higher Education.

Virginia Tech has a responsibility to engage with our history of industrial energy extraction and build better sustainability strategies into each aspect of our university. While also being a campus building consuming similar energy to other facilities on campus, VTUL is unique in its management of multiple stores of data in our institutional repositories, digital libraries, Special Collections and University Archives, and data repository. It is with this history and context in mind that we explore the current environmental impact of our digital library choices and recommendations for decreasing this impact through changes in our workflows.

III. METHODOLOGY

A. Appraisal and Digitization

Newly created digital collections at VTUL are mediated by a team of stakeholders from across library departments who review project proposals. Once approval, projects are managed by a dedicated Digital Imaging Coordinator. The core goal of content creation in the Digital Imaging Lab is to create Preservation Digital Objects (a TIFF) to serve as a surrogate to the original object. These goals are informed by the Federal Agencies Digital Guidelines Initiative (FADGI), Metamorfoze and ISO imaging guidelines. This includes not only resolution (PPI) and sharpness (sampling efficiency) requirements of the above standards but also the color accuracy and tonal accuracy requirements. By following FADGI guidelines, the Digital Imaging Lab strives to achieve consistent, repeatable, measurable digital files in an efficient and scalable manner. The FADGI Standard contains specific technical guidelines for a variety of formats. Below are the general guidelines for the TIF files captured in the Digital Imaging Lab which represent the majority of output as stored data.

Preservation File TIFF

File Type: Uncompressed TIFF

Color Depth: 24 bit Color RGB

File Compression: None

Bit Depth: 16 bit

PPI: 400

Color Profile: AdobeRGB (1998)

The latest approved revision of the FADGI guideline does not explicitly address environmental sustainability in the creation of preservation standards. On the limitations of its guidelines, the initiative defers that its quality standards are "...appropriate for most cultural heritage imaging projects, and takes into consideration the competing requirements of quality, speed of production, and cost [12]."

The Digital Imaging Lab has a production server that is backed up nightly, and upon completion, transfer the working file to the appropriate department for either metadata cleanup or deposit into the Digital Library Platform. With the variety of

² Virginia Energy: <https://vept.energy.vt.edu/index.html>

³ VT Energy Efficiency Design Guide: <https://www.facilities.vt.edu/energy-utilities/energy-reduction-efforts/energy-efficiency-design-guidelines.html>

⁴ VT Sustainability Annual Reports:

<https://www.facilities.vt.edu/sustainability/sustainability-reports/virginia-tech-sustainability-annual-reports.html>

⁵ STARS: <https://stars.aashe.org/>

projects, some which may be hosted and managed by VTUL, and some that may not, there is a likelihood of redundancy in the transfer of ownership. More copies in more places is a tenant of digital preservation, but where do diminishing returns in the realms of security and preservation cross into harmful environmental practices?

B. Fixity

In addition to evaluating archival practices, we evaluated our digital preservation choices regarding fixity, including frequency and algorithm, and storage, including number of copies and general redundancy, and their relationship to one another. Both fixity checking and mid to long-term storage are ongoing services that result in continued energy consumption. Everything in a digital preservation and access system is by name, digital, and therefore requires some form of power. Ingest, fixity, restoration, migration, distributed storage, virus checking, file format verification, access, are all functions we include in our preservation system. When we evaluate the balance between what is important to us in our preservation system, we find that fixity and distributed storage are both necessary functionalities that may also allow for flexibility that could help decrease our carbon footprint. We choose these factors because we may not be able to control factors like necessity to migrate and number or frequency of access, but we can choose fixity frequency, appraisal of content, and the number of copies we choose to maintain.

According to the 2017 NDSA Fixity Survey, 84.1% of respondents indicated that they did utilize fixity information at some point in their workflows, though the methods, schedules, and reasons are widely varied [13]. Many digital preservationists have agreed that checksum computations are an intensive energy activity [1], and may not need to be performed as frequently as the field has been practicing [14]. This is because fixity checks need to open and read the entire file to produce an accurate checksum. While there is consensus that fixity should be performed regularly, neither the NDSA Levels of Preservation [15] nor the DPC's Digital Preservation Handbook [16] provide a best practice on the optimal frequency for scheduled fixity checks, but agree that

any situation where a file is moved from one location to another should always have a fixity check. More frequent fixity checking leads to faster repair, but is energy intensive and can be cost-prohibitive especially in the cloud environment [17]. Comparatively, LOCKSS runs continuous fixity checks and uses a non-canonical fixity store [18], which requires less bandwidth as it relies on the multiple copies to self-heal rather than retrieving the entire document for a fixity check [14] to notify a manager of an error.

The Virginia Tech Digital Library Platform generates fixity at multiple points in the data lifecycle; pre-ingest, on ingest, and on a regular schedule.⁶ We have two local servers, one of which is synced to Amazon Web Services (AWS) nightly, and one as-needed. We use the MD5 hash⁷ because this is what AWS requires. Currently our AWS instance is set to run fixity on ingest and every 90 days. Our preservation storage services are the Academic Preservation Trust (APTrust)⁸ and the MetaArchive Cooperative,⁹ both with their own independent fixity policies. MetaArchive is built on LOCKSS, which runs fixity as needed in a non-canonical, self-healing fixity store [17]. We also use Figshare¹⁰ to store our data repository. Figshare contracts with Chronopolis for preservation, and we ingest our datasets into APTrust.

The following section will refer to several energy units including millijoule (mj), watt-second (W*s), watts-hour (kWh), and megatonne (MT). It will also refer to carbon dioxide equivalent as CO₂e. With an understanding of our fixity triggers and frequency, we investigated the estimated energy consumed from generating an MD5 hash. In a study examining energy measurements of standard security functions, [19] found that of a series of hash algorithms they explored, MD4 and MD5 were the least energy-consuming hash algorithms. This study examined the type of hash and the size of file, noting that "consumption increases with the size of the files." They found that a hash for a 10kb file consumed approximately 5mj and grew to approximately 40mj for a 1mb file. Energy consumption is also dependent on the energy source, meaning coal, natural gas, petroleum, or other, with coal having the highest impact at 54% of

⁶ Fixity Policy:

<https://apps.es.vt.edu/confluence/display/LIBDPLD/Fixity+Policy>

⁷ MD5 Message Digest algorithm:

<https://datatracker.ietf.org/doc/html/rfc1321>

⁸ APTrust: <https://aptrust.org/>

⁹ MetaArchive: <https://metaarchive.org/>

¹⁰ Figshare: <https://figshare.com/>

energy in the United States in 2020 [20]. The schedule of fixity checking also affects energy consumption, as running ongoing tasks during peak hours will consume more energy than running them during off-hours, such as in the middle of the night.

In an analysis of quality and energy efficiency in hashing algorithms of mobile devices, [21] highlighted the importance of low-energy hash functions' effect on battery life and found a 29% difference in battery life between choosing the highest and least energy-consuming hash. They concluded that changing the algorithm to reduce energy consumption without losing security functionality is possible. Reference [22] noted that Reference [21] did not focus on the energy consumption of hashing "from an algorithmic perspective" [22] but also concluded that MD5 is the least energy consuming algorithm. We applied this research to our own fixity practices.

C. Storage

The energy consumption of fixity checking is intertwined with digital storage choices and the number of copies. Storage is a necessary but energy-exhaustive component of preservation systems. Robust digital preservation means distributed digital preservation storage, preferably with administrative diversity, and multiple copies. The NDSA Levels of Preservation V2 recommends a minimum of 3 copies [15] and LOCKSS maintains 5-7 copies.¹¹

AWS is one of VTUL's primary storage locations. AWS claims to have a 72% reduction of carbon emissions from their data centers when compared to other enterprise data centers [23]. They have instituted multiple initiatives for renewable energy, water stewardship, supporting other organizations to increase their own sustainable initiatives.¹² Reference [24] and a team of researchers have attempted to test these claims by building a dataset of CO₂e emissions from AWS's EC2 hardware to attempt to estimate the impact of EC2 hardware on carbon emissions. They found that it was difficult to measure the distribution of emissions over time due to the limited lifespan of a server, but ultimately produced a dataset available for revalidation and

manipulation. Others have claimed that cloud computing and storage is significantly more energy-consuming than saving to a disk [25], but that any security-driven disk server will consume more energy than an energy-saving disk server [26].

Other similar work in determining the electricity usage of a storage system is at the University of Houston Libraries where Bethany Scott inventoried all of the hardware components of their access and preservation infrastructure [27]. She concluded that focusing on ZFS fixity checking and decreasing file format resolutions would be the best way to optimize their local hardware to decrease environmental impact.

D. Limitations

This paper is scoped to archival appraisal, digitization workflows, fixity frequency, and storage options. We are not exploring the energy consumption of migrations, data transfers, VTUL hardware energy consumption; we are also not examining other cloud computing actions that occur, although there is significant interest in green computing.

We are using approximate numbers to determine a broad sense of approximate impact that is not based on hard numbers and relies on others' research. Our paper is highly qualitative and meant to provide direction for exploring changes in our digital library practices. Isolating our research to the defined scope may alter the ultimate environmental impact of our practices, but still provides insight on what we may be able to modify in the short term.

IV. RESULTS

A. Appraisal and Digitization Methods

Using the plainly-stated charge of Pendergrass et al. to reduce digital content overall, appraisal and digitization practices are areas which should be scrutinized. A well defined collecting policy will help control the scope and prioritize the collecting efforts. The Special Collections and University Archives at VTUL has a mission to provide access to materials in their original form, and to offer materials in digital format "when possible".¹³ This language has an

¹¹ LOCKSS FAQ: <https://www.lockss.org/about/frequently-asked-questions>

¹² AWS Sustainability: <https://sustainability.aboutamazon.com/environment/the-cloud?energyType=true>

¹³ VT Special Collection and University Archives: <https://spec.lib.vt.edu/about/index.html>

allowance for familiar constraints to cultural heritage institutions such as time and funding. It may be beneficial to directly name environmental considerations in a future revision. The proliferation of born-digital collections presents an amplified challenge, and may require a modified collecting policy to address sustainability.

Clarity in how collections are prioritized internally for digitization is another area to address. VTUL has an Advisory Council for Digital Collections which prioritizes library and community projects for the Digital Imaging Lab. The committee's rubric for selection focuses on mission-specific projects and works through requisite technical details. This committee could be a logical check on unsustainable digital projects in the pipeline. This scrutiny should also exist within submitting library departments prior to review by the Advisory Council. Departments should clarify how collections are prioritized internally. Selection decisions may be made around privacy, access restrictions, copyright, uniqueness, as well as time and effort required. The impact of a project of sustainability goals should be given ample consideration in this list. It may also be beneficial to create a list of collections that specifically will not be digitized.

Among digitization practices, organizations should identify areas where changes can be made. In some cases this will mean going against industry standards of resolution or bit depth. While the biggest results will come from a reevaluation of standards contributing to file size, simple cleanup to digitized material can play a role in sustainability goals. For visual materials, this could mean addressing duplicate or blank pages. For audio-visual materials, editing dead air and trimming commercial/non-relevant content from digitized sources prior to repository ingest is valuable work.

B. Fixity Estimations

Given our context in the libraries, we assume most of our files will be on the larger end of the range tested by Fournier et al. [18]. If we operate under the assumption that the average MD5 hash consumes at least 40mj per hash, or 0.05 W*s, this equates to 1.11111111E-8 kW*h. This is too small to translate to CO₂e emissions, but if we calculate 1 terabyte (TB) of content, we get the following approximate results in Table 1.

Table I
Estimated energy consumption and carbon emission of hashing 1TB of data

Size	Millijoule	Watt-second	Watt-hour	Kilogram (kg)
1 MB	40mj	0.05 W*s	1.11111111E-8 kW*h	--
1 TB / 1,000,000 MB	40,000,000mj	40,000 W*s	11.1111111 kW*h	2.59 kg CO ₂ e

The number of storage locations, varying workflows, varying fixity frequencies, and general flow of storage to make exact calculations difficult. If we simplify it to whole numbers as our total TB and take into account the following table of each Virginia Tech storage location and the number of approximate TBs in each location, we find the following approximation for running fixity one time on each storage space in Table 2.

Table II
Estimated carbon emissions of VTUL storage spaces based on size

Storage Location	Fixity Freq	Size in TB	Kilogram CO ₂ e
Local high speed server	Nightly	11	28.59 CO ₂ e
Local NAS server	Nightly	10	25.9 kg CO ₂ e
AWS East Region	Every 90 days / ~4 times a year	1	2.59 kg CO ₂ e
AWS West Region	Every 90 days / ~4 times a year	.5	1.3 kg CO ₂ e
APTrust	Every 90 days / ~4 times a year	6	15.54 kg CO ₂ e
MetaArchive (LOCKSS)	Every 90 days / ~4 times a year	5	12.95 kg CO ₂ e
			86.87 kg CO₂e

The final result is simply, the environmental impact of running fixity is very complicated to define. Our results are extremely broad and validating these results would involve a time-intensive research study in collaboration with our IT division, vendors, and preservation vendors, which is a goal that we currently do not have the support or bandwidth to perform. Despite this, if our final result of 86.87 kg CO₂e for running a single fixity check on all of our approximate data is even close to accurate, this is cause for concern and an impetus to refine our workflows.

C. Storage Considerations

VTUL has designated 4 levels of preservation.¹⁴ Not all content will be maintained at all levels. In terms of the number of storage locations, the levels are as follows: Level 0 is no preservation action taken; Level 1 basic preservation is 1-2 local copies, 1 cloud copy; Level 2 extended preservation is 2 local and 2 cloud copies; and Level 3 Advanced preservation is 2 local copies, 2 cloud copies, and ingest into one of our two distributed storage locations, APTTrust or MetaArchive. Most of our content is designated at a Level 2.

To review, VTUL uses a combined storage system of the following contracted storage vendors and the approximate number of copies as shown in Table 3.

Table III
Overview of VTUL storage locations

Storage Location	Medium	Geo-graphical Location	Purpose	Copies
Local NAS	disc	Virginia	Working/staging server	1
AWS East Region	Cloud	Virginia	Primary cloud storage	1
AWS West Region	Cloud	Oregon	Secondary cloud storage	1
APTrust	Cloud	Virginia, Oregon	Preservation storage, admin diversity	3
MetaArchive (LOCKSS)	various disc	varies	distributed preservation	5
Chronopolis (Figshare)	Cloud + disc	Virginia, Oregon, California	Figshare preservation	1

This list is not scoped to include additional data points, such as our institutional repository which is run on a local DSpace instance, our learning object repository in Omeka, any other Omeka instances VTUL hosts, our Confluence spaces, GitHub instance, or our Google Drive storage. We clearly rely heavily on AWS for our own primary and secondary storage, through APTTrust, and indirectly through Chronopolis. The question we asked ourselves was whether administrative diversity between multiple services benefited us enough on a security level to justify this reliance and the number of copies we

maintained. Lots of copies and lots of checksums do keep stuff safe, but can we articulate the value of these choices and still account for the carbon footprint?

Our answer is yes, but with modifications. Our local servers are for creating and staging content for ingest and our AWS serves as both a backup for our disc servers as well as access and preservation for our digital library. All of our preservation options are for geographically distributed preservation storage, administrative diversity, and technology diversity, all of which are considered good practice in the digital preservation community. Actually defining the environmental impact of all of our storage locations is complicated due to the various workflows, number of copies, distribution of copies, and independent needs of the collection. We commit to the number of copies we maintain and the storage we have chosen, but the amount of space and energy we consume is dependent on our appraisal system, both pre-digitization and for preservation. We also found that we have not determined what Reference [1] describes as acceptable loss - the "level of acceptable loss in collection under [our] care" [1] to make better use of what resources we do have.

AWS recently released a new feature called the Customer Carbon Footprint Tool, available to all customers. This tool allows users to track their carbon emissions over time and over geographic location, specially measuring Scope 1 and Scope 2¹⁵, or direct emissions and indirect emissions, of content in AWS [28]. With the aid of our digital library's Software Engineer, we obtained results from our development server in AWS from January 2020 through November 2021. The results as seen in Figures 1 and 2, indicates that we emitted 0.6 MTCO₂e and claims that we have saved 0.4 MT CO₂e as compared to "on-premises computing equivalents." S3 is the feature generating the most carbon emissions. As it is a new feature we are still learning how to read the information and understand the true impacts of the numbers, and we will continue to monitor it as we increase activity in AWS.

¹⁴ VTDLP Preservation Policy: <https://apps.es.vt.edu/confluence/display/LIBDPLD/VTUL+Preservation+Policy>

¹⁵ EPA: <https://www.epa.gov/climateleadership/scope-1-and-scope-2-inventory-guidance>

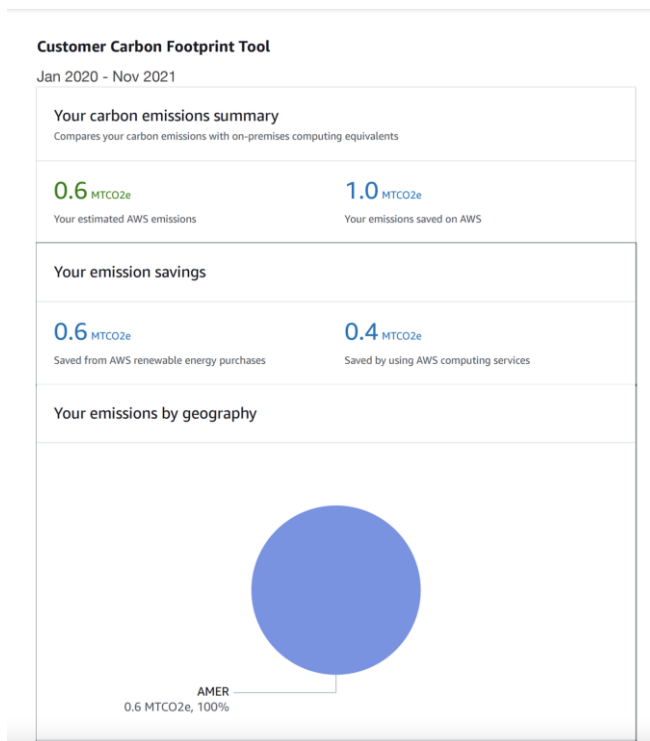


Figure 1 Virginia Tech's AWS carbon emissions summary

Service	Carbon emissions	%
EC2	0.1 MTCO ₂ e	16.67%
S3	0.4 MTCO ₂ e	66.67%
Other	0.1 MTCO ₂ e	16.67%
Total	0.6 MTCO ₂ e	100%

Figure 2 Virginia Tech's AWS carbon emissions by service

One general concept that we encountered in our work is that energy efficiency is predominantly measured by financial cost rather than environmental cost. The issue with increasing environmentally friendly systems is that it can have little to no impact on cost [29], which is a primary concern in most organizations and is often the focus of energy sustainability benefits over the environmental impact itself. The Customer Carbon Footprint Tool, for example, is found through the AWS Billing Console under Cost & Usage Reports, emphasizing financial cost. Cost is a major factor in all digital curation systems and cannot be overlooked, but it seems to be a mistake for us to only rely on financial cost as a measure of our energy sustainability.

V. RECOMMENDATIONS

Based on our preliminary results, we recommend the following actions to help reduce the carbon footprint of the VTUL digital library.

- **Include climate considerations in appraisal of digital collection projects:** The long term environmental impact of digitizing a collection should be considered alongside other factors in collection selection.
- **Revist collection policies and institutional mission:** We recommend adjusting an existing collection policy or mission to reflect a commitment to sustainably manage digital resources and guide future decision making.
- **Decrease redundancy of working files:** We recommend streamlining the transfer process to minimize the multi-department storage redundancy of working files. Understanding that redundancy is a necessity, determine which stages of the collection management process should be the most secure. Schedule a process for deletion after migration and quality assurance.
- **Reduce ongoing fixity checks:** We recommend reducing scheduled fixity checks of all AWS objects from every 90 days to every 120 days or possibly more, increase spot-checking fixity from a randomly selected subset of files in each digital collection, and to increase test restorations to account for the decreased fixity checking.
- **Determine acceptable loss:** Reducing security will reduce energy consumption. We need to determine acceptable loss for each of the storage vendors we contract with and alter our workflows with mechanisms for faster healing to compensate for any loss.
- **Preservation appraisal:** We have defined our own levels of preservation, but we recommend modifying them to include more direct appraisal strategies and a determination of acceptable loss for each level.
- **Investigate smaller object sizes:** The size of a digital object directly impacts the energy consumed in running fixity, transferring between storage locations, and ongoing storage maintenance. We recommend exploring collections or data types where there are options for creating lower resolution or otherwise smaller objects.
- **Sustainability commitment:** As an organization, we recommend that VTUL develop a Sustainability Statement for the Digital Libraries at VTUL to scope our work

and to emphasize not only the importance of but the immediate need for greener digital library curation strategies.

- **Community training:** The Libraries are responsible for keeping up with digital trends and practices and educating the University and larger community. We recommend regular Professional Development Network training sessions on ensuring good practices in personal and professional archiving that also emphasize environmental sustainability.

VI. NEXT STEPS

There are several next steps we want to pursue after this preliminary research. Exploring time and money spent on all of these steps to reinforce the areas where we need improvement on multiple levels. We also hope to explore other preservation activities including migration, restoration, transfer and syncing, file format verification, alternate storage opportunities, and appraisal.

REFERENCES

- [1] K. Pendergrass, W. Sampson, T. Walsh, and L. Alagna, "Toward environmentally sustainable digital preservation," *The American Archivist*, vol. 82 no 1, pp. 165-206, 2019. <https://doi.org/10.17723/0360-9081-82.1.165>
- [2] United Nations, "United Nations Framework Convention on Climate Change," United Nations Treaty Collection, 1992. https://treaties.un.org/pages/ViewDetailsIII.aspx?src=TREATY&mtdsg_no=XXVII-7&chapter=27&Temp=mtdsg3&clang=en
- [3] C. Durham, "The necessity of environmentally sustainable digital preservation and its effects on preservation workflow," Johns Hopkins University Sheridan Libraries, 2019. <https://jscholarship.library.jhu.edu/bitstream/handle/1774.2/62122/Durham%20Curtis%20.pdf?sequence=1&isAllowed=y>
- [4] H. Stehkämper, "'Natural' air conditioning of stacks," *Restaurator* 9, no. 4, pp. 162-177, 1988. doi.org/10.1515/rest.1988.9.4.163.
- [5] Rowoldt, Sandra. "Going archivally green: Implications of doing it naturally in Southern Africa archives and libraries," *South African Journal of Libraries and Information Science*, vol. 66, no. 4, 2014. doi:10.7553/66-4-1425.
- [6] Society of American Archivists, "SAA Core Values Statement and Code of Ethics," SAA, 2020. <https://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics>
- [7] Digital Preservation Services Collaborative, "Digital Preservation Declaration of Shared Values." https://dpcollaborative.org/shared-values_en
- [8] National Archives and Records Administration, "Climate action plan," National Archives, 2021. <https://www.archives.gov/files/about/plans-reports/sustainability/climate-action-plan.pdf>
- [9] B. Goldman, "It's not easy being Green(e): Digital preservation in the age of climate change," in *Archival Values: Essays in Honor of Mark Greene*, Society of American Archivists, 2018. <https://scholarsphere.psu.edu/resources/381e68bf-c199-4786-ae61-671aede4e041>
- [10] W. R. Hibbard, *Virginia Coal: An Abridged History*, Virginia Polytechnic Institute and State University, Ed. T. J. Clutter, 1990, pp. 20-24. <http://hdl.handle.net/10919/90196>
- [11] R. Freis, "The day the earth shook 50 years later, Montgomery's biggest tragedy resonates still," in *The Roanoke Times*, 18 April 1996, pp. 1. <https://scholar.lib.vt.edu/VA-news/ROA-Times/issues/1996/rt9604/960418/04180017.htm>
- [12] Federal Agencies Digital Guidelines Initiative, "Technical Guidelines for Digitizing Cultural Heritage Materials," September 2016. <http://www.digitizationguidelines.gov/guidelines/FADGI%20Federal%20Agencies%20Digital%20Guidelines%20Initiative-2016%20Final%20rev1.pdf>
- [13] S. Barsness, A. Collie, M. Gallinger, C. Kussman, S. Schaefer, and G. Truman, "2017 fixity survey report," National Digital Stewardship Alliance, 2017. https://ndsa.org/documents/Report_2017NDSAFixitySurvey.pdf
- [14] D. Rosenthal, "How few copies?," DSHR's Blog, April 2016. <https://blog.dshr.org/2016/04/how-few-copies.html>
- [15] Digital Preservation Coalition, "Digital preservation handbook," 2nd Ed., 2015. <https://www.dpconline.org/handbook>
- [16] J. Bailey, "File fixity and digital preservation storage: More results from the NDSA storage survey," *The Signal*, 2012. <https://blogs.loc.gov/thesignal/2012/03/file-fixity-and-digital-preservation-storage-more-results-from-the-ndsa-storage-survey/>
- [17] N. Taylor, "Lots of checksums keep stuff safer," CNI Spring Membership Meeting Proceeding, 2019. https://www.cni.org/wp-content/uploads/2019/04/CNI_Flexibility_Taylor.pdf
- [18] B. Fournier, V. Tong, and G. Guette, "Accurate measurement of the energy consumption of security functions," Scitepress, 2021. <https://www.scitepress.org/Papers/2021/105446/105446.pdf>
- [19] U.S. Energy Information Administration, "How much of U.S. carbon dioxide emissions are associated with electricity generation?," *Frequently Asked Questions*, 2020. <https://www.eia.gov/tools/faqs/faq.php?id=77&t=3>
- [20] R. Damasevicius, G. Ziberkas, V. Stukys, and J. Toldinas, "Energy consumption of hash functions," *Elektronika ir Elektrotechnika*, vol. 18, no. 12, pp. 81-84, April 2012. <https://pdfs.semanticscholar.org/7f43/7af19922bda1fe845b663290440e359d21ff.pdf>
- [21] P. Harish, "Towards designing energy-efficient secure hashes," University of North Florida Digital Commons Graduate Theses and Dissertations, 2015. <https://digitalcommons.unf.edu/etd/598/>
- [22] National Digital Stewardship Alliance, "Levels of digital preservation," version 2.0, 2019. <https://ndsa.org/publications/levels-of-digital-preservation/>
- [23] D. Bizo, "The carbon reduction opportunity of moving to Amazon Web Services," AWS, October 2019.

https://sustainability.aboutamazon.com/carbon_reduction_aws.pdf

- [24] B. Davy, "Building an AWS EC2 carbon emissions dataset," Medium, September 2021. <https://medium.com/teads-engineering/building-an-aws-ec2-carbon-emissions-dataset-3f0fd76c98ac>
- [25] J. Adamson, "Carbon and the cloud," in Stanford Magazine, Ed. N. Mmonatau, May 2017. <https://stanfordmag.org/contents/carbon-and-the-cloud>
- [26] S. Yin, M. I. Alghamdi, X. Ruan, M. Nijim, A. Tamilarasan, Z. Zong, X. Qin, and Y. Yang, "Improving energy efficiency and security for disk systems," in 2010 12th IEEE International Conference on High Performance Computing and Communications, 2010. DOI 10.1109/HPCC.2010.26
- [27] B. Scott, "Estimating energy use for digital preservation, part 1," Bloggers!, October 2020. <https://saaers.wordpress.com/2020/10/06/estimating-energy-use-for-digital-preservation-part-i/>
- [28] J. Barr, "New – customer carbon footprint tool," AWS News Blog, 2022. <https://aws.amazon.com/blogs/aws/new-customer-carbon-footprint-tool/>
- [29] D. Rosenthal, "Why is green preservation hard?" in Green Bytes: Sustainable Approaches to Digital Stewardship at the National Digital Information Infrastructure and Preservation Program partners meeting, July 2013. https://digitalpreservation.gov/meetings/documents/ndiipp_13/Rosenthal.pdf

AIN'T NO MOUNTAIN HIGH ENOUGH:

Developing a New Competency Framework for Digital Preservation

Sharon McMeekin

Digital Preservation Coalition
Scotland
sharon.mcmeekin@dpconline.org
[0000-0002-1842-611X](tel:0000-0002-1842-611X)

Amy Currie

Digital Preservation Coalition
Scotland
amy.currie@dpconline.org
[0000-0001-9099-8457](tel:0000-0001-9099-8457)

Abstract – A skilled workforce is essential to successful digital preservation. But how do we define what “skilled” means? Attempts to define knowledge and competencies for digital preservation have, so far, largely focused on the issue from the point of view of educators. This paper describes work carried out by the Digital Preservation Coalition to create a new competency framework that can be deployed for a range of purposes, including facilitating recruitment, structuring professional development, auditing skills, and reviewing curricula.

Keywords – skills, training, staffing, education, professional development

Conference Topics – Community; Resilience

I. INTRODUCTION

In her iPres 2019 paper “People Get Ready: Building sustainability into digital preservation workforce development”, Sharon McMeekin stated that “A skilled workforce is essential to digital preservation and should be [...] a key part of strategies for development” [1]. She identified the failure to clearly define “digital preservation practitioner” as a distinct profession as a key barrier to developing successful pathways for workforce development in the field.

The paper finished with a call to improve and expand on the current digital preservation workforce development resources. This included support for recruitment, training aimed at new and more advanced audiences, increased collaboration and sharing of knowledge and resources, and the

creation of a digital preservation competency framework that reflected current good practice.

This paper provides an overview of a project undertaken by the Workforce Development team of the Digital Preservation Coalition (DPC) to develop a competency framework as described in the 2019 paper. It will start by offering context for the activity through a brief overview of the community's previous efforts to define digital preservation competencies and curricula. The paper will go on to describe the methodology used to research and design the new competency framework, which brought together the insights of previous work and good practice guidance from key models such as the NDSA Levels of Digital Preservation¹ and the DPC's Rapid Assessment Model (DPC RAM)². It will then detail the competency framework that has been developed and accompanying resources to facilitate its practical use. The paper concludes with ideas of future work that will continue to expand the suite of resources to support recruitment for digital workforce development.

II. DIGITAL PRESERVATION REQUIRES SKILLED STAFF

In their seminal 2003 essay “The Five Organizational Stages of Digital Preservation”, Anne Kenney and Nancy McGovern were among the first to highlight the dangers of considering digital preservation as simply a technological problem [2]. They championed a balanced approach that gives equal consideration to issues relating to both organizational context and resourcing, with skilled

¹ <http://doi.org/10.17605/OSF.IO/QGZ98>

² <http://doi.org/10.7207/dpcram21-02>

personnel and defined responsibilities being a key component. In the years since, there has been much discussion about, and several attempts to define, the skills, knowledge, and competencies required to facilitate digital preservation activities and fulfil this need for skilled personnel.

The breadth of skills required has been a constant issue discussed by those developing resources in this area. For example, Fulton, Botticelli, and Bradley (2011) identify the importance of understanding digital preservation to be a “distinctly interdisciplinary undertaking” when developing the curriculum for their Digital Information Management (DigIn) graduate certificate program [3]. They also discuss the importance of information professionals acquiring “a common foundation of technological literacy” to allow them to effectively undertake digital preservation activities and to collaborate with colleagues, particularly those who specialize in relevant areas of information technology.

Likewise, the “Matrix of Digital Curation Knowledge and Competencies” developed by Cal Lee and colleagues in 2009 to help with “identifying and organizing the material to be covered in a digital curation curriculum” shows that digital preservation requires a complex array of skills from different disciplines [4]. The Matrix includes a diverse list of skills and knowledge from areas including administration, advocacy and communication, information management, legal considerations, and a broad range of technological activities. The Matrix is comprehensive in its coverage, but it is also focused on what should be included in a digital preservation curriculum which makes it difficult to parse if trying to identify the skills needed by a digital preservation practitioner.

Aiming to bridge the gap between course curricula and professional training and development needs, the 2013 Digital Curator Vocational Education Europe Project (DigCurV) developed a framework with three “lenses” onto the competencies required at different stages of a digital preservation career. Indeed, Molloy, Gow, and Konstantelos share that the project “aimed to address two types of vocational training: for those aiming to enter the profession (including Master’s-level qualifications), and for existing staff (such as in-house skills training or CPD provided by professional organisations)” [5]. The DigCurV framework also explicitly recognizes the

complexities of the skills required for digital preservation noting that for “successful professional performance, staff must demonstrate domain-specific and technical competencies, generic professional and project skills, and personal qualities in a blend appropriate to their particular professional context”. The framework was described as an aspirational model and that they did not “expect an individual [...] to possess every skill, ability or piece of knowledge enumerated in the Framework”.

The competency framework developed by the DigCurV project has been a popular touchstone for those interested in education and training for digital preservation since its publication in 2014, but even at that time, its authors suggested further work that should be undertaken. This included defining “a set of core knowledge and skill elements” from across the “Practitioner”, “Manager”, and “Executive” lenses defined within the framework. The framework also represents a snapshot of digital preservation practice at the time of its development and is now a step or two behind current good practice, suggesting updates or a new framework would be desirable.

In another attempt to define the skills for digital preservation, Blumenthal et al.’s 2016 paper “What Makes a Digital Steward: A Competency Profile Based on the National Digital Stewardship Residencies”, aimed to provide “a profile of the skills, responsibilities, and knowledge areas that define competency in digital stewardship” [6]. The authors list seven categories of competence they derived from an analysis of the project proposals of the National Digital Stewardship Residences and a survey of participants. The seven categories are as follows:

1. Technical Skills
2. Professional Output Responsibilities
3. Communication Skills
4. Research Responsibilities
5. Project Management Responsibilities
6. Knowledge of Standards and Best Practice
7. Personality Requirements

In addition to defining these categories, they also captured data on the importance of different skills within respondents’ organizational contexts. Responses showed that skills relating to

communications, project management, research, and knowledge of standards and best practice were more important across the cohort than technical skills. As with previous work on digital preservation skills, the emphasis is on the need for professionals with interdisciplinary competencies. Additionally, they found that “being an effective steward of digital material requires more extensive and specialized training than can be acquired through traditional means”.

The need for a clear definition of the skills required for digital preservation to facilitate professional development was further confirmed by the survey results detailed in the “Staffing for Effective Digital Preservation 2017” report from the NDSA. Watkins et al. share that 68% of organizations who responded to the survey source some or all of the staff to work on digital preservation from their existing staff complement, retraining them to work in the area [7]. Without knowing what skills are required, how can these staff be successfully trained to carry out their new responsibilities effectively? Their survey results also upheld previous descriptions of digital preservation practitioners as multi-faceted professionals who are required to have a wide-ranging skillset. Knowledge of standards and best practices for digital preservation was amongst the most important identified by survey respondents, but the other high-ranking skills were all of a more generic nature. These included communications, project management, and collaboration skills.

Finally, in their 2020 article “What’s Wrong with Digital Stewardship”, Blumenthal et al. list ideas shared by the practitioners they interviewed for better orientating digital stewardship towards its ultimate goal [8]. These included the need to “reorganize existing staff, change existing job descriptions, redistribute responsibility for digital stewardship, and implementing more effective decision- and policy-making protocols”. These ideas closely mirror requests the DPC has received from its members to help support their workforce development. Members have indicated the need for a resource that would support recruitment, develop job descriptions, and help structure ongoing professional development for staff.

Members have also indicated a need for a resource that would align with other DPC tools such

as the Rapid Assessment Model (DPC RAM), reflect current good practice, offer an optimal balance of detail so that it would be widely applicable but still be useable, and would be flexible enough to be used for a number of different purposes. With these goals in mind, we (the DPC Workforce Development team) began the development of a new competency framework for digital preservation.

III. METHODOLOGY

We employed a team-based approach and iterative, agile methodology, gathering and analyzing qualitative data from previous work on digital competencies and drawing from the experiences and expertise of those working in digital preservation. There were three main phases of research and development: (1) in-depth assessments of existing literature and resources with identification of key skills, knowledge or competency areas relating to digital preservation, (2) a series of concept mapping exercises for framework design and development, and (3) an iterative feedback and review process with other DPC colleagues, DPC Members, and through a pilot of the accompanying resource, the Competency Audit Toolkit (DPC CAT).

A. Phase One – Research and Data Collection

We began with a short but intensive phase of qualitative research and data collection, gathering relevant articles and resources on digital competencies and curricula to compile a shared reading list. We then each conducted in-depth reviews of those readings. This assessment took place from July to August 2021 and involved separate readings and analysis; we each identified, collected, and assessed direct or indirect references to digital preservation skills. Following these individual reviews, we held a face-to-face meeting in August to compare our findings and compile a preliminary list of common skills, knowledge, and competencies relevant to digital preservation based on the discussion.

B. Phase Two – Concept Mapping and Design

Our next phase of research focused on developing and designing the framework itself, conducting a series of concept mapping (or mind mapping) exercises for further analysis and structuring of the framework. The first mapping exercise took place in August during the face-to-face meeting. The aforementioned list of skills, knowledge and competencies were written onto

post-it notes, then arranged and rearranged into groups. Our first version of the framework, drafted from this exercise, resulted in 37 distinct skills elements arranged into seven overarching skill areas. These elements and areas were entered into a spreadsheet to record them and to allow us to input more detail by adding corresponding example statements and example activities to clarify their meaning and practical applications. Following the completion of these additions, a second mapping exercise was conducted in September, resulting in a more expanded scope, and structuring and 74 skill elements under six skill areas.

The use of concept mapping exercises at various points throughout the research proved useful for the early stages of framework design; not only did they help identify interrelationships of skills required for digital preservation, and skill areas and elements where overlaps occur, but they also provided a way to structure and present early drafts of the framework in a meaningful way to facilitate feedback and refining of findings.

C. Phase Three – Review and Refinement

From November 2021 to January 2022, drafts of the framework were shared with DPC colleagues for iterative review and refinement based on their feedback. In light of the feedback received, a final mapping exercise was completed in February, resulting in a revised version, with 27 skill elements listed under five competency areas.

In June 2022, this version of the framework and the accompanying Competency Audit Toolkit (DPC CAT) were shared with DPC Members through a members-only preview on the DPC website and an online webinar. Additionally, a pilot of DPC CAT was conducted with five DPC Member organizations to gather practical feedback to help further refine the resource before its public release. Participants were asked to complete a competency audit process for their organization using DPC CAT during June and July and then provide general feedback on how it progressed, what went well, and what could be improved.

The five DPC Member organizations participating in the pilot were based in different locations (in Europe, Asia, and Australasia) and encompassed a range of different organizational contexts (higher education, research, national collecting body, financial, governmental, interinstitutional). The

number of staff participating in each organization's competency audit process also varied (from one to eleven individuals).

The outcomes of the pilot were generally very positive and constructive. The feedback received from participants was supplementary rather than corrective. Recommendations included the addition of Framework tables into the CAT workbooks, additional explanatory text on differences when assessing digital preservation specific versus more generic skill elements, and the addition of a computer programming skill element under the information technology competency area. All of these recommendations were taken on and incorporated into the framework and CAT. There are also plans to create simple short "quick start" guides for using CAT in different scenarios following suggestions from DPC Members.

IV. A NEW COMPETENCY FRAMEWORK

The new framework [9] aims to be a reference point for anyone interested in understanding the skills required to undertake digital preservation activities. This might be an individual wishing to benchmark their own skills as part of planning their professional development, or when considering an advertisement for a post they would like to apply for. It could be an educator looking to evaluate the curriculum of a digital preservation course they teach. Or the framework may also be used by an organization revising job descriptions for staff, recruiting new employees, or auditing current skills across a team or department.

The framework presents information on the skills required for digital preservation in a hierarchical structure, from generic to granular, and aims to offer as much flexibility as possible for users. The information is organized into the following:

- Five high-level competency areas that offer an overview of and quick reference to the breadth of competencies required to undertake digital preservation work.
- Twenty-eight skill elements, organized in groups under the skill areas, which break down the knowledge and competencies into more clearly defined units.
- An example descriptive statement for each skill element to show how it might be defined in a job description or advertisement.

Table I
DPC Skills Framework

Competency Area	Skill Element No.	Skill Element
Governance, Resourcing, and Management	1	Policy Development
	2	Risk Management
	3	Resource Management
	4	Staff Management
	5	Strategy and Planning
	6	Analysis and Decision-Making
Communications and Advocacy	7	Effective Communication
	8	Collaboration and Teamwork
	9	Stakeholder Analysis and Engagement
	10	User Analysis and Engagement
	11	Advocacy
	12	Training
	13	Producing Documentation
Information Technology	14	General IT Literacy
	15	Computer Programming
	16	System Procurement
	17	Storage Infrastructures
	18	Information Security
	19	Workflow Development and Implementation
Legal and Social Responsibilities	20	Legal and Regulatory Compliance
	21	Environmental Impact
	22	Inclusion and Diversity
	23	Ethics
Digital Preservation Domain Specific	24	Metadata Standards and Implementation
	25	Information Management Principles
	26	Approaches to Preservation
	27	DP Standards and Models
	28	Managing Access

- Between three and seven example activities for each skill element to show how that element might be deployed in practice.

The five competency areas and twenty-eight elements included in the framework are shown in Table One. As with previous endeavors to define the skills required for digital preservation, the five competency areas represent a broad range of interdisciplinary skills, with only one of the five areas specifically referencing digital preservation knowledge and competencies. The other four competency areas cover issues relating to ensuring sustainable organizational infrastructures, communications, technological skills, and proactive management of legal and social consideration.

Echoing the structure and use of maturity models, five skill levels have also been defined, against which an individual might rate their experience and capabilities with regards to a particular skill element. These have been loosely aligned with the five levels of maturity defined in DPC RAM (Minimal Awareness, Awareness, Basic, Managed, and Optimized) [10]. The five levels of experience are as follows:

1. Novice - Limited awareness of the skill element.
2. Beginner - A basic understanding of the skill element. May have received some training, but little or no practical experience.
3. Intermediate - A sound understanding of the skill element and some experience of its practical application
4. Advanced - A thorough understanding of the skill element and significant experience of its practical application.
5. Expert - An in-depth understanding of the skill element and a leader in the development of approaches to its practical application.

As mentioned above, the framework has been structured as described to allow for flexibility in how it is used. In particular, that flexibility means that the framework might be deployed to understand and assess the skills needed in any context from an individual role through to all of the staff involved in an organization's digital preservation activities. With this in mind, it is important to note two pieces of guidance with regards to using the framework.

The first piece of guidance is that it is not intended that any individual member of staff should be competent in all of the skills included in the framework. Digital preservation is a collaborative undertaking and as such responsibilities for different areas of work, and the corresponding skills required to effectively fulfil those responsibilities, should be spread across a number of roles. When assessing skills for a particular role, this should be done in reference specifically to the skill elements that align with the related job description and/or the responsibilities and activities carried out by the individual in the role. Other skill elements are likely only to be considered as part of reformulating job descriptions or to facilitate the professional development of those looking to expand their current skill set.

The second piece of guidance relates to the level of experience required for each skill element. It is unlikely that there will be a requirement to reach expert level for all skill levels, either for an individual or across a group of staff. The appropriate level of skill to facilitate the organization's digital preservation activities should be identified and used as the benchmark against which to measure skills. For example, few practitioners will be required to become experts in the development of metadata standards and implementation. An intermediate or advanced level of knowledge of how metadata standards are deployed within their own organizational context is likely to be more than sufficient. Indeed, there are benefits to be gained from aligning the skill levels required within an organization with the results of a maturity modelling exercise, and this is facilitated by the accompanying Competency Audit Toolkit (DPC CAT).

It is also important to note that, as with maturity models such as DPC RAM, the information in the framework aims to be illustrative and not exhaustive. An early attempt in the development process to make the framework as thorough as possible resulted in a resource that was frankly too detailed and likely unusable in its complexity. Therefore, while it is hoped that the framework is relevant across the digital preservation community, some customization may be required for individual contexts.

Due to the complexity and interrelationships of skills required for digital preservation, there are also some skill areas and elements where overlaps occur. We spent a significant amount of the development

time attempting to successfully tease out the individual skill elements and decide under which skill area they should sit, so that the framework would be clear and usable. For example, there are clear links between the skills within the Information Technology area and the Digital Preservation Domain Specific Area. With this in mind, users of the framework may need to make their own judgements as to which skill(s) relate to a particular activity they undertake if it might be related to more than one skill area.

It is expected that the skills framework will continue to develop over time, as good practice within digital preservation continues to develop. Technological solutions will change and may require the development of new skills, and new areas of specialization may evolve. The DPC is committed to the continued management and development of the skills framework to ensure it remains relevant and usable for the digital preservation community. With this in mind, the DPC welcomes feedback from practitioners on the Competency Framework on how they used the framework, what worked well, and what could be improved¹.

V. ACCOMPANYING RESOURCES

Creating a framework that could be practically applied to a variety of workforce development issues was a key aim of this project. To facilitate practical implementation two accompanying resources have been developed: the DPC Competency Audit Toolkit (DPC CAT) [11] and a set of Example Role Descriptions [12].

DPC CAT has been developed with the support of the UK's Nuclear Decommissioning Authority² and provides practical structured processes for assessing competencies at individual and group levels. The toolkit itself contains three components: a guide to using the toolkit and two Excel workbooks, one for auditing both individual skills and role descriptions (the Individual Audit Workbook), and one for auditing the skills of a group of staff members (the Organizational Audit Workbook).

The individual audit allows a practitioner to benchmark their current skills, set targets for development, and plan activities and training to meet those targets. The process can be used independently or folded into ongoing review processes such as staff appraisals. The role

description audit provides a process for evaluating an existing role description in relation to the reality of day-to-day tasks and responsibilities carried out by a role holder. Results of this process might be used to provoke an update to an existing job description, or as evidence when making the case for additional staff or increased compensation. Finally, the organizational audit is directly linked to DPC RAM, allowing organizations to identify if they possess the required skill levels to support their current and target digital preservation capabilities. This is completed by entering the information from a RAM assessment exercise and results from individual skills audits, from which a report is generated indicating required skills levels, the current highest and average skill levels amongst staff members, and where gaps exist.

The second resource, the Example Role Descriptions, aims to provide an illustration of how the framework can be used to help formulate role descriptions for digital preservation roles. Each example role description identifies which skill elements are relevant, what skill level might be expected, and provides an example statement of how the skill element might be expressed with a role description. These resources are intended to be guides for those drafting role descriptions for current staff and potential new hires and should not be considered prescriptive. The role descriptions cover the following role types:

- Graduate
- Trainee
- Digital Preservation Officer
- Digital Preservation Archivist/Librarian
- Web Archivist
- Digital Preservation Developer
- Digital Preservation Program Manager
- Senior Executive/Administrator

These are the first of the resources developed to accompany the new framework and in the next section we will discuss some of the complementary resources that we hope will be developed in the near future.

¹bit.ly/CATFeedback

²<https://www.gov.uk/government/organisations/nuclear-decommissioning-authority>

VI. WHAT'S NEXT?

The DPC is committed to the continued support and development of the competency framework as part of its increasing suite of tools and resources for digital preservation continuous improvement.

A key planned resource will aim to help practitioners identify how to “level-up” in relation to particular skill elements. This will include information on suitable training courses, funding opportunities, or suggested tasks or projects they could undertake to gained practical experience. It is not yet clear what format this resource will take but one potential option would be incorporating information in a registry such as COPTR³.

Next on the list of future developments are resources to support recruitment for digital preservation posts. Anecdotal evidence suggests that many find the current digital preservation labor market difficult to navigate and role descriptions included in advertisements to be intimidating, whilst employers often struggle to assemble a viable pool of candidates. This will begin a new iteration of our labor market analysis work, and will lead to guidance covering issues such as role titles, job advertisements, salaries, interview methodologies, and more. The hope is this will ease the process for those recruiting new employees but will also remove some of the uncertainty for those applying for a new position.

Finally, we hope to engage with those offering training and education opportunities to encourage the use of the competency framework to aid in the development and review of their courses and curricula. We will also be using the competency framework inhouse for this purpose, reviewing current DPC training modules and resources and planning for the development of new content in line with gaps identified against the framework. This will be of particular use as we build on our existing online training offering, the Novice to Know-How learning pathway⁴.

VII. CONCLUSION

The purpose of this project by the Workforce Development team at the Digital Preservation Coalition was to develop a new digital preservation competency framework as described in Sharon

McMeekin's 2019 iPres paper on sustainability and digital preservation workforce development. While there is a great deal of previous work on digital preservation competencies and curricula, a new, more defined competency framework for digital preservation practitioners was necessary to ensure successful pathways for those in the field. With this in mind, the project aimed to develop a competency framework that balances detail with flexibility--providing enough detail to be applicable by digital preservation practitioners across different organizational contexts while also having enough flexibility to be used for a number of distinct purposes such as recruitment, training, or benchmarking models.

This was not an easy feat, given the scale of the task and high/lofty aims, but we employed a team-based approach with qualitative research methods to build on the paths laid by previous efforts and drew from the expertise of DPC colleagues to develop and design our framework. Throughout the research and design process, we found that the use of concept mapping exercises was useful for designing and refining the framework in a meaningful way and that feedback from colleagues who bring different perspectives is invaluable. To continue this collaborative approach to development, we present the competency framework here in this paper to facilitate feedback from the iPres and broader digital preservation community as we continue to expand the collection of additional resources that complement the framework. To borrow a lyric, we hope that there “ain't no mountain high enough” to keep us from this momentous task...

ACKNOWLEDGMENTS

We would like to acknowledge the support of our DPC colleagues for their input to and feedback on early drafts of the skills framework, particularly Jenny Mitcham, Paul Wheatley, Michael Popham, and William Kilbride. Also, the support, insight, and encouragement of the DPC's Workforce Development Sub-Committee, chaired by Susan Corrigan of the National Records of Scotland. And finally, the DPC members who participated in the pilot of DPC CAT and provided us with constructive

³https://coptr.digipres.org/index.php/Main_Page

⁴<https://www.dpconline.org/digipres/train-your-staff/n2kh-online-training>

feedback that has been incorporated into the final versions of the resources.

REFERENCES

- [1] McMeekin, Sharon. "People Get Ready: Building Sustainability into Workforce Development." Proceedings of the 16th International Conference on Digital Preservation, Amsterdam, September 2019. <https://osf.io/dtqe8/>
- [2] Kenney, Anne, and McGovern, Nancy, "The Five Organizational Stage of Digital Preservation" in Hodges, Patricia; Bonn, Maria; Sandler, Mark; and Price Wilkin, John (2003) Digital Libraries: A Vision for the 21st Century: A Festschrift in Honor of Wendy Lougee on the Occasion of her Departure from the University of Michigan <http://dx.doi.org/10.3998/spobooks.bbv9812.0001.001>
- [3] Fulton, Bruce, Peter Botticelli, and Jana Bradley. "DigIn: A Hands-On Approach to a Digital Curation Curriculum for Professional Development." Journal of Education for Library and Information Science 52, no. 2 (2011): 95–109
- [4] Lee, Cal, "Matrix of Digital Curation Knowledge and Competencies" (2009) <http://web.archive.org/web/20100616210630/http://ils.unc.edu/digccurr/digccurr-matrix.html>
- [5] Molloy, Laura; Gow, Ann; Konstantelos, Leo. "The DigCurV Curriculum Framework for Digital Curation in the Cultural Heritage Sector", International Journal of Digital Curation, 2014, Vol. 9, Iss. 1, 231-241 <https://doi.org/10.2218/ijdc.v9i1.314>
- [6] Blumenthal, Karl-Rainer, et al. "What Makes a Digital Steward: A Competency Profile Based on the National Digital Stewardship Residencies." Open Science Framework, March 4, 2016. <https://osf.io/zndwq/>
- [7] Atkins, Winston, et al. "Staffing for Effective Digital Preservation, 2017: An NDSA Report." National Digital Stewardship Alliance, September 13, 2017. <https://osf.io/zndwq/>
- [8] Blumenthal, Karl; Griesinger, Peggy; Kim, Julia Y.; Peltzman, Shira; and Steeves, Vicky (2020) "What's Wrong with Digital Stewardship: Evaluating the Organization of Digital Preservation Programs from Practitioners' Perspectives," Journal of Contemporary Archival Studies: Vol. 7, Article 13. <https://elischolar.library.yale.edu/jcas/vol7/iss1/13>
- [9] Digital Preservation Coalition, "Digital Preservation Competency Framework", 1st Edition, September 2022, <https://doi.org/10.7207/dpccf22-01>
- [10] Digital Preservation Coalition, "Rapid Assessment Model", Version 2, March 2021, <http://doi.org/10.7207/dpcram21-02>
- [11] Digital Preservation Coalition, "DPC Competency Audit Toolkit", 1st Edition, September 2022, <https://doi.org/10.7207/dpccat22-01>
- [12] Digital Preservation Coalition, "Example Role Descriptions for Digital Preservation", September 2022, <https://www.dpconline.org/digipres/train-your-staff/dp-competency/dp-roles>

SUPPORTING RESILIENCE OF INTERNET ART THROUGH THE EXECUTABLE ARCHIVE FRAMEWORK

Case-study of Virtual Reality Modeling Language & Flash Artwork

Natasa Milic-Frayling

*Intact Digital Ltd
United Kingdom
natasamf@intact.digital*

Michael Takeo Magruder

*Takeo.org
United Kingdom
m@takeo.org*

Abstract – We present a case study of an artist-led reconstruction of Internet art, triggered by the obsolescence of the Virtual Reality Modeling Language (VRML) and Adobe's Flash software. The study provides insights into the ongoing management of technology configurations throughout an artwork's life cycle in order to maintain consistency in its presentation and interaction. Guided by the artist's requirements for the integrity of the artwork, we evaluated multiple software configurations to achieve quality (Q), stability (S), longevity (L) and scaled online access (A) of the artwork installations. These installations were explored within the Executable Archive framework, centered on long-term artwork integrity and continual maintenance of software environments to ensure reliable access. The study reveals the artist's priorities that guided preservation actions and the importance of access requirements as an integral part of the artwork reconstruction. The study demonstrates that the Executable Archive framework can make Internet art more resilient.

Keywords – Internet art, VRML, Cortona3D, Flash
Conference Topics – Resilience. Innovation.

I. INTRODUCTION

In December 2020, Adobe ceased support for Adobe Flash and, from 12 January 2021, blocked Flash content from running in the Flash Player to protect users from security risks. This sent shockwaves through artists' and authors' communities with concerns about impact on collections of digital artworks that use Flash and Flash Player to reach Web audiences ([1], [2], [3]). From its inception, Flash played an essential role in the online creative landscape, providing a then missing capacity of the Web to support animation

and visual-design consistency across platforms. With the standardization and adoption of the Virtual Reality Modeling Language (VRML), authors could also specify platform-independent 3D objects with rich structures, textures, sounds and interaction. ParallelGraphics' Cortona3D viewer for VRML

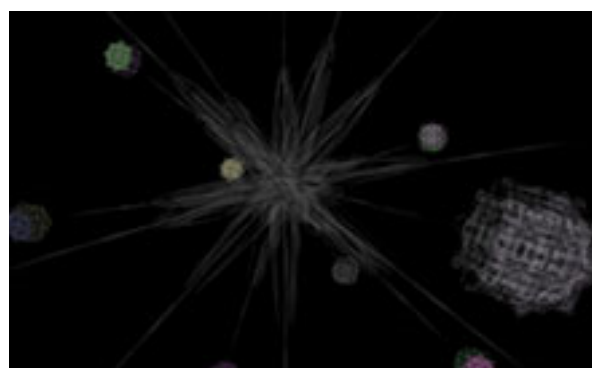


Figure 1 Star shape configuration that connects a series of infinitely complex virtual sculptures generated from the single word 'world' translated into society's most common languages.

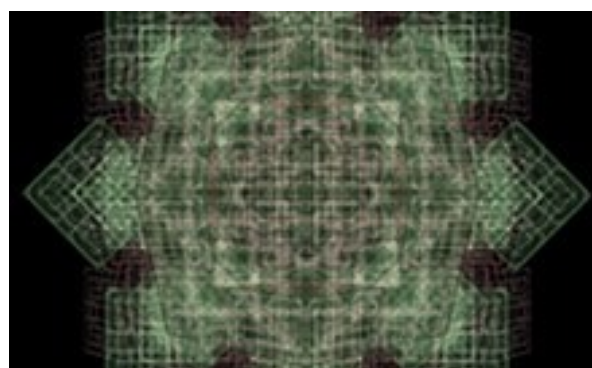


Figure 2 Intricate and infinitely complex texture of individual sculptures generated using VRML and Flash.

provided additional creative opportunities by enabling artists to combine VRML with Flash textures. All artworks using these technologies are now affected. This has prompted actions by memory institutions and specialist preservation organizations, such as Rhizome [4], who provide processes and tools to preserve complex digital artefacts. Contemporary artists who manage their artwork throughout its entire life cycle, from inception to publishing and archiving, must now find ways to keep their art alive.

We present a case study of reconstructing *World[s]*, an Internet art piece (Fig. 1 & 2) by contemporary artist Michael Takeo Magruder. *World[s]* is representative of the artist's works that blend VRML and Flash technologies. The artist and the team of experts at Intact Digital Ltd engaged in a joint effort to explore principles of reconstructing and extending use of this Internet artwork.

A. Artwork Reconstruction Approach

Considering the critical role that software components play in *World[s]*, we conducted the study within the Executable Archive framework [5] that complements traditional archives with hosting and long-term care of software environments needed to use archived digital media.

Prior work on digital art reconstruction raised practical questions around authenticity ([6], [7], [8]) and preservation decisions ([9], [10], [11], [12], [13], [14]). As a result, we paid particular attention to the artist's preferences and priorities when similar issues emerged. As it transpired, the artist's concerns revolved around four art reconstruction objectives: quality (Q), stability (S), longevity (L) and scaled online access (A). Q.S.L.A. objectives are an integral part of the artist's creative practices and publishing strategy. Thus, throughout the reconstruction process, we revisited the artist's past practices and considered how to extend them to increase the resilience of the artist's Internet artwork in the face of the changing technological ecosystem and risks to long-term artwork integrity. These four objectives guided our reconstruction work.

Furthermore, the previous use of Executable Archive framework [5] demonstrated quality assurance processes for specialized software installations that are needed in highly regulated domains where archived data must remain usable for decades. Such software is subject to rigorous testing by software vendors and well-established

Computer System Validation practices by IT support staff. We expanded the Executable Archive framework with processes to deal with complexities of artwork integrity that include dependencies among multiple software with different life-cycle and artist's idiosyncratic use of technologies to create unique artwork experiences.

B. Study Contributions

The artist led reconstruction of *World[s]* provided an opportunity to apply the Executable Archive framework to a bounded, performant, browser-based artwork dependent on multiple software technologies. As a case study, it complements prior conservation efforts of similar art pieces by considering the end-to-end process, from archiving to reconstruction and active use. It demonstrates how this approach leads to specific practices that make Internet art resilient to obsolescence risks:

- 1) *Artwork maintenance.* The artist typically modifies the artwork technical configurations during its prime performance period, making pragmatic choices that are guided by the sense of authenticity of the artwork experience. This justifies the approach of replacing obsolete technical components during artwork reconstruction and long-term maintenance, subject to the quality assessment of the intended user experience.
- 2) *Artwork access.* Active use of artworks is an essential artistic objective, and the reconstruction process must take into account the interaction between legacy software installations and contemporary technologies for hosting and remote access to the artwork installations.
- 3) *Artwork integrity metrics.* Due to the complexity and intricacies of the art pieces like *World[s]*, it is challenging to arrive at an effective way of characterizing artwork integrity requirements and mapping them onto specifications for artwork installations. In our study, we adopted an iterative process that evolved the artwork specifications through testing and evaluating different installation configurations.

In the following, we provide background information about the problem at hand and reflect on related research. We describe the artwork reconstruction process and then discuss open

problems of characterizing the artwork installations and importance of ongoing IT support.

II. BACKGROUND

A. The Art Collection

The internet art collection created by the artist

Michael Takeo Magruder is an example of performant digital artworks from the early 21st century. The artwork *World[s]* (2006(v1.0), 2009(v1.1)) is a representative piece that combines VRML and Flash plug-ins to enable textured 3D rendering of audio-visual art elements.

The artist maintains a Web portal with detailed descriptions of his art pieces, including documentation, videos, and still images (see Appendix A). The artist also manages a repository of digital media files and selected versions of software used to create and publish the art pieces. The Web portal serves as an archive of the artist's work. In the past, the artworks have been displayed both in situ, in galleries and museums, and online. The artist maintains old PCs with these original installations. However, due to the recent obsolescence of Flash, online installations are not possible anymore.

The objective is to revitalize the Internet art collection for online use. This requires a careful technical set up since the original operating system, browser versions, and VRML and Flash plugins are neither supported nor secure. We used a local, isolated instance of the *World[s]* installation on a physical PC in the artist's studio as a *reference installation* and a benchmark for specific aspects of quality and stability. However, even this reference installation needed to be extended and modified to achieve the Q.S.L.A. objectives.

B. Software and Artwork Integrity

In *World[s]*, the artist uses multiple software technologies: Cortona3D VRML viewer and Flash Player plug-ins for the Internet browser with DirectX rendering and GPU acceleration to achieve the artwork aesthetics and the requisite user interaction. *World[s]* 3D interactive sculptures are presented through intricate visual and audio effects (Fig. 1 & 2). The reconstruction and long-term use of *World[s]* thus requires a principled approach to managing artwork installation complexities and adoption of processes and procedures to ensure artwork integrity over time, as supported by the Executable Archive Framework (Fig. 3).

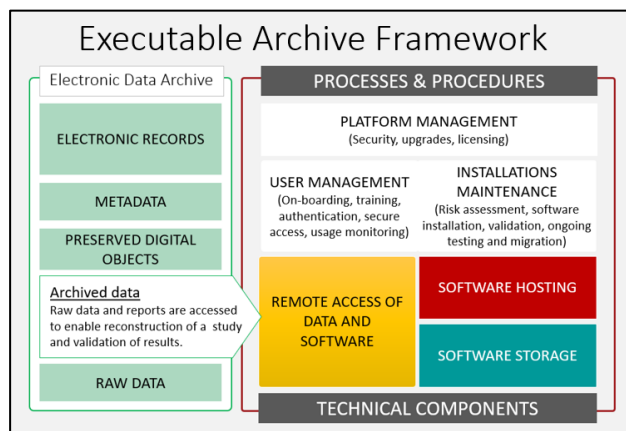


Figure 3 The Executable Archive framework complements electronic data archives with technical components and processes to manage legacy software installations and environments.

Artwork by Michael Takeo Magruder

World[s] 2006(v1.0), 2009(v1.1)

<http://www.takeo.org/nspace/ns018/>

Artwork elements

- 106 files @ 13.9MB, in a single directory
- 28 VRML (wrl): compressed and signed (no Cortona3D logo)
- 26 Flash (swf):
- 52 Audio (wav): lossless, 16bit, stereo, 8kHz
- start file = ns018.wrl

1. Executable Archive

The Executable Archive Framework highlights data and software integrity as key requirements for long-term digital preservation [5]. Archived data integrity is commonly achieved through secure storage, reliable access control and regular file fixity checks. In contrast, software integrity for archival use is given less attention, particularly by archives that restrict their practices to a pre-defined set of formats (e.g., PDF) with broadly used readers (e.g., Adobe PDF Reader). However, in highly regulated sectors, like life-sciences, data must remain immutable and Software integrity is required by Good Laboratory Practices (GLP) [15] regulations to ensure that scientists can reconstruct decades old studies from archived data.

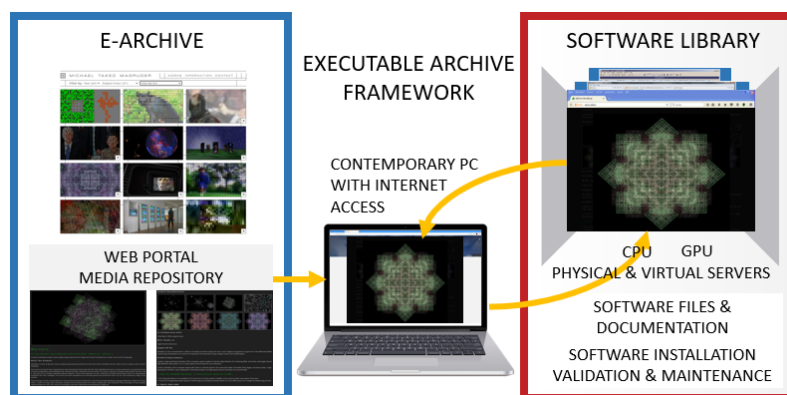


Figure 4 The Executable Archive framework extends traditional e-Archive and digital media repositories with a Software Library to manage software installations on physical and virtual servers. This enables safe use of art installations that require non-secure software components.

Within the Executable Archive Framework, software use is enabled through a Software Library platform that hosts validated legacy software installations, provides secure remote access and includes IT support services for long-term software maintenance [5].

1. Software Validation

Complex scientific protocols typically involve use of sophisticated instruments to collect data and specialized software to interpret and analyze the instrument data. The instruments and the software are subject to rigorous Computer System Validation procedures. Software integrity is strictly monitored; no changes during the software operation are allowed without a well-documented change process.

These Computer System Validation instructions are helpful when creating software installations for long-term archival use. IT specialists create virtualized and secure software installations within a Software Library and maintain them for reliable processing of study data that is kept in an electronic archive. Scientists then use software installations through remote access technologies, such as Virtual Desktops (Citrix) and Remote Desktops (Microsoft) to reconstruct studies.

These end-to-end considerations, from archiving to active use, are important to determine effective reconstruction approaches and set-up appropriate validation processes. Each technical component,

from hosting machines to remote access technologies, can affect the software performance and user experience. Thus, the validation process needs to be systematic and comprehensive, covering all the components that support use of data and software installations.

2. Artwork Integrity

The Computer System Validation practices used in scientific research typically involve software applications that are widely deployed and tested by vendors following standardized procedures. In contrast, validation of Internet art installations is complex and non-standard. First, artists use multiple software technologies that are managed separately, with different levels of support and different release and obsolescence schedules. Second, artists are likely to combine and apply software in non-standard ways, to explore new creative opportunities. Thus, the previous applications of Executable Archive Framework to individual software integrity and Computer System Validation are helpful but not sufficient.

From discussions with the artist, it was clear that the quality of interactive experiences is an important aspect of artwork integrity. The quality refers to the visual, audio and interactive experience that results from the computing environment (hardware, operating system, network connectivity, etc.), the software components, and the digital media. When preparing *World[s]* installations for exhibition, the artist would tailor the computing environment to produce the intended artistic effects. The installations would change across art exhibitions as

the artist made adjustments to achieve consistency in the quality of the artwork expression.

With this in mind, reconstructing an artwork affected by software obsolescence can be viewed as a task to identify a configuration that retains the intended quality and increases the artwork's resilience to emerging technical issues. Considering the complexity of multiple software interactions in *World[s]*, we extended the Executable Archive procedures from pre-specified validation of individual software to combinatorial validation of multi-software installations by applying *emerging qualification criteria*.

Indeed, in the case of *World[s]*, the reference installation that the artist recreated on a standalone PC was helpful to convey the *correct* artistic expression and to reason about *acceptable departures* from that reference expression as we considered factors related to Q.S.L.A. objectives and compared new installations against the initial qualification criteria. Explorations of Q.S.L.A. objectives were conducted within a Software Library (Fig. 4) that effectively extended the artist's archive into an Executable Archive. The Software Library provided a secure platform for hosting software environments with the artwork installations and remote use of the artwork through virtual desktops that are accessed via standard Internet browsers.

III. RELATED WORK

A. Preservation of Digital Art

Concerted efforts, platforms, and tools have brought significant advances in preserving digital art ([4], [9], [12], [16], [17], [18], [19]). Researchers have identified key issues with unbounded and networked Internet artwork and explored approaches to engage meaningfully with their scale, complexity and dynamic nature. We have seen successful efforts to preserve self-contained (i.e., bounded) digital art using migration, virtualization, emulation and porting. We have also seen success with reverse engineering digital art installations.

A relatively recent obsolescence of Flash (31 Dec 2020) motivated a significant effort in the preservation of electronic literature and net artworks. Our work contributes to efforts to address the obsolescence of Flash ([10], [20]). We focus on the operational aspects of art presentation and delivery (i.e., on its active use), while managing the security risks of out-of-support software and

complex interactions between contemporary and legacy technologies.

The notion of authenticity in relation to digital artefacts and experience has been essential for assessing the quality of the digital preservation activities ([6], [7], [8]). Through the consideration of artwork integrity, we show how the artist sets the boundary between experiential and technological aspects of the digital artwork and decides which aspects are essential to maintain.

B. Management of Software Obsolescence

Internet art and computational art forms in general rely on technologies that are produced and used within the global software ecosystem. Thus, it is instructive to consider how software obsolescence is managed in a broader context.

In engineering and electronic systems management, software obsolescence is considered alongside a more general concern of Diminishing Manufacturing Sources and Material Shortages that affects maintenance and leads to the decommissioning of systems ([20], [21], [22], [23]). Software, including open source and Commercial-Off-The-Shelf (COTS), becomes unusable due to functional, technological and logistical obsolescence ([24], [25]). With software, we are particularly aware of issues with:

- Software vendor no longer producing a software product (end-of-sale)
- Inability to extend or renew licensing agreements (legally unprocurable)
- Software vendors, distributors and other third parties ceasing to provide support (end-of-support).

COTS software, in particular, has end-of-sale and end-of-support dates that may be separated by long periods of time. That is taken into account in the system 'sustainment' practices that involve maintenance, support, and upgrade to improve the system capacity to endure. By maintaining and upgrading the system, its availability is maximized while controlling the cost and footprint [21].

The end-of-sale and end-of-support are key events that affect artwork integrity and trigger the reconstruction activities. With a lack of planning for sustainment, most of digital art suffers [2]. Conducting the artwork reconstruction and enabling long-term use, requires constant awareness of the changes in the technology landscape and planning

for the component replacement as licensing, operational and security issues arise.

At the same time, the obsolescence cannot be stopped or reversed unless all the stakeholders are engaged ([25], [27], [28], [29]). Thus, the focus is on measures for mitigating the impact, depending on the type of obsolescence. In the case of software obsolescence, one may consider re-developing or modifying software to work in a new development environment or hosting it within a virtual environment.

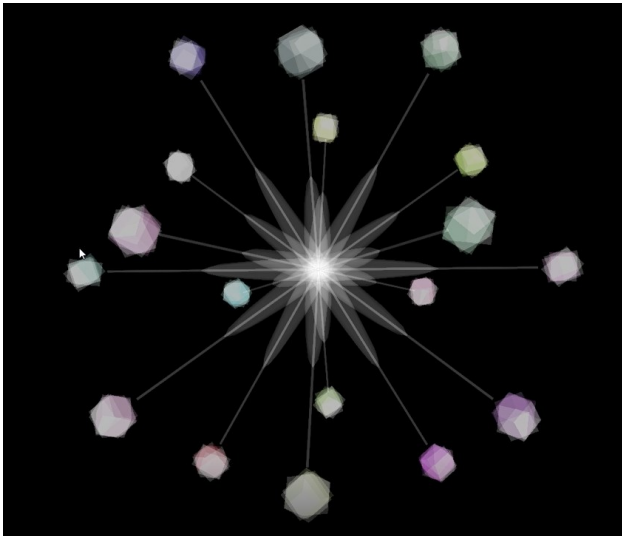
Our work complements the past preservation efforts by considering artwork resilience and long-term sustainment in the context of the creative process. This includes economic aspects of enabling Internet art access at scale which were considered at the time of artwork creation and publishing and remain essential for the preservation planning and long-term availability.

IV. CASE STUDY: WORLD[S]

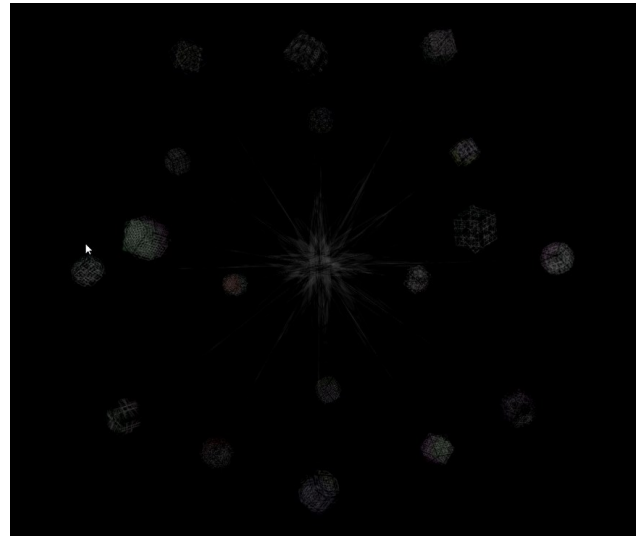
A. *Method*

For the case study we adopted a hybrid method, combining ethnographic and co-creation activities that involved the artist and the Intact Digital team comprising an IT specialist and a computer scientist. Initial scoping of work was established through exchange of information about the system requirements for *World[s]* installation and recorded online sessions where the artist provided a historical account of the *World[s]* creation and publishing, rationale behind the selection of technologies and a demonstration of an on-premises installation. The artist created recordings of the local installation and transferred software and digital media to Intact Digital for the purpose of the artwork reconstruction. The reconstruction activities were divided among the team members to cover installation, qualification and documentation of the created artwork instances.

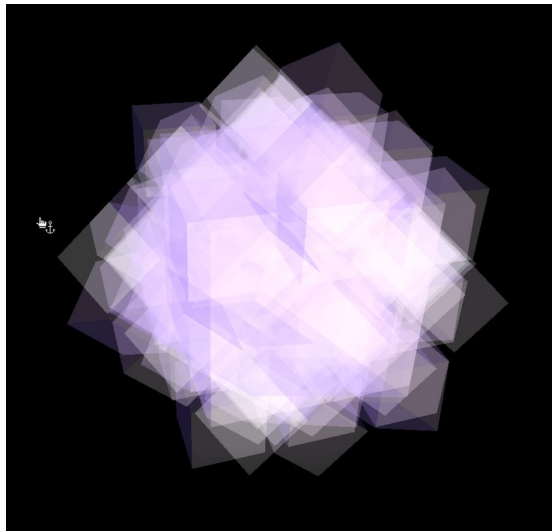
Since each installation is carefully managed in dedicated physical and virtual environments, we optimized resources by selecting software versions based on the artist's experience with technical issues in the past and gradually substituted obsolete components to meet Q.S.L.A. objectives. During this process, the artist's instructions and requirements became more specific and the artist's relative priorities of Q.S.L.A. objectives became more crystalized. For example, one of the installations deemed acceptable involves a trade-off between the



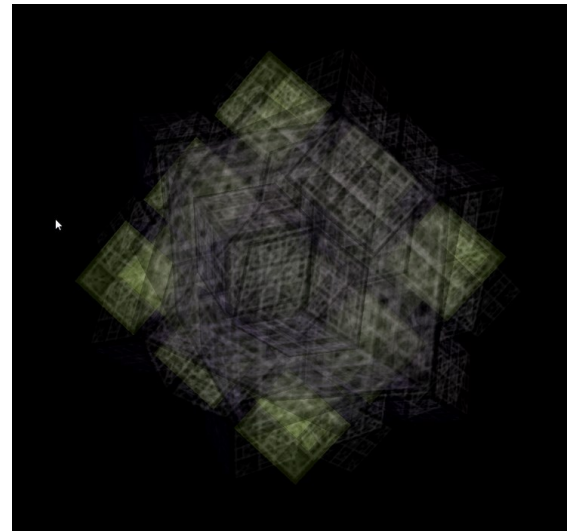
(a) The main interface star structure with connected digital sculptures.



(b) The start structure correctly rendered with Flash textures.



(c) Cortona3D rendering of one of the individual sculptures.



(d) The same digital sculpture correctly rendered with Flash textures.

Figure 5 A combination of Cortona3D v.5.0 and earlier versions of Flash causes delays in the rendering of textured sculptures. Resizing Cortona3D window causes textures to disappear altogether.

economically and technologically sustainable global access and the rendering quality.

The work was conducted in two phases:

Phase I. Technical feasibility assessment, involving (a) identification and sharing of artwork digital assets, (b) specifying initial artwork installation and performance requirements and (c) assessing the feasibility of artwork installation.

Phase II. Performance qualification, involving iterations of physical and virtual installations: (a) creation of local installations by the artist, in the artist's studio, on a dedicated physical computing device and (b) re-creation of the artist's artwork configurations by the IT specialist within a local IT environment and in the data center, within the

Software Library. Work on the data center installations was facilitated through using virtual desktop facilities and included use of virtualized and physical computing resources.

B. *Phase I: Technical feasibility*

Preparations for the work began with the considerations of the artist's immediate concerns about visual and interactive aspects of the *World[s]* artwork that depend on DirectX, Cortona3D VRML and Flash plugins with supporting GPU capabilities. In *World[s]*, the Cortona3D plug-in is used to render 3D geometry that the artist specified using VRML, overlaying them with textures generated through Flash. When running an artwork instance, a browser uses Cortona3D for VRML rendering and Cortona3D

receives textures from Flash. Generally, past VRML renderers did not support import of Flash textures. Cortona3D was unique in providing that extended feature and therefore, it is considered an essential component of the artwork reconstruction that cannot be substituted.

For scaled Internet access, one important factor was GPU cost. In the past, the artist would purchase hardware for individual PCs used for art display. Similarly, during the study, no virtual GPU resources were used within the Software Library platform. All the GPU units were physical, on individual PCs, or attached to a physical server that was accessed remotely. Furthermore, since server configurations can support multiple user sessions, they provide more economical scaling of artwork access. That immediately implies that an optimal artwork installation would need to be compatible with the server operating system (OS), e.g., MS Server 2019, as opposed to the Windows 7 OS used in the artist's reference installation.

We started with the technical specifications of the reference installation created by the artist that included Microsoft Windows 7 64-bit, Adobe Flash Player 11.2.202.228, ActiveX 64-bit, Cortona3D Viewer 7.0 r185 64-bit and Internet Explorer 11. The first objective was to investigate interactions among four technical components and their behavior with and without a GPU. Thus, we experimented with different installation environments:

- Physical PC with Windows 7 and a GPU
- Physical PC with Windows 10 and a GPU
- Virtual PC with Windows 7 and no GPU.

The initial findings suggested that interactions between Cortona3D, DirectX and Flash needed to be given a careful consideration:

- The Flash based textures sometimes failed to load and disappeared (Fig. 5)
- Opening VRML files themselves led to inconsistent results.

Furthermore, experiments with DirectX and OpenGL showed a significant degradation of OpenGL rendering, thus ruling out a possibility of replacing DirectX.

C. Phase II: Performance qualification

The second phase focussed on identifying a configuration that satisfied Q.S.L.A. objectives, including the *World[s]* user experience through a

INITIAL TECHNICAL SPECIFICATION

World[s] - VRML + Flash Internet Artworks

Test System (Software)

- Windows 7 64-bit pro Service Pack 1 (last updated 20/02/2015)
- Internet Explorer 11: v 11.0.9600.17728
- Cortona3D Viewer 7.0.185 64-bit

Test System (Hardware)

- CPU: Intel Core2 6600 (dual core @ 2.4GHz)
- RAM: 3GB DDR2-800
- GPU: AMD Radeon HD7750 800MHz - 1GB DDR5
- Storage: 60GB SATA SSD
- Display: 1920x1080p, 60Hz, 32-bit colour
- Audio: onboard stereo

IE 11 & Flash Setup

- both set to no updates

Cortona3D Preferences

- General: Background Color = black
- General: Gradient Color = None
- General: CPU load = Highest Frame Rate
- General: Display Frame Rate = None
- General: Console Mode = Auto Launch
- Scene: NA
- Renderer: Direct X9. Select:
 - o Anti-aliasing real-time (if possible),
 - o Optimize textures for quality
 - o Use Textures for Mip-mapping
 - o Limit Texture Size Disabled
- Navigation: Default
- Skin: Default

browser and at scale. For all practical purposes, we treated O.S.L.A. as the artwork integrity requirements that the installations within the Software Library needed to meet. While fully functional, the artist's local installations of *World[s]* could not be exposed to the Internet due to the Flash obsolescence and therefore could not achieve longevity and access objectives.

We divided artwork implementation activities into two streams with related goals: (1) creation of local installations on a physical PC to guide the quality assessment and (2) creation of installations in the Software Library environment to satisfy Q.S.L.A. criteria. During the implementation work, the team

Table 1 Versions of software used in the explorations of the *World[s]* installations. Considering that 4 components are involved in each configuration, one is dealing with a large optimization space.

Flash	VRML plug-in	Renderer	Browser
Macromedia Flash 8.0.42.0 (2006)	Cortona3D 5.0 r150 (2006)	DirectX 5 (1997)	Firefox 1.5.0.1 (2007)
Adobe Flash 11.2.202.228 (2012)	Cortona3D 5.1 r157 (2007)	DirectX 7 (1999)	Firefox 52.9.0 (2018)
Adobe Flash 32.0.0.101 (2018)	Cortona3D 5.1 r161 (2007)	DirectX 9 (2002)	Internet Explorer 8 (2009)
	Cortona3D 6.0 r179 (2009)		Internet Explorer 11 (2013)
	Cortona3D 7.0 r185 (2011)		

consulted the artwork documentation, audio-visual recordings and still images and, in addition to email communication, conducted four 2-hour discussions about specific issues encountered with various configurations of the technical components. The initial focus was on the quality and stability. This required navigating through the space of options by combining multiple versions of Flash, Cortona3D, DirectX and Browsers (180 possible configurations). Table 1 shows a selection of technology releases that were selected for explorations.

1. Components Interaction

Influence of Cortona3D releases. The artist explored a number of system configurations on the local PC, evaluating their quality and stability. The selection of Cortona3D 5.0 and DirectX versions were made to mimic the historical installations. The core components comprising Cortona3D, DirectX and Flash, including GPU, were considered as essential and non-interchangeable with alternatives. The operating system and the Internet browser were treated flexibly.

The artist's familiarity with the Cortona3D releases, led to an informed decision to stay with the Cortona3D 5.0 version and, in case of required upgrade, skip Cortona3D 6.0 which was known for a number of technical problems. That decision naturally led to the selection of the browser. Historically, Cortona3D 4.0 and earlier versions had plug-ins for Internet Explorer (IE) separate from other browsers. Starting with Cortona3D 5.0 version, the same plug-in was used across browsers. In fact, even for Cortona3D 5.0, the installation interface had a checkbox to indicate whether the plug-in is used with IE or other browsers, suggesting that the plug-in might work differently for other browsers.

Influence of Flash versions. As early observations suggested, the overlaying of VRML geometry

elements with Flash textures suffered from inconsistent behaviour. The rendering qualities were resolved by the use of Cortona3D, 5.0 r150 with the later version of Flash 11.2. The artist preferred that configuration from the quality and stability perspective.

2. Access Requirements

In order to allow secure use of non-supported browsers with Flash and Cortona3D plug-ins, the IT expert set up physical computing devices with remote access through a virtual desktop that enabled use of the artwork while blocking the Internet inbound traffic. With this security protection, the user could conveniently interact with the artwork through a virtual desktop within a modern browser while the platform and the installations were safe from cyberattacks (Fig. 4).

We tested the performance of IE 11 with Cortona3D 5.0 r150, and Flash 11.2 running on MS Windows Server 2019, equipped with a GPU. The interaction was enabled through a Citrix virtual desktop. We discovered that a combination of a multi-session environment using MS Windows Server with remote access, Cortona3D 5.0 r150, and GPU was incompatible. We identified Cortona3D 5.0 r150 as a likely cause. More precisely, the way Cortona3D 5.0 r150 plug-in establishes a context for the use of the GPU (e.g., through a remote desktop) was not compatible with a multi-session use of Windows. We tested a modified configuration Cortona3D 7.0 r185, a newer version of the plug-in, and confirmed that it allowed the GPU to be used within the multi-session Windows environments.

Thus, the final configuration for meeting Q.S.L.A. objectives comprised Cortona3D plug-in v7.0 r185, Internet Explorer 11.1790.17763.0 and Macromedia Flash ActiveX plug-in 8.0 r42. Since the original *World[s]* artwork was created and aesthetically

optimised for Cortona3D 5.0, we needed to conduct detailed quality and stability assessments to determine the implications of using Cortona3D 7.0. It turned out that Cortona3D 7.0 introduces flicker during artwork rendering. However, from the artist's perspective, the trade-off between the changed visual effects and cost-effective online access was deemed acceptable.

This completed the reconstruction of the *World[s]* artwork that satisfied the Q.S.L.A. objectives:

- Quality and stability are ensured through the compatible versions of Flash, Cortona3D, DirectX and Internet Browser.
- Longevity is extended through the use of MS Server 2019 OS which is fully supported and secure.
- Access is enabled through secure and fully supported Citrix virtual desktops, from any contemporary browser; thus, staying true to the nature of Internet art. MS Server environment supports scaling through multi-user access.

V. DISCUSSION

The *World[s]* reconstruction case study provided a number of important insights for the digital art management practices.

Since fundamentally dependent on digital technologies, Internet artwork can be sustained only through carefully managed computing environments. Our collaborative effort, involving the artist and IT specialists, resulted in a clear understanding of the core and supporting technologies. By exploring the dependencies among them, we identified technological components that must be retained and those that could be replaced. The choice of the operating system and the browser, for example, were never seen as an integral part of the *World[s]* artwork and therefore could be chosen more flexibly to achieve longevity in terms of vendor support, licensing and security updates. Having multiple options for implementing the artwork installations increases its resilience to the technology obsolescence.

At the same time, dealing with a combinatorial set of possible configurations for the artwork installation (Table 1) required reliance on the artist's experience and intuition about ways a specific technology would process art media, i.e., audio-

visual material, and programming scripts. Based on the *World[s]* reconstruction effort we can affirm:

1) *Importance of considering end-to-end use scenario*

It is critical to take a holistic view and ensure that the full set of requirements for artwork use are included in the artwork reconstruction process. As we have seen, publishing *World[s]* as an online Internet art installation, through remote access to a hosting server with GPU and browser plug-ins, requires all the components to work in concert. We had to make a change to the original specifications and include a more recent version of the Cortona3D plug-in to enable GPU usage in remote access sessions.

2) *Importance of ongoing maintenance and support*

As physical and virtual computing platforms change and remote access technologies evolve, Internet artwork installations will need to be revisited and adjusted. Ideally, the art creation process would involve *sustainment plans* that include maintenance, support, and upgrade. This is already common in architecture, engineering and manufacturing, and can be introduced as part of the art appraisal and value retention efforts.

3) *Importance of comparison and benchmarking*

The *World[s]* reconstruction process confirmed practical challenges in capturing artwork characteristics and providing an operational guide for re-installations. In contrast to scientific scenarios where the integrity of software and data are coupled to provide standard presentations of results, digital art involves intricate audio-visual effects and interactive properties that are difficult to define and capture. During the *World[s]* reconstruction we have observed a few important aspects:

- (a) *Attention trigger*—Considerations of a specific artwork property, e.g., lighting or colour intensity, increased when an artwork configuration showed unexpected outcomes, e.g., a significantly better or a significantly worse quality compared to previous instances or set expectations.
- (b) *Evolving quality criteria*—The artist's view is considered as ultimate criteria for the quality of the artwork installation. However, that view may evolve with the artist's exposure to different options and opportunities for changes that are perceived as improvements.

(c) *Benefits of version management*—The iterative process of optimizing the artwork configuration and assessing the quality of installations was facilitated by easy access to previous attempts. Within the artist's studio, the artist systematically explored and stored different versions of the artwork installations and discussed them with the IT specialist. Similarly, the IT specialist used the Software Library platform to set up convenient remote access to the server hosted installations. This suggests that the version management of the artwork configurations should be an essential part of the long-term care and quality assurance. Executable Archive framework supports that practice and incorporates Software Library as a platform to facilitate access to prior installations.

VI. CONCLUDING REMARKS

Our study demonstrated the use of the Executable Archive framework to reconstruct and increase the resilience of Internet art affected by technological obsolescence. It illustrated a principled way of identifying core and supporting components to achieve Q.S.L.A. objectives. By treating the Internet art reconstruction as an extension of the artist's ongoing care of art installations, the artist and the technical team ensured the use of legacy software through contemporary technologies for remote access that are supported and secure. Compared to related efforts that use cloud resources and virtualization ([4], [19], [30]), such technologies are used as a means of scaling, stability and audience reach rather than preservation.

Generally, the Executable Archive framework extends the standard archiving and digital preservation practices with IT processes and procedures that support installation, validation, monitoring and long-term maintenance of software essential for the artwork use. The *World[s]* case study demonstrated that such an approach can re-vitalize and protect the artistic and cultural value of the Internet art from the Flash obsolescence. Our future work will build on the insights from the case study and explore opportunities to generalize the approach to a broader range of digital artworks.

ACKNOWLEDGMENT

We acknowledge the contributions of M. Willcock who implemented artwork installations and a value

of the Internet Archive in identifying and collecting past versions of software used in digital art, including Cortona3D, DirectX and Flash installations. We thank Adam Farquhar for insightful feedback on our study and advice on improving the manuscript.

REFERENCES

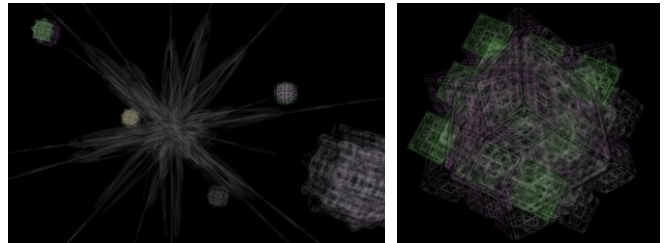
- [1] Grigar, Dene M., and Stuart A. Moulthrop. "Pathfinders: Documenting the Experience of Early Digital Literature." (2015).
- [2] Mladentseva, Anna. "Responding to obsolescence in Flash-based net art: a case study on migrating Sinae Kim's Genesis." *Journal of the Institute of Conservation* 45, no. 1 (2022): 52-68.
- [3] Salter, Anastasia, and John Murray. "E-Lit after Flash: The Rise (and Fall) of a "Universal" Language." *Electronic Literature as Digital Humanities: Contexts, Forms, and Practices* (2021): 267.
- [4] Dragan Espenschied, 'Emulation or it Didn't Happen', Rhizome, <https://rhizome.org/editorial/2020/dec/21/flash-preservation/> (accessed 28 February 2021).
- [5] Milic-Frayling, Natasa, and Marija Cubric. "EXECUTABLE ARCHIVES: Software integrity for data readability and validation of archived studies." (2021).
- [6] Innocenti, Perla. "Rethinking authenticity in digital art preservation." (2012): 63-67.
- [7] Innocenti, Perla. "Bridging the gap in digital art preservation: interdisciplinary reflections on authenticity, longevity and potential collaborations." (2012): 71-83.
- [8] Innocenti, Perla. "Bridging the gap in digital art preservation: interdisciplinary reflections on authenticity, longevity and potential collaborations." (2012): 71-83.
- [9] Ensom, Tom. "Revealing hidden processes: instrumentation and reverse engineering in the conservation of software-based art." In *AIC 46th annual meeting*, Houston, Texas, USA. 2018.
- [10] Fiadotau, Mikhail. "Growing old on Newgrounds: The hopes and quandaries of Flash game preservation." *First Monday* (2020).
- [11] Hedstrom, Margaret and Christopher A. Lee, 'Significant Properties of Digital Objects: Definitions, Applications, Implications', *Proceedings of the DLM-Forum* (2002): 221.
- [12] McGarrigle, Conor. "Preserving Born Digital Art: Lessons From Artists' Practice." *New review of information networking* 20, no. 1-2 (2015): 170-178.
- [13] Phillips, Joanna. "Reporting iterations: a documentation model for time-based media art." *Revista de Historia de Arte-Performing Documentation in the Conservation of Contemporary Art* 4 (2015): 168-179.
- [14] Quaranta, Domenico. "From Context to Content: On the Preservation of Net-based Art." In *Science and Art*, pp. 452-476. 2020.
- [15] OECD Guidance on Principles of GLP Data Integrity [https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/cbc/mono\(2021\)26&doclanguage=en](https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/cbc/mono(2021)26&doclanguage=en)

- [16] Guez, Emmanuel, Morgane Stricot, Lionel Broye, and Stéphane Bizet. "The afterlives of network-based artworks." *Journal of the Institute of Conservation* 40, no. 2 (2017): 105-120.
- [17] Laurenson, Pip. "Authenticity, change and loss in the conservation of time-based media installations." *Tate papers* 6, no. Autumn (2006).
- [18] Lurk, Tabea, Dragan Espenschied, and Juergen Enge. "Emulation in the context of digital art and cultural heritage preservation." *PIK-Praxis Der Informationsverarbeitung Und Kommunikation* 35, no. 4 (2012): 245-254.
- [19] Rechert, Klaus, Patricia Falcao, and Tom Ensom. "Introduction to an emulation-based preservation strategy for software-based artworks." *Tate Research Publications* (2016).
- [20] Poppe, Erik, Eduard Wagner, Melanie Jaeger-Erben, Jan Druschke, and Marina Köhn. "Is it a bug or a feature? The concept of software obsolescence." (2021).
- [21] Sandborn, Peter, and William Lucyshyn. "Sustainment Strategies for System Performance Enhancement." In *Handbook of Adv. Performability Engineering*, pp. 271-297. Springer, 2021.
- [22] Starling, James K., Youngjun Choe, and Christina Mastrangelo. "Identifying DMSMS availability risk at the system level." *International Journal of Production Research* 59, no. 10 (2021): 2905-2925.
- [23] Tomczykowski, Walter. "DMSMS Acquisition guidelines: Implementing parts obsolescence." 2001. US Department of transportation, Federal Aviation Administration. *Obsolescence and Life Cycle Management for Avionics*. 2015.
- [24] Feldman, Kiri, and Peter Sandborn. "Integrating technology obsolescence considerations into product design planning." *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2007.
- [25] Jensen, Peter Byrial, Linda Nhu Laursen, and Louise Møller Haase. "Barriers to product longevity: A review of business, product development and user perspectives." *Journal of Cleaner Production* 313 (2021): 127951.
- [26] Boissie, Kevin, Thomas Vigier, Marc Zolghadri, and Sid-Ali Addouche. "Business Intelligence and Obsolescence Engineering: Prediction, Performance and Innovation, Linked Destinies." In *Int. Design Engineering Technical Conferences and Computers and Information in Engineering Conf.*, vol. 85413, p. V005T05A019. American Society of Mechanical Engineers, 2021.
- [27] Bartels, Bjoern., Ermel, Ulrich., Sandborn Peter. Et Pecht, G., *Strategies to the prediction, mitigation and management of product obsolescence*, John Wiley & Sons, 2012.
- [28] Box, Jo., *Extending Product Lifetime: Prospects and Opportunities*, *European Journal of marketing*, 34-49, 1983.
- [29] Van Nes, Nicole, and Cramer Jaqueline., *Influencing Product Lifetime through Product Design*, *Business Strategy and the Environment*, 286-299, 2005.
- [30] Dirk von Suchodoletz, Klaus Rechert, and Isgandar Valizada, 'Towards Emulation-as-a-Service: Cloud Services for Versatile Digital Object Access', *The International Journal of Digital Curation* 8 (2013): 132.

A. Internet artwork collection

World[s] v1.1 by: Michael Takeo Magruder with Drew Baker, 2006(v1.0) - 2009(v1.1)
<http://www.takeo.org/nspace/ns018/> is part of the artist's collection of VRML-Flash Internet artworks produced between 2004 and 2014.

Description of the artwork from the <https://takeo.org>:



In Collaboration with:

Drew Baker [VRML programming]

With Thanks to:

Hugh Denard [discourse]

Supported by:

World[s] v1.0 was commissioned in 2006 for *Soundtoys.net* with funding from *Arts Council England* and generous support from *The Watershed Media Centre*; *King's Visualisation Lab*, *Centre for Computing in the Humanities*, *King's College London*; and *ParallelGraphics*.

Artwork Requirements:

[gallery] *High-specification Windows 7/8/10 computer system capable of real-time high definition 3D rendering (VRML and Flash); multi/single-channel high definition video system; 5.1/2.1 audio system; and HCI device for user interaction.*

[online] *Windows 7/8/10 computer system with Firefox or Internet Explorer; the Cortona3D Viewer and Adobe Flash plugins; and stereo audio. A high-specification CPU/GPU, colour display with ≥1024x768 resolution and high-speed Internet connection are recommended.*

World[s] <http://www.takeo.org/nspace/ns018/>
 v1.1 by: Michael Takeo Magruder with Drew Baker,
 2006(v1.0) - 2009(v1.1)

About the Artwork:

World[s] is a series of dynamic virtual sculptures generated exclusively from the word 'world' translated into the native script of society's most common languages.

Each word in its text format is imported into a two-dimensional 32x32 pixel Flash file. The embedded characters are then vectorized, re-proportioned into a square configuration, and multiplied at 90° intervals and

their respective mirrored states. The result is a group of mandala-esque entities less than 1KB in size that can be infinitely expanded without pixilation. These visual elements are then rasterized as 64x64 pixel bitmaps which are subsequently translated into sonic analogues. The visual and audio equivalents are inherently paired and provide the basis for the next evolutionary stage of the artwork.

These pairings are then incorporated into a three-dimensional space defined by a set of Virtual Reality Modeling Language (VRML) files. Within this virtual realm, a series of simple cubic structures oscillate at the terminal points of a central rotating star. Each structure is the summation of four possible rotational states (0°, 45°x, 45°y, 45°z) of a prototype cube that is texturized and auralized by a single pair.

When a viewer selects one of these basic elements, the entire realm is destroyed and a new complex formation is created within the void. The newly generated architecture is derived entirely from the single prototype cube that was selected by the viewer. This cube is multiplied and arranged into a perfect 3x3x3 lattice. The lattice is then quadruplicated in a manner identical to its basic precursory structure, and an exponentially more complex 'world' is formed.

Interacting with the furthest extremities or the innermost depths of the construct initiates a mechanism of self-destruction and an ensuing regeneration of the interface star. Though this process, a cyclic relationship between the work's evolutionary states is created.

"WE'RE ALL DOING THE BEST WE CAN WITH WHAT WE'VE GOT"

Preservation practices of Data Curation Network members

Hoa Luong

University of Illinois
Urbana Champaign
U.S.A.
hluong2@illinois.edu
[0000-0001-6758-5419](tel:0000-0001-6758-5419)

Mikala Narlock

Data Curation Network,
University of Minnesota
U.S.A.
mnarlock@umn.edu
[0000-0002-2730-7542](tel:0000-0002-2730-7542)

Jon Petters

Virginia Tech University
U.S.A.
jpetters@vt.edu
[0000-0002-0853-5814](tel:0000-0002-0853-5814)

Abstract – Over the course of six weeks, members of the Data Curation Network were interviewed by the Assistant Director to discuss their research data preservation practices. Through these semi-structured interviews, several commonalities emerged, including key challenges that will need to be addressed to ensure the long-term reusability of research data as well as the similar mentality many institutions expressed: that they are doing the best they can with what they have. The authors conclude by identifying areas of potential future research as well as practical collaboration opportunities.

Keywords – Research data, Peer comparisons, Data curation, Data preservation

Conference Topics – Community; Exchange

I. INTRODUCTION AND BACKGROUND

Data plays an important role in scientific research to facilitate innovations and drive the economy. By sharing research data, researchers contribute to the scientific community and the public [1]. With the move of science and technology, the sheer quantity of data has grown at an increasing rate leading to challenges around long term preservation of said data. For example, in a study done by Melero and Navarro-Molina, researchers in food science and technology expressed concerns about how long the data should be preserved and the obsolescence of the research data [2]. Researchers in data-intensive fields, such as biomedical sciences, face significant

preservation challenges that arise from the volume of the data, as well as diversity, complexity and multimodal nature of data generated by and for the researchers. [3] The long term reusability and reproducibility hangs in the balance without robust preservation interventions.

To support researchers, many institutions have developed data repositories that not only comply with funders and journals' requirements, but also take the burden off their researchers by offering reliable storage to steward and preserve the data. Academic institutions additionally offer research data management support, including data curation services that support researchers throughout the research data lifecycle by providing consultations and education during the planning, collection, and sharing phases [4]. While good data sharing requires active data management throughout a research project [5], data curation is an integral part of research data sharing– in particular for enabling data reuse– as the final review before publication. Data Curation Network (DCN) members are leveraging standard curation practices, like the CURATE[D] checklist created and leveraged by DCN members [6] to support researchers in sharing their research outputs. This robust curation– with critical tasks like documentation, format migration, and other curation activities– provides a notable difference between saving research data (i.e., bit-level) and *preserving* data, which is necessary for enabling open

science and Findable, Accessible, Interoperable, and Reusable (FAIR) data [7] while supporting long-term access.

The Data Curation Network [8] is a member-funded cooperative organization that employs a shared-staffing model to collaboratively curate data, and boasts an active community of practice that continuously develops, refines, and advances the art of data curation. The DCN has been active as a community since 2016, officially launched as a grant-funded organization in 2018, and transitioned to member-funded in 2021. As curation practitioners, these data stewards are at the forefront of open science, enabling effective data sharing, use, and reuse for communities locally and globally [9].

Currently, the DCN consists of fifteen sustaining member organizations with nearly fifty (50) curators, working together to support reusability and reproducibility in open science (note: there is also a sixteenth member of the organization that is beta-testing membership in the DCN). The DCN aims to enhance long-term access to research data through the practice of curation and regularly collaborates on different projects to address data management issues. This invaluable effort helps to ensure that data are as aligned with the FAIR and CARE [10] principles as possible. The efforts of the DCN are primarily focused on data access and use in the pursuit of open science—while invaluable, there is a lingering concern among curation professionals around the long-term preservation of these assets.

In 2017, DCN members wrote a paper comparing the data repository and curation services among the six DCN members at the time which covered some of the preservation aspects [11]. This type of project, in which institutional representatives freely share information and practices has internally been termed “peer comparison.” To understand deeper the current preservation practices and further build on the 2017 work, representatives from DCN’s members were invited to participate in this peer comparison project in which they shared and discussed their practices through a set of interview questions. This preliminary assessment revealed many key themes, as well as pain points that will need to be addressed by both the research data management and the preservation community in tandem. It is intended to help others learn what peer institutions have implemented to preserve their research data.

II. LITERATURE REVIEW

A literature review and supplementary discussion with preservation practitioners and other data stewards suggests that there is little, if any, existing literature on how organizations are addressing the unique needs of research data. There are numerous conversations happening constantly around data use, reproducible science, and open access. Countless toolkits, listservs, working groups, and professional associations discuss active research data management, curation, and access [12]. Missing from these conversations is a robust discussion of the preservation practices of these institutions. There exist guidelines [13] and tools for preserving content, especially research data, but few studies on how this is actually done— if at all— in US institutions with a demonstrated commitment to data curation. This report seeks to fill this critical research gap [14] by building on existing DCN processes: namely, peer comparisons. Information sharing sessions, in which members freely share current practices, and similar collaborative efforts are the foundation of our network, which stems back to the early investigation and report from the DCN [15]. Given the DCN’s previous work in this vein, as well as their demonstrated commitment to data curation and reproducible science, our member institutions provide an appropriate starting point for this exploration.

To better understand research data sharing and preservation, it is important to discuss the current and well-documented practices of academic libraries. Librarians at academic institutions engage researchers in the research data management lifecycle through numerous key activities, including: research data management planning; workshops and other educational offerings; individual consultations; data curation support; and institutional and/or data repositories in which researchers can deposit their data, receive curation support, and reserve a digital object identifier (DOI) to provide publishers and funding agencies [16]. In particular, educational efforts, such as training researchers and students [17], as well as outreach and engagement early in the life cycle [18] are the primary services that academic libraries are offering support to researchers in research data management [19].

The process of collecting, providing access to, and preserving these research data are similar to those that have long been utilized by libraries and

archives. In particular, data management has striking similarities to archival science and records management practices— from appraisal, arrangement, and description, through provenance and chain of custody tracking, to ensuring long-term access to content [20]. Libraries are well suited to serve as stewards of these unique and locally created digital collections, and serve as data stewards in the FAIR and open science movements.

The early educational interventions discussed above mirror pre-custodial work, a term from archival research, that describes the point at which there is room for intervening in the creation, management, and deposit of records prior to accessioning them into a repository [21, 22, 23]. In research data management, this includes education and outreach, such as establishing relationships with data creators while the research projects are still in-progress. This provides an opportunity to document their research projects prior to the publication of the data, at which point any funding has likely been spent, research partners have transitioned to new projects, and questions about project nuances (e.g., variable definitions, coding schemas, etc) will become increasingly difficult to answer, even for those most intimate with the project.

For all of these reasons, understanding the preservation practices and policies of research institutions, specifically as they apply to data and research outputs (i.e., beyond cultural heritage materials) is critical. This exploration serves as a snapshot in time, and draws attention to the extensive work currently happening and potential areas for future collaborations and research.

III. METHODOLOGY

Information about preservation practices at partner institutions was collected from representatives in a semi-structured interview. The same set of questions was sent to all institutional representatives, including a member beta-testing membership in the DCN, (see Appendix A for email and questions) with a virtual meeting request. Some members chose to provide information in an email, and some forwarded the request to others in the organization who might be more posed to answer the questions. In total, 14 representatives agreed to a brief meeting and 2 provided written summaries.

Over the course of six weeks (January-February 2022), the DCN Assistant Director Mikala Narlock

conducted 14 interviews. To encourage free information sharing, the conversations were not recorded, and Narlock took notes throughout the conversation. After the conversation, a summary of the preservation activities per institution was sent to the respective institutional representative for confirmation. At that time, representatives could correct the notes, including redacting information when necessary.

We could have applied quantitative methods to analyze the collected information. However, since this was considered as an initial exploratory assessment, and we allowed flexibility in information sharing, we have adopted a more qualitative approach. Moreover, due to the fact that the interviews were incredibly open-ended in the way that partners could take the conversation in any direction they chose, it made direct comparisons less accurate. Therefore, the results in this paper are discussed based on the interview questions and detailed answers in a table format in the appendix.

IV. RESULTS

Through these various conversations, many key themes emerged, including key pain points that will need to be addressed by the research data management and preservation communities to ensure long-term reusability of the content partner institutions are preserving. A full table of respondent information is provided in Appendix B.

A. Q1: *Preservation activities currently being taken on research data*

All partner institutions expressed the importance of reusability, access, and the needs of future users in these conversations. In other words, descriptions of preservation activities were all prefaced and described through the lens of data reusers. Preservation activities began once the datasets entered the institutional (data) repositories. Curation activities, though performed at different levels depending on the institutions, are designed to follow the CURATE(D) and FAIR model to focus and tailor to the dataset long term preservation mission. For instance, the effort of checking and adding documentation to a dataset were described as an critical step to record key contextual and reproducible information, such as how the data were collected and processed, necessary software and versioning information (especially for proprietary formats and softwares), and how to reuse and cite

the data. This action ensures and enhances the value of the data that the institutions steward.

Preservation should happen throughout the data's lifecycle. As Navale and McAuliffe [24] suggested, researchers need to become more proficient in understanding and managing research data. To do this, all of the academic institutions employ education and outreach efforts in their curation and data services. This pre-custodial preservation seeks to educate researchers on the importance of actively managing files and research outputs so as to reduce the burden of curation and preservation at the end of the project.

Backing up data is an important factor in data management as well preservation. Nearly all institutions are storing geographically distributed copies of the data (especially via AWS), with a few reporting that they have numerous copies but in a more narrow geographic distribution. Those that are not storing multiple copies of the research outputs are creating routine back-ups that, even if costly, would make data recovery possible. Checking the fixity of files (e.g., bit-level preservation) via checksums are performed by the majority of the institutions in the DCN.

Challenges in preserving research data often lies in its diversity, e.g. format. Mindful of this, all member institutions consider file transformation as an important curatorial step to create data upfront that is easier to maintain in the long-term (e.g., converting to open source or more sustainable formats, when applicable). The Michael J. Fox Foundation is slightly different due to the nature of its data and organizational mission – as a funding and research institution, the data created and managed by the Michael J. Fox Foundation is created by affiliated researchers, and no external researchers can deposit into their repositories unless funded by the foundation.

Of the DCN members that accept data deposits, some institutions offer converting files upon ingestion into the repository and preserve the original formats in their preservation back-end or make it available along with the non-proprietary format. If the repository is not automatically converting files to more stable formats, they are providing it as a recommendation sent to researchers. This is in line with what DCN member refers to as “format agnostic but not format blind.” In other words, while they might accept a wide-variety

of file formats, and may not require researchers to convert to open-source formats, they are paying close attention to the formats for future preservation needs.

B. Q2: Data retention/deaccession policies and period

As for retention, most institutions are, or would like to be, preserving content indefinitely– even if there are stated retention and review policies. If organizations have review periods, they are most often understood to provide the institutions with some flexibility: stated review policies provide a timeline in which content could be reviewed if necessary (e.g., due to rising storage costs, lack of use, or other concerns). However, at this point, reviewing content is incredibly difficult and expensive – and storage costs and demands have not tipped the point at which organizations need to expend energy and effort to remove content. All institutions noted that data would only be removed or deaccessioned in the event of contractual obligations, ethical or legal concerns, or an accidental deposit prior to formal publication and DOI minting.

Some members noted that preservation practices have evolved over time, and will continue to do so. This may pose a problem for previously curated content, which may be less robustly described or have been curated to different standards than are currently available. These research outputs will need to be addressed on a case-by-case basis.

The biggest difference between the academic institutions and the non-profit organizations (i.e., the Michael J. Fox Foundation and Dryad) is the sheer size and quantity of data to be managed. These institutions see a significant amount of data, either deposited into their repository or that needs to be managed across distributed platforms.

C. Q3: Dataset size

With the exclusion of the Michael J. Fox Foundation, which routinely uses and manages terabytes of data for longitudinal studies, all member institution representatives reported that the largest datasets were less than 10TB – and many were significantly smaller. In fact, all of the academic institutions reported that the average size of datasets could range from MB through to GB, but often the datasets over 1TB were outliers.

Many interviewees noted that there were often hurdles for researchers to deposit more than a threshold amount of data– such as 50GB– including cost-recovery models that start after a certain amount of data, or technical considerations that required researchers to connect with repository managers to deposit the data. All institutions reported that dataset sizes were increasing, and dealing with “Big Data” would likely become a more acute problem.

D. *Q4: Is software preserved alongside a dataset?*

Many repositories are seeing an increase in related research outputs– this primarily looks like code that is preserved alongside datasets. Moreover, of the repositories that are accepting code, many are seeing an increase in blended code and data. In particular, this poses problems for reusability– there are pressing concerns about the longevity of this code (especially with regards to backwards compatibility, such as the migration from Python 2 to Python 3, that required code to be adjusted for use).

A few institutions are participating in or closely monitoring other software (i.e., executable files) preservation (e.g., Software Preservation Network [25]) and emulation efforts (e.g., Emulation-as-a-Service Infrastructure, EaaSI [26]). There is not a concerted effort of this in the DCN, though.

E. *Q5: Cost -recovery model on dataset*

Very few organizations are adopting a cost recovery approach to data storage. For those that are, the costs only go into effect if a dataset reaches a particular threshold or amount per year; many of the institutions that have this policy written, though, expressed that there is often flexibility in the cost depending on conversations with the researchers. All of the academic members reported that they are increasingly working with researchers in the grant planning phase so the costs associated with long-term data preservation can be incorporated into grant budgets.

V. DISCUSSION

A. *Areas of overlap between DCN partners*

Many institutions expressed the sentiment, even if in different words, that they are doing the best they can with the resources – time, personnel, funding, etc. – that they have on hand. This often means preserving content in the areas they can, such as bit-level preservation, and kicking the proverbial can

down the road to advocate for and develop other tools, resources, and support to more effectively preserve content. For this reason, data curation is highlighted and reinforced as a critical preservation activity: curation activities like documentation and format migration provide more support than would otherwise be available, and thereby increase the likelihood that the research data that is deposited in institutional repositories will be accessible and reusable longer than data that does not receive the same curation interventions.

This is incredibly significant, because preserving research data deposited by researchers is vital. Losing research data would result in significant reputational harm to the institutions that are viewed as trustworthy stewards, as well as loss of researcher trust and damage to key relationships. Researchers giving their data to others to steward (in particular librarians, archivists, and other data management professionals) are implicitly trusting that their research outputs are being stewarded well. The long-term sustainability of curation services largely depends on this trust, and this trust could well be lost if the data are lost.

All representatives of the academic institutions reported that they are engaging in and leveraging pre-custodial preservation efforts to help create research data that is more preservable early in the research lifecycle. This includes key educational offerings, like webinars and trainings, as well as point of need consultations, and the emotional labor of forming relationships with faculty, staff, and students. As with archival practice, these interventions early in the research data lifecycle have significant positive impacts. In particular, during these interviews, DCN members reported that support they provided for researchers often meant their current and future research projects were significantly improved. For example, if a student learned how to properly create a codebook, or which formats would likely be more reusable (e.g., storing data as a csv instead of xlsx), it meant that they would be able to use that information throughout the course of their career.

Similarly, most interviewees reported that, at the point of need curation, data stewards would either automatically convert file formats when possible (e.g., transforming a word document to a PDF/A) or suggest alternative formats to researchers. In instances where content was transformed, often both versions would be retained and made available:

for example, a spreadsheet might be converted to csv files, but also made available as a Microsoft Excel Document so data re-users could also view images, formula, and macros that were original to the research. By adopting this approach, data curators are demonstrating their commitment to preserving content as best they can with what they have [27], and further garnering trust from researchers– by not just transforming data and deleting original content, researchers can see the commitment to preservation without feeling like their data have been manipulated without their consent.

Lastly, all interview participants tied the conversations back to reuse and the needs of future users. Throughout the conversations, when discussing their curation work, their policies and practices, and especially in discussing data retention, all interviewees and respondents emphasized that curating and preserving the data is done with the goal of access, reuse, and reproducibility. This builds on the sentiment set forth by Normand Charbonneau, International Council of Archives, 2019, that “Preservation without access is merely hoarding,” [28] and argues that access alone– while critical [29]– is not enough. Considering the needs of future data reusers is fundamental and is what drives curation and preservation activities for academic libraries.

B. Key Challenges

Despite the fact that each interview with institutional representatives was conducted individually, there emerged numerous opportunities to collaborate and solve common problems. While a few are presented here, it is worth noting that more challenges could emerge in wider discussions with more diversity of institutions represented.

A shared issue that emerged was the challenge of increasingly complex datasets. In particular, datasets that blend software, code, and other data formats, that also rely on the existence of one another, are increasingly being deposited into institutional repositories by researchers. These pose intellectual issues, as the data, software, and code are subject to different copyright and licensing requirements. While data stewards continue to provide content to the best of their present abilities, future repositories will need to accommodate these increasing complications.

An area of future collaborative efforts is preservation metadata, or PREMIS [30]. While this metadata standard is crucial for ensuring the long-term preservation of data, the burden of creating this metadata by hand is a significant deterrent for many data curators– and in fact, the standard is not intended or designed to be created by hand. The standard is complex, and without a system automatically generating PREMIS, it is both difficult and time-consuming to attempt to create this standardized metadata on a one-off basis and is not sustainable for long-term, large-scale preservation efforts. In addition to collaborating with preservation professionals to better understand when and how to create this metadata, there is the opportunity to develop tools to better integrate a Curator’s Log into any preservation metadata, which is a plain text object that can be used by data stewards to record any changes made to the research outputs, as well as correspondence with data authors [31].

The final opportunity for collaboration is in review and retention policies, workflows, and tools. Many institutions noted that they had no plans to review content in the next 5-10 years, and even those that have stated review policies remarked that these were more to give the library the leeway to remove content as needed, and would likely not be implemented unless storage costs grew unwieldy. Participants noted that, at present, storage costs were low enough that it was likely more expensive to review materials, in terms of labor costs, than to just continue paying for storage. Moreover, the guidelines by which content would be reviewed for removal were unclear – usage and size were two that were frequently mentioned, but with the caveat that those are potentially flawed, as past use does not indicate future use of content, and reviewing content by data size will result in a disproportionately higher number of large datasets being reviewed and potentially removed. Data stewards, librarians, archivists, and other information professionals can and should collaborate to not only develop tools and workflows for reviewing content– this could look like automating review of content leveraging machine learning, rubrics that encourage a holistic evaluation of the reusability of data, and documented practices that other institutions can leverage to develop their own. Academic libraries should also adopt and employ clear policies to ensure the crucial trust built up with researchers over time is not lost if and when research outputs are deaccessioned.

VI. LIMITATIONS

This project had significant limitations– in particular, the isolated nature of the conversations did not allow other DCN members to participate in collaborative discussions. In other words, members could not ask follow up questions of one another, identify shared challenges together, serendipitously remind one another of practices, etc. While this method was adopted to provide flexibility and to avoid the struggle of scheduling 16 individuals, it significantly limited the ability for building off one another and engaging in peer to peer comparison. Moreover, the semi-structured interviews meant that conversations often went in different directions: for example, if an institution has recently or is in the process of applying for the CoreTrust Seal [32], conversations were far more likely to be focused on technical specifications. This means that some institutions reported more technical information than others, or not mentioned the information was asked, making it difficult to compare. Future research projects can and should engage in more technical conversations with partners to understand these needs in greater detail.

Moreover, these 16 institutions – 14 academic, 1 generalist repository, and 1 funding agency– are not representative of research data management writ large. There are other repositories, academic institutions, and research data stewards that are not represented in this exploratory report. More holistic research on preservation practices across the research lifecycle and from different institution types would be of significant value. Similarly, more structured and quantitative investigations, such as by using the NDSA Levels of Digital Preservation [33], to more directly compare repositories will be essential in identifying areas of challenge and opportunities that could have a wide-reaching impact. For example, if many different institutions would benefit from tools that automate PREMIS or support in the systematic review for retention work, it will provide more incentive for collaboration.

Lastly, due to the exploratory nature of this report, partners were not asked to describe in detail the preservation staffing of their institution. Given that preservation also requires a significant commitment of personnel [34, 35, 36] and maintenance labor [37], future research should seek to understand how, in addition to the work of data curators, data intensive repositories and institutions are committing personnel to the work of preserving

research data and supporting open science and data reusability.

VII. CONCLUSION

Despite the limitations, this undertaking revealed some significant challenges our members are facing with research data preservation. Through conversations with representatives from DCN partner institutions, it is clear that all partners are preserving research data to the best of their current abilities within their institutional confines. By engaging in critical pre-custodial work, and supporting researchers in curation at their point of need, Data Curation Network members are improving the overall FAIRness of research outputs.

While the benefits of data curation will likely make these research data more preservable in the long run, there are a few areas that need to be collaboratively addressed for more robust preservation of research data across both DCN members as well as research institutions. This includes, but is not limited to, retention and review policies, tools, and workflows, creating and managing PREMIS metadata, and the increasing complexity and intricacies of research outputs.

In addition to the limitations and future research described above, data stewards and preservation practitioners can and should collaborate to better understand which curation activities, in particular, enhance the value of preserved datasets as well as which activities most directly impact the preservability of datasets.

Preservation is an ongoing process– and one that is evolving with new technologies, data formats, and tools to support librarians and archivists. By sharing information about our current practices, our pain points, and identifying opportunities for collaboration, we can enhance our capacity and knowledge. Much like the DCN's approach to curation, by engaging in peer comparisons and other information sharing efforts, we can collaborate to continuously improve our best practices and standards.

ACKNOWLEDGMENTS

The authors would like to thank the interviewees and other colleagues who provided feedback on this work in various stages. This includes: Lisa Johnston, Jake Carlson, Wendy Kozlowski, Josh Gottesman,

Daniella Lowenberg, Andrew Johnson, Renata Curty, Vicky Rampin, Katie Wissel, Scout Calvert, Wind Cowles, Hannah Hadley, Matt Chandler, Sophia Lafferty-Hess, Joel Herndon, Seth Erickson, Chen Chiu, Jennifer Moore, Peggy Griesinger, Heather Coates, and Carol Kussmann.

REFERENCES

- [1] Office of Science and Technology Policy. (2021). "2021 Public Access Congressional ReportPublic Access Congressional Report". Accessed March 04, 2022. https://www.whitehouse.gov/wp-content/uploads/2022/02/2021-Public-Access-Congressional-Report_OSTP.pdf
- [2] Melero, R. and Navarro-Molina, C. (2020), Researchers' attitudes and perceptions towards data sharing and data reuse in the field of food science and technology. *Learned Publishing*, 33: 163-179. <https://doi.org/10.1002/leap.1287>
- [3] Navale, Vivek, and McAuliffe, M. "Long-term preservation of biomedical research data." *F1000Research* vol. 7 1353. 29 Aug. 2018, doi:10.12688/f1000research.16015.1
- [4] Murray, M., O'Donnell, M., Laufersweiler, M., Novak, J., Rozum, B., & Thompson, S. (2019). A survey of the state of research data services in 35 US academic libraries, or "Wow, what a sweeping question". *Research Ideas and Outcomes*, 5, e48809. <https://doi.org/10.3897/rio.5.e48809>
- [5] Higgins, S. (2012). The lifecycle of data management. Managing research data, 17-45.
- [6] Data Curation Network (2018). "Checklist of CURATED Steps Performed by the Data Curation Network." <http://z.umn.edu/curate>.
- [7] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
- [8] Johnston, L.R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., Stewart, C., Blake, M., Herndon, J., Mcgeary, T.M., Hull, E., and Coburn, E. 2018. Data curation network: A cross-institutional staffing model for curating research data. *International Journal of Digital Curation*, 13(1), pp.125-140. <https://doi.org/10.2218/ijdc.v13i1.616>.
- [9] Johnston, L.R. (2020) How a network of data curators can unlock the tremendous reuse value of research data. *OCLC Next blog*. <http://www.oclc.org/blog/main/data-curators-network/>
- [10] Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., ... Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19(1), 43. DOI: <http://doi.org/10.5334/dsj-2020-043>
- [11] Johnston LR, Carlson JR, Hswe P, Hudson-Vitale C, Imker H, Kozlowski W, Olendorf RK, Stewart C. (2017) Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions' Data Repository and Curation Services. *Journal of eScience Librarianship* 6(1): e1102. <https://doi.org/10.7191/jeslib.2017.1102>.
- [12] Hudson-Vitale, C., Hadley, H., Moore, J., Johnston, L., Kozlowski, W., Carlson, J., & Herndon, J. (2020). Extending the Research Data Toolkit: Data Curation Primers. *International Journal of Digital Curation* 15 (1).
- [13] For example: "Storing and preserving research data," published by Utrecht University. Accessed March 6, 2022: <https://www.uu.nl/en/research/research-data-management/guides/storing-and-preserving-data>
- [14] Lavoie, B. (2021) "Preserving Research Data." *Digital Preservation Coalition Blog*. Accessed March 6, 2022: <https://www.dpconline.org/blog/wdpc/preserving-research-data>
- [15] Johnston LR, Carlson JR, Hswe P, Hudson-Vitale C, Imker H, Kozlowski W, Olendorf RK, Stewart C. (2017) Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions' Data Repository and Curation Services. *Journal of eScience Librarianship* 6(1): e1102. <https://doi.org/10.7191/jeslib.2017.1102>.
- [16] Tenopir, C., Sandusky, R. J., Allard, S., & Birch, B. (2014). Research data management services in academic research libraries and perceptions of librarians. *Library & Information Science Research*, 36(2), 84-90.; Gowen, E., & Meier, J. J. (2020). Research Data Management Services and Strategic Planning in Libraries Today: A Longitudinal Study. *Journal of Librarianship and Scholarly Communication*, 8(1).
- [17] Carlson, J., & Johnston, L. R. (Eds.). (2015). *Data Information Literacy: Librarians, Data, and the Education of a New Generation of Researchers*. Purdue University Press. <http://www.jstor.org/stable/j.ctt6wq2vh>;
- [18] Johnson KA, Steeves V. (2019) Research Data Management Among Life Sciences Faculty: Implications for Library Service. *Journal of eScience Librarianship* 8(1): e1159. <https://doi.org/10.7191/jeslib.2019.1159>.
- [19] Johnston, L. R. & Carlson, J. & Hudson-Vitale, C. & Imker, H. & Kozlowski, W. & Olendorf, R. & Stewart, C., (2018) "How Important is Data Curation? Gaps and Opportunities for Academic Libraries", *Journal of Librarianship and Scholarly Communication* 6(1), p.eP2198. doi: <https://doi.org/10.7710/2162-3309.2198>
- [20] Johnston, Lisa R; Carlson, Jake; Hudson-Vitale, Cynthia; Imker, Heidi; Kozlowski, Wendy; Olendorf, Robert; Stewart, Claire. (2016). Definitions of Data Curation Activities used by the Data Curation Network. Retrieved from the University of Minnesota Digital Conservancy, <https://hdl.handle.net/11299/188638>.
- [21] Cunningham, A. (1994). The archival management of personal records in electronic form: Some suggestions. *Archives and Manuscripts*, 22(1), 94-105.
- [22] Weisbrod, D. (2016). Cloud-supported preservation of digital papers: A solution for special collections?. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 25(3), 136-151.
- [23] Searcy, R. (2017). Beyond control: Accessioning practices for extensible archival management. *Journal of Archival Organization*, 14(3-4), 153-175.
- [24] Navale, Vivek, and McAuliffe, M. "Long-term preservation of biomedical research data." *F1000Research* vol. 7 1353. 29 Aug. 2018, doi:10.12688/f1000research.16015.1
- [25] Vowell, Z., Hagenmaier, W., Rios, F., & Roke, E. R. (2017). The software preservation network (SPN): a community effort to ensure long term access to digital cultural heritage. *D-Lib Magazine*, 23(5/6).
- [26] Cochrane, E., Peer, L., Gates, E., & Anderson, S. (2019). Saving Software and Using Emulation to Reproduce Computationally Dependent Research Results.
- [27] Johnston, L. R. (2014). Developing a data curation service: Step #1: Work with what you've got. *Bulletin of the*

Association for Information Science and Technology, 40(4), 45–47. <https://doi.org/10.1002/bult.2014.1720400416>

- [28] Normand Charbonneau. (2019) "SESSION 5.4 / P100 Public Services & Outreach and Appraisal: two new expert groups." Presented at the International Council on Archives.
- [29] The Committee For Film Preservation and Public Access. (1993) "Preservation without Access is Pointless." Accessed March 6, 2022: <https://www.loc.gov/static/programs/national-film-preservation-board/documents/fcmtefilmprespubaccess.pdf>
- [30] PREMIS Editorial Committee (2015). PREMIS Data Dictionary for Preservation Metadata, version 3.0. Accessed March 6, 2022: <https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf>
- [31] See "Curator's Log" in: Johnston, Lisa R; Carlson, Jake; Hudson-Vitale, Cynthia; Imker, Heidi; Kozlowski, Wendy; Olendorf, Robert; Stewart, Claire. (2016). Definitions of Data Curation Activities used by the Data Curation Network. Retrieved from the University of Minnesota Digital Conservancy, <https://hdl.handle.net/11299/188638>.
- [32] Dillo, I., & Leeuw, L. D. (2018). CoreTrustSeal. Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen & Bibliothekare, 71(1), 162-170.
- [33] National Digital Stewardship Alliance (NDSA). Levels of Digital Preservation 2.0. Accessed March 6, 2022: <https://osf.io/2mkwx/>
- [34] D. Waters and J. Garrett, "Preserving digital information: Report of the task force on archiving of digital information," The Commission on Preservation and Access, Washington, DC, 1996. [Online]. Available: <https://www.clir.org/wp-content/uploads/sites/6/pub63watersgarrett.pdf>
- [35] A. Kay Rinehart, P.-A. Prud'homme, and A. Reid Huot. Overwhelmed to action: Digital preservation challenges at the underresourced institution. *OCLC Systems & Services: International Digital Library Perspectives*, vol. 30, no. 1, pp. 28–42, Apr 2014.
- [36] T. Owens, *The Theory and Craft of Digital Preservation*. Baltimore: Johns Hopkins University Press, 2018.
- [37] D'ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT press.

VIII. APPENDIX A: RECRUITMENT EMAIL AND INTERVIEW QUESTIONS

[Greetings]

I'm reaching out today as I am diving into my projects for this year, and starting off is an exploration of the digital preservation capabilities of DCN Partners. This builds on previous work conducted in the DCN around technical capabilities and specifications, but focuses specifically on preservation practices. I am reaching out to all DCN Member institutions to learn more about preservation efforts to help understand what resources or support would be most beneficial. I am

hopeful to report the results to DCN members, as well as in a publication and/or presentation later this year.

To that end, I was wondering if you and/or someone at [member institution name] would be able to meet briefly with me to discuss your current preservation practices? I anticipate that this would take about 45 minutes– I've pasted my general questions below, to give you an idea of what I am looking for, but might also have follow-up questions specific to your institution.

Thank you for your consideration– please let me know if I can provide you with any additional information!

Questions:

-With regards to research data, what does preservation look like at your institution? What preservation activities are currently being taken on research data? (related: Is the data in the institutional repository? Or a standalone repository?)

-How long are data retained? Are there retention and deaccession policies, or retention review periods? If so, who is responsible for making deaccession decisions? Have these been implemented and used to delete data?

-Could you provide an approximate range of dataset size (e.g., from 1 GB to 7 TB)? Is there a maximum amount researchers can deposit?

-Is software preserved alongside a dataset, when appropriate?

-Do you employ a cost-recovery model on datasets?

APPENDIX B: SUMMARY TABLE

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software Preservation alongside data	Q5 - Cost Recovery Model
Cornell University	<ul style="list-style-type: none"> - Research data kept with other content types in IR. - Working on integrating library-managed platform for preservation in later 2022 - IR contents backed up daily, in a separate building than where the repository servers are physically located. 	<ul style="list-style-type: none"> - Retained "indefinitely". - At this moment, there is no retention/ deaccession and review policies in place. - Seeing the need to develop a deaccession policy as the growth of collection size to reduce the storage cost. 	<ul style="list-style-type: none"> - Per policy, file size is less than 5GB and total submission limits of 50GB. Exceptions made in rare circumstances. - Currently, datasets range from single file spreadsheets in MB to the largest of 107GB. 	<ul style="list-style-type: none"> - Currently, not a regular practice. - Custom software written for analysis which is considered as a part of the dataset package, will be made available, if applicable. - Request version and access information in documentation to facilitate access to the software. 	<ul style="list-style-type: none"> - No cost recovery model.
University of Colorado Boulder	<ul style="list-style-type: none"> - Datasets are stored in IR, leveraging Samvera/Fedora. - 2 copies are stored in AWS and 1 in PetaLibrary. - Files over 10GB are loaded directly to PetaLibrary. - Backup and fixity checks are managed by PetaLibrary. 	<ul style="list-style-type: none"> - Retained "indefinitely". - Funded projects that generate datasets over 500GB are kept for 10 years and reassessed. - No review policy yet. - Deaccession happens when there is a violation of copyright/ethical issue. Data will be transferred to cold storage. 	<ul style="list-style-type: none"> - Range from MB to 500 GB. 	<ul style="list-style-type: none"> - Currently, no preserve software/executable files separately. - Having many instances of blended code and data. 	<ul style="list-style-type: none"> - Fees applied for over 500GB datasets.

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software Preservation alongside data	Q5 - Cost Recovery Model
	<ul style="list-style-type: none"> - Globus is used for uploading and downloading large files. - Curators encourage file format transformation at ingestion stage. 				
Dryad (information shared via email)	<ul style="list-style-type: none"> - Rely on CoreTrustSeal certified repository (Merritt) for preservation 	<ul style="list-style-type: none"> - Retained "indefinitely". 	Not mentioned	Collaboration with Zenodo, which preserves the software.	Yes.
Duke University	<ul style="list-style-type: none"> - Data stored in a repository separate from IR, but preservation approached at Duke in a holistic manner. - Stores 4 copies of content, all based in Durham. Planning to move one copy to the cloud. - Fixity is checked prior to upload to repository; leverage BagIt; repository technology generates checksums. Virus checking 	<ul style="list-style-type: none"> - Currently drafting a Retention Policy (considering 25 years as a minimum unless the depositor is paying for less). - Deaccessioning would likely only occur due to takedown requests (e.g., legal or disclosure reasons). - Acknowledge that the review process would be time consuming and may need to involve machine learning. 	<ul style="list-style-type: none"> - Largest data is 100s of GB, most under 100GB, many under 10 GB. - Leveraging Globus for large upload and downloads. 	<ul style="list-style-type: none"> - Lots of code files, but not many executables; watching environments/containers (e.g., code ocean). - Recommending GitHub+Zenodo workflow for software archiving 	<ul style="list-style-type: none"> - Cost recovery after 100GB.

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software Preservation alongside data	Q5 - Cost Recovery Model
	<p>is also automated via ClamAV but cannot support very large files.</p> <ul style="list-style-type: none"> - Not using PREMIS, but using different components of preservation metadata in other ways/fields, work with a metadata specialist. - Working on further standardization of normalization processes. - File formats are tracked and recommendations are provided to all library end users (link); when able and appropriate normalize files on ingest and no formal policy around future migration. - Recently completed a self-assessment using the NDSA levels of preservation, 				

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software alongside Preservation data	Q5 - Cost Recovery Model
	preparing for CoreTrustSeal application.				
University of Illinois, Urbana-Champaign	<ul style="list-style-type: none"> - Data is stored in a data repository separate from IR. Built in-house, on top of Medusa, the Library preservation system. - Medusa manages fixity checks. - Multiple copies: in Medusa and on AWS. - Does not automatically convert files, but suggests alternative formats. - Does not automatically convert files, but suggests alternative formats. - Submitting a CoreTrustSeal application recently. 	<ul style="list-style-type: none"> - Minimum data retention is 5 years, then go through a robust review and the 5 years starts again. - Deaccession happens due to ethical or legal reasons. 	<ul style="list-style-type: none"> - 2 TB/per/faculty - Most datasets are in the GB range and less than 100GB, with some outliers. 	<ul style="list-style-type: none"> - Accepts files in any format, including software. 	<ul style="list-style-type: none"> - Currently no cost recovery
John Hopkins University	<ul style="list-style-type: none"> - Files (and documentation) submitted to OneDrive, JHU Data 	<ul style="list-style-type: none"> - 5 year retention and review, but in reality, data are likely retained indefinitely. This might change with 	<ul style="list-style-type: none"> - All dataset submissions are under 1TB right now, with the largest being 	<ul style="list-style-type: none"> - Yes, code and data. 	<ul style="list-style-type: none"> - Cost recovery (e.g., charge for storage) after 1TB of data

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software Preservation alongside data	Q5 - Cost Recovery Model
	<p>Management Consultants curate, upload and enter metadata to Dataverse (JHU Data Archive). Use a homegrown packaging tool (https://github.com/DataConservancy/dcs-packaging-tool) to package the data and submit to dark storage (managed by Libraries IT).</p> <ul style="list-style-type: none"> - Data are removed from OneDrive after data collection is published on JHU Data Archive and an archival package is made. - No file format migrations after curation; migrations happen during curation (e.g., xlsx to csv) 	<p>upcoming changes to JHU University wide policies that mandate 7 year retention of data.</p> <ul style="list-style-type: none"> - In the extremely rare event of deaccessioning data, that decision would be made by the data librarians and other consultants as appropriate 	<p>video files.</p> <ul style="list-style-type: none"> - Most datasets are under 10GB. - Using Globus for transferring larger datasets. - Still working on a solution for users to download big dataset. 		
Michael J. Fox Foundation	<ul style="list-style-type: none"> - Research data are created / collected primarily in two ways: 	<ul style="list-style-type: none"> - Retained indefinitely, unless contractually obligated to delete. - Deaccession depends on the 	<ul style="list-style-type: none"> - Largest datasets are hundreds of TB; not infrequently in the ~1TB range; many 	N/A	N/A

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software Preservation alongside data	Q5 - Cost Recovery Model
	<p>research studies funded and operated by MJFF, and that from studies operated by external researchers and funded by MJFF.</p> <p>- Data is decentralized, focused on supporting reusability of data; data access is priority.</p> <p>- There are MJFF repositories available: MJFF can deposit data in these, but the data can also be shared in other repositories (typically only for MJFF operated studies, at this time - i.e., data generated by researchers using MJFF funding but in studies not operated by MJFF are not necessarily able to be shared in multiple</p>	<p>governance structure of a given data set, but likely some combination of MJFF + study investigators.</p>	<p>others are under this threshold.</p>		

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software alongside Preservation data	Q5 - Cost Recovery Model
	places). - Preservation practices have evolved over time; currently have two AWS options, one for nearline data and one for cold-storage (AWS Glacier); some physical servers remain.				
University of Michigan	- Three different levels of preservation for all repositories: 1-- open source, more easily preserved; 2-- proprietary but popular (e.g., PDF, excel), we will do our best to preserve based on the information available to us; 3-- closed, can only promise bit-level preservation. - Leveraging a digital preservation team and task force on	- Retained for 10 years. - Unclear review policy for now.	- 100 files or 5GB, more than this and the researcher will need to contact Deep Blue for upload help. - For content more than 1TB in size, we ask researcher to complete the Large Data Conversation form; evaluating both technical cost of time to manage/upload, as well as financial considerations	Yes, and seeing more blended software and code.	- No cost recovery model

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software Preservation alongside data	Q5 - Cost Recovery Model
	<p>digital formats moving forward-- library-wide efforts.</p> <ul style="list-style-type: none"> - There is redundancy in storage, but not as geographically dispersed as ideal; Currently exploring the use CLOCKSS for more redundancy. - Deep Blue Data relies on a mediated deposited; researchers deposit their own data (depending on size and number of files), but admin determine when to publish; have removed datasets that were in draft mode (not yet published), but would only remove published data if ethical/legal concerns (but that has not happened yet) 				

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software Preservation alongside data	Q5 - Cost Recovery Model
University of Minnesota	<ul style="list-style-type: none"> - CoreTrustSeal certified. - Format agnostic but not format blind- in other words, depositors can share any formats, but curators make recommendations. - Through the web interface, users can deposit 150 GB; can deposit more through repository staff. - Working copies of data are also retained in a dark archive. - Preemptive preservation through curation and documentation 	<ul style="list-style-type: none"> - 10 year review period (but not required to remove content). - Concern: if/when time to review will be prejudiced against file formats and size. - Deaccession for legal or ethical reasons at the moment 	150 GB limit through self-deposit interface; larger datasets accepted with collaboration with repository staff.	Yes, and seeing more blended software and code.	-No cost recovery model
University of Nebraska Lincoln	<ul style="list-style-type: none"> - Use Rosetta (Ex Libris) for preservation; fixity checks, file format validation, and technical metadata are 	<ul style="list-style-type: none"> - Depositors can select 5 year or 20 years for storage upon deposit into the repository; but not currently deaccessioning. 	<ul style="list-style-type: none"> - Datasets are in the GB range right now; those with larger datasets tend to have their own infrastructure already. 	<ul style="list-style-type: none"> - Accepting software and code, but promise is bit level preservation; migration of files 'as resources allow'-- gives flexibility to migrate, but not required to. 	<ul style="list-style-type: none"> - Cost recovery model only begins at 1TB.

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software Preservation alongside data	Q5 - Cost Recovery Model
	all automated processes. - During curation (fully mediated deposit) data curators advise on file formats and descriptive practices; data is backed up with copies on self-hosted library servers.				
New York University	- Promising bit-level preservation; monitoring fixity, numerous copies and geographic distribution - Content deposited into institutional repository (which also houses data, code, etc.) is replicated to library tech for maintenance; they do not add additional metadata or documentation. - NYU does not, as a practice, automatically convert file formats.	- Deaccession for legal or ethical reasons	- No file size limit	- Accepts code and data-- format agnostic	- No cost for storage

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software Preservation alongside data	Q5 - Cost Recovery Model
	<p>Instead, suggests potential file transformations and documentation when appropriate.</p> <ul style="list-style-type: none"> - Adopts a "curation up front" mentality, and leverages education and consultation to help researchers build data and code better before deposit; also uses the time of need as an educational opportunity when possible. Pre-custodial education and outreach. 				
Pennsylvania State University	<ul style="list-style-type: none"> - Data stored in the IR, Scholarsphere (locally built and maintained); unmediated self-deposit, curation often happens post ingest; some things can be done without contact depositor, but 	<ul style="list-style-type: none"> - Retained for 10 years. - Developing guidelines for how to assess content after the 10 year mark. - Developing guidelines for what content is sent to the preservation system (in progress). - Deaccession due to content concerns (e.g., doesn't fit in the content policy) 	<ul style="list-style-type: none"> - Most datasets are less than 100GB-- primary user base is in the 'middle data' -- not big data, but bigger than small data. - Considering how to grow to accommodate large datasets-- know this will be a need 	<ul style="list-style-type: none"> - Have many instances of data that is blended with code (e.g., R scripts); - Also part of the Software Preservation Network. - Licensing is a challenge for software preservation. 	<ul style="list-style-type: none"> - No cost to deposit.

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software Preservation alongside data	Q5 - Cost Recovery Model
	<p>some require connecting with depositor (which can be difficult); concern is capacity-- not enough time.</p> <p>- Scholarsphere is understood more as an access platform; minimum preservation (bit level- fixity and multiple copies in AWS).</p> <p>- Leveraging Globus.</p>				
Princeton University	<p>- Multiple copies, checksums, suggesting alternative formats for long-term access or use</p> <p>- The library, as a policy, does not change formats, but encourages researchers to change formats when appropriate. Can store both formats if needed/requested. Interested in automatic</p>	<p>- No deaccession policies, but also no firm commitments to researchers for how long data will be preserved</p>	<p>- Most datasets are less than 200MB; the current largest dataset is 375 GB.</p> <p>- No technical limit to datasize, but practical limit in download/upload.</p> <p>- Public Globus access point to support those.</p>	<p>- Accepts data and code-- mild increase in number of code deposits; no requirements for docker, code ocean, etc; curators encourage documenting environment, additional libraries, code versions, etc.</p> <p>- Challenge when the code and data are more blurred</p> <p>- Licensing data and code together is a real challenge</p>	<p>No cost to researchers-- the repository is funded by the Provost.</p>

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software Preservation alongside data	Q5 - Cost Recovery Model
	<p>conversions.</p> <ul style="list-style-type: none"> - Fully mediated deposit - Currently, data deposited into a DSpace instance mixed in with other content 				
University of California, Santa Barbara	<ul style="list-style-type: none"> - Leverages Dryad institutional membership. See Dryad. 	See Dryad.	See Dryad.	See Dryad.	See Dryad.
Virginia Tech	<ul style="list-style-type: none"> - Leverages Figshare for Institutions for data (which is separate from the IR). - Upon submission, data and metadata is bagged and stored on Google Drive and in an AWS bucket - At this point, curation happens: suggesting documentation, file formats, tracking emails and changes in provenance); no standard 	<ul style="list-style-type: none"> - Retained for 5 years. - Not worry about review at the moment due to the small size range. - Deaccession due to for ethical / legal issues (outside of a reappraisal process) 	<ul style="list-style-type: none"> - Datasets range in size from MB to 370GB. 	<ul style="list-style-type: none"> - Store some code and data-- instances of blended code and data. 	<ul style="list-style-type: none"> - No cost for depositing datasets, but can be tricky technically after 50GB. - Have totaled up some costs for grant applications

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Software alongside Preservation data	Q5 - Cost Recovery Model
	<p>or automatic file format conversions</p> <ul style="list-style-type: none"> - All files and curation provenance log (and metadata) are bagged into AIP, stored on Google Drive and AWS; hopeful for APTTrust as a storage point, but still exploring 				
Washington University in St. Louis	<ul style="list-style-type: none"> - Want to implement Archivematica or other more robust preservation tool in the future - Digital assets in the library are somewhat siloed, but hopeful the Digital Preservation Librarian (currently being hired) will support in aligning repositories - SIPs are combined with DIPs, provenance information, and other documentation 	<ul style="list-style-type: none"> - Retained for 10 years, at minimum. - Deaccession due to error in deposit (e.g., meant to deposit in IR) or for legal/ethical concerns. 	<ul style="list-style-type: none"> - Average dataset size falls below 150GB which is the maximum amount researchers can deposit for free. - Some large datasets are in progress, but have not been deposited yet 	- Some instances of code and data but few	No cost recovery model.

Institution	Q1 - Preservation Activities	Q2 - Data Retention/ Deaccession	Q3 - Dataset Size	Q4 - Preservation alongside data	Q5 - Cost Recovery Model
	<p>to create an AIP, which is deposited on a storage gateway managed by campus OIT -- there is currently no fixity run on these</p> <ul style="list-style-type: none"> - Pain point: PREMIS. How to manage/create this without a tool like archivematica and without writing xml every time? - In curation step, suggesting changes to file formats; but, if there is a working relationship with the depositor or a low labor ask, may convert for the researcher 				

FEASIBLE, ADAPTABLE, AND SHARED:

A Call for A Community Framework for Implementing ML and AI

Abigail Potter

*Library of Congress
USA
abpo@loc.gov*

**Eileen Jakeway
Manchester**

*Library of Congress
USA
ejakeway@loc.gov*

**Meghan Ferriter,
PhD**

*Library of Congress
USA
mefe@loc.gov*

Jamie Mears

*Library of Congress
USA
jame@loc.gov*

Abstract - Through open research, experimentation and convenings with LAM sector peers and colleagues, a foundational need has emerged for a broadly shared and evidenced set of guidelines for implementing ML and AI technologies that centers the long-term stewardship and ethical responsibilities of cultural heritage organizations. Inspired by community guidelines that rationalize complex information into an understandable framework like the NDSA Level of Digital Preservation and the Data Nutrition Project, LC Labs is proposing a step toward collaboratively generating a LAM-specific framework for understanding and implementing ML and AI technologies.

Keywords - Community guidelines, machine learning, artificial intelligence, transparency, experimentation

Conference Topics - Community, Resilience

I. INTRODUCTION

Another wave of technical change is at the door of many libraries, archives and museums (LAMs). The promise and claims of artificial intelligence (AI) systems to transform organizations and solve entrenched challenges with data-driven results and solutions are enticing. Especially when users are expecting consistent and sophisticated search and discovery systems across all formats and content types. In addition to the shared challenge of user expectations, cultural heritage and research

organizations have limited budgets, technical staff and expertise in implementing AI-driven services. As a result, formal and informal networks are forming to develop and share strategies and practices for dealing with this latest wave of transformation.

Through open research, experimentation and convenings with LAM sector peers and colleagues, a foundational need has emerged for a broadly shared and evidenced set of guidelines for implementing ML and AI technologies that centers the long-term stewardship and ethical responsibilities of cultural heritage organizations. Inspired by community guidelines that rationalize complex information into an understandable framework like the NDSA Levels of Digital Preservation [1] and the Data Nutrition Project [2], LC Labs is proposing a step toward collaboratively generating a LAM-specific framework for understanding and implementing ML and AI technologies.

II. RESILIENCE THROUGH COMMUNITY

Despite outstanding efforts to digitize and preserve historical materials, the information they hold remains difficult to use computationally, fragile to sustain, and unwieldy for systems that serve modern user needs. AI systems hold promise for solving our technical and data challenges. However,

digital library, archive and museum collections generally need to be transformed to be used by data-centric technologies and tools like machine learning (ML) and artificial intelligence (AI). And these systems are generally sold by vendors. In any AI or ML process there is potential for distortion or loss of context at each stage of transformation; this risk is exacerbated when proprietary algorithms are used. Understanding the potential for generating positive impacts for the users of LAM collections verses the potential for harm in using untested technologies has spurred the Library of Congress Labs team (LC Labs) to sponsor public experiments in machine learning to gather evidence about benefits and risks before making large scale investments in implementing what is often proposed by vendors as a soup-to-nuts AI solution.

With each technical advance, be it creating machine-readable bibliographic records, digitizing collections, making content available online, navigating digital publishing and social media, and managing and preserving digital collections, the LAM community has developed shared tools, practices and standards to respond to a changing technical landscape. These tools and standards, like the MARC record standard, FADGI digitization guidelines, the WARC format for preserving web archives, and various file format identifiers were developed through trial, error and committee agreements to benefit users, improve institutional practices and give guidance to staff who often have to train themselves on the latest technologies and advancements.

III. COMMUNITY AI PRACTICE

The LAM community has seen immense benefit from reports demonstrating the imperative of ethical adoption and research agendas to inform use and proliferation of AI in cultural heritage [3] [4] [5] [6]. In addition to these resources, events and workshops have brought together practitioners and leadership to highlight the practical challenges in this problem space [6] [7] [8]. Ongoing communities of practice continue to share the outcomes of their regularly convening, including the AI4LAM community. Additionally, there are vibrant existing and emerging disciplinary collaboratives, which present opportunities for the LAM community to engage more deeply around how ML and AI can perpetuate legacies of silence, harm, and structures of power. These activities offer the potential to synthesize

practices and facilitate knowledge exchange and evaluation.

Concurrent to these community initiatives and knowledge sharing activities, LC Labs sponsored research and experimentation has generated evidence, surfaced complexities in applying methods, and produced recommendations shared widely to benefit the community. Through experimentation, research, collaboration, and reflection, LC Labs works to realize the Library's vision that "all Americans are connected to the Library of Congress" by enabling the Library's Digital Strategy [11]. While pursuing this line of experimentation and convening practitioners, LC Labs staff have encountered challenges shared by a wider community. Before discussing LC Labs experimentation toward a community framework for ML and AI, we will briefly summarize some of those challenges.

IV. CHALLENGES TOWARD A FRAMEWORK

Despite these promising activities and the shared needs surfaced from these community activities and events, challenges remain. The landscape of available and effective methods is rapidly evolving, as are organizations as they test and even expand capacity to adopt and implement these methods. A shared framework would likely address a range of challenges, including these challenges in taking a first set of steps into this practice.

Distortion of and the loss of context in the production of digital collections are issues [12] that go back to collection acquisition, selection, description, digitization, management, online availability and then mass digitization. One of the key challenges of the transformations done by ML and AI technologies is the lack of transparency in decision-making at the human and systems level. Interpreting viability and nuance within the results of ML applications requires human expertise, as well as clear articulation of each step in a project's lifecycle; to include decisions of inclusion, exclusion, availability, and source and training data dimensions.

Experimentation and iteration should be essential to adopting approaches and a framework and its support for implementation. However, we acknowledge that creating space and leadership buy-in for experiments and pilots—and even

prototyping—is more complicated in practice than on paper. At this time, it is particularly difficult to convey—in advance of undertaking initiatives—consistent predictions about resource requirements, risk, complexity, and user and organizational needs; precisely because these types of information are gathered through the process of undertaking this work. Following the completion of an ML or AI project, it can be a challenge to immediately assess impact and define coherent next steps in advance of broader evidence.

Moving beyond project level implementation to more systematic exploration remains a specific challenge in fields in which resources have not yet been allocated for wider programmatic implementation. Additionally, as frequently shared outcomes and related effective practice tend to represent discrete projects rather than broad implementation, comprehensive approaches will require greater preparation. As communities of practice like AI4LAM gather people for knowledge exchange and comparison of approaches at that project level, transitioning to broader adoption within an organization would benefit from a shared framework.

Even with community reporting, the parts of the work most often highlighted in these community presentations represent the outcomes of those projects and the methods employed. However the team and organizational dimensions are less frequently foregrounded, which leaves opaque essential methods for integrating subject matter expertise, staff competencies, and other critical considerations for the people involved in undertaking these projects.

This brief discussion of these challenges suggests that a shared framework may allow staff and leadership to take steps into practice. A shared framework might further present opportunities to experiment with intention, document data transformation and consequences, and suggest starting points to evaluate approaches for broader implementation.

V. LC LABS EXPLORING ML

Recent LC Labs initiatives have demonstrated the complexity inherent to benchmarking. Furthermore, it is imperative that the intersection of project and organizational objectives include opportunities to

assess resources, collections, risk, and people encountered at each step. Committing to effectively centering people including users, staff and subjects of digitized items means that we must move with intention and integrate moments and mechanisms to ask critical questions of the approaches we are applying.

For the last several years, the LC Labs team has explored dimensions of machine learning through events, initiatives and experiments. We have hosted events, sponsored experiments and research, explored user needs from a range of angles, and frequently shared the outcomes of our work as part of our practice at LC Labs. We hosted a Machine Learning + Libraries Summit, alongside US- UK Digital Scholarship workshop in 2019 which also surfaced ML + crowdsourcing threads. From internal experimentation with Speech to Text Viewer to recommendations around socio-technical assessment and planning with the Intelligent Data Analytics report and a state of the field report on Machine Learning and Libraries; and from wildly successful and entertaining IIR experiments Newspaper Navigator and Citizen DJ, to the Collective Wisdom Project, Experimental Access initiative, and Humans in the Loop experiment, the LC Labs team and partners continue to investigate methods, models, and resources in context. Outcomes from this series of events and experiments have demonstrated that subject matter expertise is essential, that we must center approaches on humans and their real needs, and that we should experiment and iterate, while sharing outcomes [13].

Many of these endeavors were themselves informed by the work of the Digital Scholarship Working Group report [14]. Its foundational findings articulate essential needs for item-level metadata and rights assessment to enhance usability of digital collections - approaches that require human expertise and computational methods to address challenges of scale. Fundamentally, that work is iterative and woven together with many threads of collaboration and participation of colleagues.

These recurring recommendations have emerged from the ML-focused initiatives that LC Labs has sponsored:

- Cultivate responsible practices
- Develop appropriate solutions via iteration
- Make available training data for wider use

- Combine machine learning and crowdsourcing
- Sponsor interdisciplinary and interagency collaboration
- Support staff skills development
- Explore infrastructure, policy, and capacity

If and when implemented, these recommendations would benefit not only the Library of Congress but the wider library and archives field – so we continue to share them publicly via labs.loc.gov.

VI. LC LABS PROPOSED FRAMEWORKS

The overarching themes from LC Labs experiments and reports on ML focus on developing a statement of values to guide decision-making around implementing AI in your organization, reinforcing that there is no one-size-fits-all solution when it comes to AI and LAMs and that the operationalizing of any kind of AI system will require more AI expertise across the organization. Building from these very important starting points, further frameworks are needed to help prioritize action and investment.

LC Labs AI and ML experiments have demonstrated and recommended several frameworks, including developing checklists, risk assessments and data archeologies, that encourage reflection and assessment of AI and ML goals against the capabilities and performance of existing models and data.

Practices are evolving across interdisciplinary sectors, accompanied by calls for implementation guidelines. Seeking useful examples that give structure to the community of practice and professional activity has surfaced tools and frameworks that offer practical and aspirational pathways to assess readiness, get started, think critically, and share practice, methods, tools, and insight. Examples include the NDSA Levels of Preservation, NIST AI Risk Framework, Collections as Data, Responsible Operations, and grant funded scholar-practitioner networks[15].

Additionally, methods of documenting datasets and models continue to be refined in interdisciplinary exchange [16] [17] [18]. AI model cards, for example, are lightweight documentation for AI models and are meant to support an informed

decision about the use of a model by a non-expert, inspired by a nutrition fact label--you don't have to be a dietician to know the cautions around the food you're eating. Model cards fit into a larger ecosystem of AI documentation. Documentation of the AI lifecycle helps support understanding, collaboration, sustainability, transparency, reusability. Components of an AI model card include [19]:

- Context: Express the intended user and use of the model, can also include what the model is not intended for
- Ethical Considerations - express the risks possible downstream considerations - environmental and for populations or groups, highlighted for non-technical stakeholders
- Data description: source, size of data and limitations of data (e.g. over 60% males represented)
- Quantitative analysis - overall quality of predictions in use cases

LC Labs developed an experimental framework to help rationalize what users and organizations have to benefit from specific ML or AI-enabled capabilities and to help gain insight into when and how to move toward implementation. The draft framework outlined below is for public comment, review and collaborative improvement.

A: AI Capability Inventory and Assessment

In a spreadsheet, in column one, we are tracking types of ML or AI capabilities that have the potential to transform LAM digital services and categorizing them. The first broad category is divided between front of the house and back of the house services. Some capabilities are processes that are performed behind the curtain and then made public selectively, like an OCR process that helps to generate metadata to enhance search but is not displayed to users. Or, a process that creates one-second audio clips and sorts them by starting note so that they can be remixed and downloaded from an application. We categorized these capabilities as “enabling discovery at scale.” Additional capabilities in this category are generating granular metadata for items, pages, articles, and paragraphs to enhance search services, creating non-English language OCR, handwriting recognition, object classification, name entity identification and linking, and generating bibliographic data, among other tasks.

Another back of the house category is a group of tasks that are processes for local management and preservation of digital content and collections, we are calling these “enhanced collections processing and analysis.” Examples tasks here include using AI to assist in rights assessment of born-digital content, categorizing unstructured born-digital content like web archives and email, assisting in general document sorting for internal business processes, helping with inventory control systems, and creating data that feeds customizable presentations, exhibits and visualizations.

AI and ML can also be used by LAMs in the front of the house to further “augment and extend the user experience” by letting the public directly interface with AI-enabled services like recommending systems, text chat bots, voice recognition and answer systems, voice search services, or visual search tools.

LAM users are also employing AI tools themselves to analyze collections that are made available as data. This front of the house service we are calling “enabling research use” includes a wide range of processes a researcher would perform themselves, including corpus creation, technical methods research and network analysis, among others. The questions that have arisen with this area of capability, are around the surrounding reference services that would be required to support these uses in a responsible way.

In the rows of the spreadsheet we name the specific task or process being considered and in the further columns we capture aspects of the AI process that was examined. These are:

- user story,
- tools or methods tested,
- collection data utilized,
- benefits and risks for users, staff and the organization,
- evidence about the performance of the data or model,
- user or subject feedback and impact, and
- staff or training implications.

To try and summarize the assessment and to get at potential next steps in the exploration of a specific AI task or process, we developed a rough scoring

system rating from one to five, one being the a process that could be closest to implementation.

1. Ready for large-scale implementation with guidelines.
2. Ready for small scale implementation with guidelines.
3. Build on current evidence and do more experimentation.
4. Design an initial experiment and engage stakeholders.
5. Identify and scope potential methods and services.

The evidence gathered through our experimentation to date points to the most potential for small-scale AI implementation in the “enabling discovery at-scale” category, followed by the ‘enabling research use’ group of tasks. These are not surprising (or scientific) results because these are the categories we have done the most experimenting in. A broader set of use cases and feedback from other organizations testing this framework would be required to assess if this is a useful assessment model.

B: A Data Processing Plan Template

One of the key lessons-learned from experiments involving machine learning or artificial intelligence is that characteristics of the data used to train models and how well it aligns to the target data (or data that will be processed with the model) directly indicates the quality of the model’s output. Most models are trained with contemporary born-digital data and don’t perform well when used to process historic or digitized content. The model and all data utilized in a processing task must be documented at each stage so the results can be analyzed--especially before implementing at scale. LC Labs developed a Data Processing Plan template as a starting point for a required set of documentation that technical staff, researchers, or vendors can compile before and after processing, transforming or generating any Library of Congress data. This documentation can help to ensure Library staff have more comprehensive information when deciding how to utilize data generated from experiments. The information will allow for responsible experimentation with Library of Congress data and the opportunity for Library staff to learn about how ML and AI can be effectively implemented.

The elements of the Data Processing Plan template are proposed below. The plan is a work in progress and is also shared with the goal of receiving feedback and community contribution. It is based on recommendations and existing data and algorithmic impact assessment guides. The goal for this plan would be to have staff, partners or vendors fill out an initial draft of the template for review and discussion. A final version of the template would then be compiled after the data had been processed. Each distinct data set that is used in an experiment would require a unique data processing plan.

Section A: General

Describing the goals of the experimental data processing or transformation, the scope of the intended workflow or pipeline, the data delivery format and specifications, and the description of the intended use of the generated data.

Section B: Data Documentation

Describing the data that will be processed, its title, technical composition, including file type, content type, number of items and relative size. The language of the dataset, the time period it covers, the genre and other description information about what intellectual content the dataset contains. Document any copyright, licensing, rights and/or privacy restrictions that could affect the Library's (or the public's) subsequent use of any data processed.

The relevant background context about the composition of the dataset. For example, a dataset may be organized as a single spreadsheet containing metadata about a collection or it may be a series of folders containing images derived from a particular source. The data's provenance, or where it originated, how it was compiled, when, and by whom, and how the dataset is/was technically compiled, for example via an API query or bulk download. This section also covers the preprocessing steps. How has the dataset been classified, cleaned or otherwise prepared for the experiment? How was material selected for inclusion or exclusion in the dataset? Is the data organized according to a schema, content standard or other standards? If yes, which one?

Also document if there are any potential risks to people, communities and organizations if the dataset is used in the experiment and what are the

strategies for risk mitigation. For example, searchable access to individual names and places could expose personal identifying information of private citizens. How will the experiment team mitigate these risks? For example, the team will select data that is over 125 years old to include in the experiment. How will the experiment team address outdated or potentially offensive terms or elements of data that may be harmful if encountered by human users?

Section C: When documenting a dataset for machine learning or artificial intelligence processes, describe the purpose of this dataset with relation to the ML/AL workflow. Explicitly address if it is being used as training, validation or test data. For training data, if the model is pre-trained, describe the data on which it was trained. If the model will be fine-tuned, outline the data involved in this process. If the model is being trained from scratch, outline the plan for creating training data. If creating training data using volunteers or paid participants (e.g. via crowdsourcing), please describe the workflow and incentive structure. If validating training data using volunteers or paid participants (e.g. via crowdsourcing), please describe the workflow and incentive structure. Document any known gaps in the dataset, such as missing instances or forms of representation. Address possible sources of bias in the dataset resulting from these discrepancies. Describe any steps taken to remediate or address gaps or bias in the dataset used in the ML/AL processing or the experiment overall.

Section D: Documentation of the model or models used, including the intended use of the model, the known limitations for the model and its copyright and licensing details. Before processing, document the predicted performance metrics of the model and after each stage of processing and fine tuning, document the actual performance metrics. Establish an audit schedule for how often and how many times the performance metrics will be checked and define a range of successful algorithmic performance. Draw a workflow or pipeline description and diagram, including plans for conducting annotation and validation process, including an overview of supervised or unsupervised machine learning passes.

VII. A COMMUNITY CALL TO ACTION

With accessible and computable collections data, ML and AI methods can be used to enable discovery at scale, enhance collections processing and analysis, enable computational research and augment user experiences. This is the promise that has yet to be realized in LAMs. In sharing these frameworks, we want to continue a community discussion about developing structures that support informed decisions about emerging technologies in LAMs. Developing these initial assessments has been clarifying for prioritizing the next experiments in LC Labs and our hope is that a fuller set of use cases and input could make them useful for more organizations. We invite you to test it out and experiment with different use cases and designs and figure out what works and what does not work in your context. In the coming months, we will aim to come together again to continue iterating on these frameworks together. We are continually inspired by the work of our peers and colleagues and eager for feedback, particularly from the recently formed groups who will evaluate AI and ML practice in LAMs with a specific focus on equity and inclusive justice. The NDSA Levels of Digital Preservation are an excellent model for very actionable and digestible documentation. Extending this concept to AI and ML could help to ensure the informed and responsible adoption of these technologies across the LAM sector.

ACKNOWLEDGMENT

The authors wish to thank the researchers, collaborators, and peer community for participation in LC Labs research, experimentation, events, and outcomes. Furthermore, they wish to express gratitude to Library of Congress colleagues who have enabled exploration of uses and methods through their work to acquire, describe, support, manage, provide access to, and preserve collections.

REFERENCES

- [1] M. Phillips, J. Bailey, A. Goethals, T. Owens. "The NDSA Levels of Digital Preservation : An Explanation and Uses." Library of Congress. 2013.
https://www.digitalpreservation.gov/documents/NDSA_Level_s_Archiving_2013.pdf
- [2] The Data Nutrition Project. <https://datanutrition.org/>
- [3] T. Padilla. "Responsible Operations: Data Science, Machine Learning, and AI in Libraries." Dublin, OH: OCLC Research. 2019. <https://doi.org/10.25333/xk7z-9g97>.
- [4] ExLibris. "Artificial Intelligence in the Library: Advantages, Challenges and Tradition." 2018.
<https://cdn2.hubspot.net/hubfs/2909474/Ex%20Libris%20Artificial%20Intelligence%20White%20Paper.pdf>
- [5] R. Cordell. "Machine Learning + Libraries: A Report on the State of the Field." Library of Congress. 2020.
<https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf>
- [6] E. Lorang, L. Soh, Y. Liu, and C. Pack. Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project. Library of Congress. 2020.
https://labs.loc.gov/static/labs/work/experiments/final-report-revised_june-2020.pdf
- [7] E. Jakeway, L. Algee, L. Allen, M. Ferriter, J. Mears, A. Potter, K. Zwaard. Machine Learning & Libraries Summit Event Summary. Library of Congress. 2019. [ML-Event-Summary-Final-2020-02-13.pdf](https://labs.loc.gov/static/labs/work/reports/ML-Event-Summary-Final-2020-02-13.pdf) (loc.gov)
- [8] W. A. Ingram, S. Johnson. Ensuring Scholarly Access to Government Archives and Records. Virginia Tech University Libraries in partnership with Virginia Tech Center for Humanities and the U.S. National Archives and Records Administration. 2021. <https://vtechworks.lib.vt.edu/handle/10919/108067>
- [9] AEOLIAN Network. <https://www.aeolian-network.net/>
- [10] The Collective Wisdom Project.
<https://collectivewisdomproject.org.uk/>
- [11] Library of Congress Digital Strategy. 2018.
<https://loc.gov/digital-strategy>
- [12] M. Vajcner. "The Importance of Context for Digitized Archival Collections." The Importance of Context for Digitized Archival Collections, vol. 11, no.1, April 2008.
- [13] S. Averkamp, K. Willette, A. Rudersdorf, M. Ferriter. Humans-in-the-Loop Recommendations Report. Library of Congress. 2021.
<https://labs.loc.gov/static/labs/work/reports/LC-Labs-Humans-in-the-Loop-Recommendations-Report-final.pdf>
- [14] A. Potter, G. Harris, K. Zwaard, D. Brunton, S. Garfinkel, J. Hessler, C. Maher, J. Mears, N. Saylor, S. Stillo, C. Townsend. Digital Scholarship at the Library of Congress: User demand, current practices, and options for expanded services. Library of Congress. 2020.
- [15] O. Murphy, E. Villaespesa. The Museums + AI Network. AI: A Museum Planning Toolkit. Goldsmiths, University of London. 2020.
https://themuseumsainetwork.files.wordpress.com/2020/02/20190317_museums-and-ai-toolkit_rl_web.pdf
- [16] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughn, Hanna Wallach, H. Daume III, K. Crawford. "Datasheets for Datasets." v8. 2021 <https://arxiv.org/abs/1803.09010>
- [17] M. Mitchell et al., "Model Cards for Model Reporting." *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–29. 2019.
<https://doi.org/10.1145/3287560.3287596>

- [18] E. M. Bender & B. Friedman. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." *Transactions of the Association for Computational Linguistics* 6: 587–604. 2018. https://doi.org/10.1162/tacl_a_00041.
- [19] B. Kopp. 2022. Framework presented as part of the General Services Administration Responsible AI Community of Practice presentation. Draft documents are only available to community members at the time of submission/publication.

A DIGITAL PRESERVATION WIKIBASE

Katherine Thornton

*Yale University Library
United States
katherine.thornton@yale.edu
du
[0000-0002-4499-0451](tel:0000-0002-4499-0451)*

Kenneth Seals-Nutt

*Yale University Library
United States
kenneth.seals-nutt@yale.edu
[0000-0002-5926-9245](tel:0000-0002-5926-9245)*

Abstract – The Wikidata for Digital Preservation (WikiDP) Wikibase is a project of the Yale University Library's department of digital preservation. The WikiDP Wikibase is an open knowledge base that is publicly available on the web. We outline the relationship of Wikibase to other software of the Wikimedia Foundation, and provide examples of where it is being used. We describe the data models, data sources, and connections between the WikiDP Wikibase and the Wikidata knowledge base. We discuss our decision to use Wikibase for this project which involves transforming a data set related to software into a knowledge base using technologies of the Semantic Web.

Keywords – Wikibase, Wikidata, software metadata, Shape Expressions, Semantic Web

Conference Topics – Community, Exchange, innovation

I. INTRODUCTION

We introduce Wikidata for Digital Preservation (WikiDP), a Wikibase instance related to the domain of computing. This knowledge base contains structured metadata about software, file formats, and configured software environments. Data can be searched via a search bar in the user interface, an application programming interface (API) and a SPARQL endpoint. The knowledge base is publicly available on the web¹.

The fact that this knowledge base is available on the web in a way that is accessible to both humans and machines enables collaboration between large numbers of people around this resource [1]. Making structured data available to machines is part of Tim Berners-Lee's vision for the Semantic Web [2]. Incorporating technologies of the Semantic Web in the field of digital preservation allows us to improve

the interoperability of digital preservation systems with a broad landscape of other systems and data sources. This increases the utility and the value of our data [3]–[7].

Created in 2019, the WikiDP Wikibase contains data from the National Software Reference Library (NSRL) structured to support the description of configured software environments. It contains data about thousands of software titles including information about when they were published, who developed them, what operating systems they are compatible with, and the human languages in which they are available.

We designed the WikiDP Wikibase to support the work of the Emulation as a Service Infrastructure (EaaS) program of work at Yale University Library [8]. The EaaS program of work aims to provide a broad range of configured software environments using a range of software emulators. EaaS users can then interact with legacy software titles which may require outdated operating systems, or other software, that may be inconvenient to access. The EaaS team creates metadata descriptions for configured software environments and stores them in the WikiDP Wikibase.

We outline the steps we took to design and populate this Wikibase. We describe how we mapped the data in the WikiDP Wikibase to Wikidata, and share some example federated queries that allow us to ask questions of the WikiDP Wikibase and Wikidata at the same time.

¹ <https://wikidp.wikibase.cloud>

II. WIKIDATA

Wikidata is a community-curated knowledge base of structured data [9]. Tens of thousands of volunteer editors contribute data to Wikidata relating to a broad range of topics [10]. Data published in Wikidata is available under a Creative Commons Zero (CC0) license. Anyone is free to reuse data from Wikidata for any purpose.

There are multiple options for data reuse from Wikidata. Data in Wikidata can be accessed via the API². Data can also be accessed via SPARQL. SPARQL is a query language for RDF data [11]. RDF is an acronym for Resource Description Framework, a graph-based data model [12]. Wikidata has a SPARQL endpoint that allows anyone with access to the internet to submit queries and get results³. Users can select a format for downloading the results of a query. The available formats are JSON, TSV, CSV, HTML and SVG⁴.

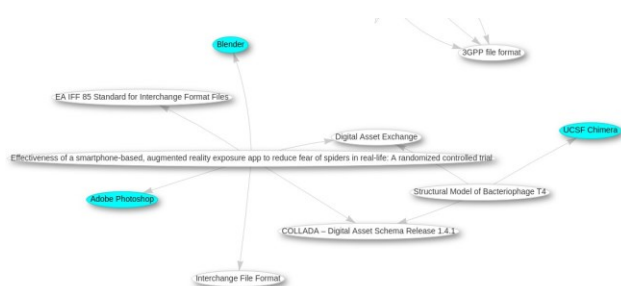


Figure 1 Graph visualization of a SPARQL query illustrating connections in Wikidata between a scholarly publication, a software title (in blue), a file format and a technical specification for that format.

Wikidata contains hundreds of thousands of items related to the domain of computing [13]. It contains data about topics from software titles to software development companies, from file formats to operating systems, even computer hardware. As members of the Wikidata community contribute statements to these items, the set of structured data describing the domain of computing becomes more complete. As members of the Wikidata community use more properties to connect items to one another, we can trace context from a scientific article that describes a project that uses a particular piece

of software to a general set of information about that software title, to a list of the file formats with which that software title can interact, to a technical specification for the file format itself, as seen in Figure 1. One way to get data out of Wikidata is to write SPARQL queries and run them on the Wikidata Query Service SPARQL endpoint [14].

Not only is the data in Wikidata free for anyone to reuse, the software used to create Wikidata is also available for reuse. The Wikimedia Foundation (WMF) has stewarded the MediaWiki software which is used across the many projects of the WMF. The well-known Just solve the problem project⁵ uses Mediawiki software, and the popular Coptr project⁶ uses Semantic Mediawiki, which itself is based on Mediawiki.

III. WIKIBASE

Wikibase is an extension of MediaWiki. MediaWiki is the software used by projects of the Wikimedia Foundation, familiar to most people as the software that powers the different language versions of Wikipedia. Wikibase is the software that enables Wikidata [15]. The German chapter of the Wikimedia Foundation, Wikimedia Deutschland (WMDE), made a docker image available that includes Wikibase in addition to other software [16]. It is available under a free software license allowing anyone to reuse Wikibase to build their own knowledge base.

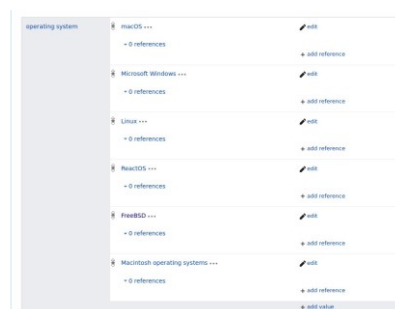


Figure 2 The six operating systems listed on the item for SimulaBeta Q110377565.

Anyone can use Wikibase to design a system tailored to their data⁷. People who want to create their own properties to express relationships different from those available in Wikidata can use Wikibase to do

² https://api.wikimedia.org/wiki/API_reference

³ <https://query.wikidata.org/>

⁴ https://www.mediawiki.org/wiki/Wikidata_Query_Service/User_Manual

⁵ [http://fileformats.archiveteam.org/wiki/Statement of Purpose](http://fileformats.archiveteam.org/wiki/Statement_of_Purpose)

⁶ https://coptr.digipres.org/index.php/Main_Page

⁷ Wikibase documentation available [here](https://www.wikidata.org/wiki/Wikidata:Main_Page).

so⁸. People who want to structure data that isn't appropriate for inclusion in Wikidata can use Wikibase to do so. The fact that Wikibase includes a SPARQL endpoint means that it is possible to run federated queries across a Wikibase and Wikidata itself which allows people to combine the data in their Wikibase with data from Wikidata. An example of a Wikibase related to digital preservation is the ArtBase created by Rhizome [17]. Additional examples of projects using Wikibase can be found in the Wikibase Registry, itself an instance of Wikibase, that provides details on Wikibase usage⁹.

We selected Wikibase for the WikiDP knowledge base because of our familiarity with it from curating data in Wikidata [13], [18], [19]. We wanted to be able to reuse parts of the Wikidata graph in the WikiDP Wikibase. We also wanted to use the Wikibase data model so that we could contribute parts of this data to Wiki- data at some point in the future, if the community de- cides it would be valuable. Wikibase is appropriate for our project because it allowed us to easily make this data available on the web, and it provides a SPARQL endpoint for querying the data.

We decided to create a Wikibase instance for this data because the level of detail required to describe configured software environments involves greater expressivity than is currently possible in Wikidata. We decided that this data model extended too far beyond that of Wikidata, and thus would not be appropriate for inclusion. An example of differences in the level of de- tail is the way software titles and operating systems are described. In Wikidata, multiple operating systems are listed for a software title to indicate those with which the software is known to be compatible. An example of a Wikidata item with multiple compatible operating systems listed is SimulaBeta (Q110377565) as seen in Figure 2.

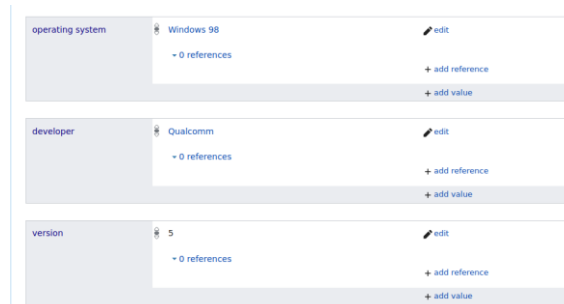


Figure 3 Screenshot of Eudora with Windows 98 listed as compatible operating system in the WikiDP Wikibase.

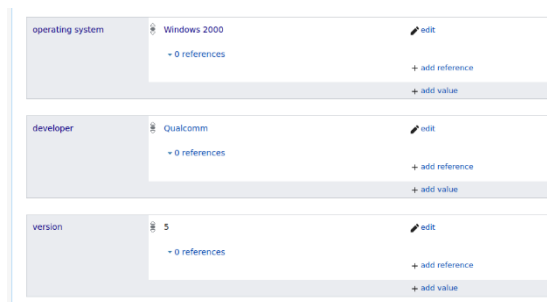


Figure 4 Screenshot of Eudora with Windows 2000 listed as compatible operating system in the WikiDP Wikibase.

In the WikiDP Wikibase we create new items for each software title and operating system combination. This is because we are interested in describing configured environments available in EaaSI and what they contain. Driven by this use case, it is helpful to have each soft- ware title and operating system combination modeled as distinct items. This can be seen in Figure 3, showing the software title Eudora with a single value for the operating system property in the WikiDP Wikibase and Figure 4, a distinct item for the software title Eudora with a different operating system listed. As users of EaaSI use pre-configured environments, it is helpful to have different items for each software tile and operating system combination.

Wikibase has worked very well for this use case. All of the data is public, so there are no issues with the data being available on the web. The process of creating items and properties is familiar to editors of Wikidata. Reusing data from Wikidata allowed us to make useful and meaningful connections between the NSRL data, which was previously siloed, with a general-purpose data set describing computing resources.

⁸ There is also a feature known as 'federated properties' which allows Wikibase users to seamlessly reuse properties from Wikidata as described [here](#).

⁹ https://wikibase-registry.wmflabs.org/wiki/Main_Page

IV. WIKIDATA SUBSETTING

Data published in Wikidata is available under a Creative Commons Zero (CC0) license¹⁰, meaning anyone can reuse any data from Wikidata for any purpose. When creating a new Wikibase, it is sometimes desirable to reuse one or more subsets of Wikidata in the new knowledge base. Creating a subset involves identifying the items and statements about those items you are most interested in and writing a query to extract them from Wikidata. Due to the coverage of items related to the domain of computing, we were able to reuse data from Wikidata to populate our WikiDP Wikibase with structured data. Reusing subsets of Wikidata reduces time needed to source and structure that data. Reusing subsets of Wikidata in Wikibase instances is also convenient because of the fact that they share the same underlying data model.

We used WikidataIntegrator (WDI) to fetch subsets of Wikidata and to populate the WikiDP Wikibase with that data. WDI is a Python library for interacting with data from Wikidata [20]. WDI was created by the Su Lab of Scripps Research Institute and published under an open-source software license via GitHub¹¹. WDI can be used to pull data from Wikidata or to populate Wikidata with data. Similarly, WDI can also be used to get data from or write data to a Wikibase.

We created direct mappings to corresponding Wikidata items for several classes in WikiDP. We reused a subset of Wikidata covering human languages, creating items for each language in WikiDP, and creating a mapping back to Wikidata. We added these items so that we could use them to indicate the languages in which the user interfaces of software titles are available. We also reused the file format subset of Wikidata so that we could reuse them in the Wikibase. Each of the file format items also has a statement containing a mapping back to Wikidata.

Maintaining these mappings is useful for writing federated SPARQL queries. A federated SPARQL query requests information from two or more endpoints in a single query. For example, because of the mappings between file format items in the WikiDP Wikibase and their counterparts in Wikidata, we can ask questions about the file formats in the WikiDP Wikibase and also retrieve data from Wikidata in a single query. Figure 5 shows a SPARQL

query that asks for file formats in the WikiDP Wikibase that have a mapping to Wikidata, and then uses that mapping to find the equivalent file format items in Wikidata that have been used as a value for the property 'main subject' on scholarly article items in Wikidata. The query allows us to see a list of scholarly articles that describe file formats.

Another example of a federated query between the two systems allows us to retrieve user manual links for certain software titles, as seen in Figure 6. The software in the NSRL collection does not include user manuals for the majority of titles. Users of EaaS may need to consult the user manual for the software they are using in a given configured environment. Some of the software titles in Wikidata contain links to a copy of their user manual. By combining data from both knowledge bases we can supply user manual links for many of the NSRL software titles. The software titles are from the NSRL collection in the WikiDP Wikibase, but the user manual links are from Wikidata.



```
1 PREFIX wd: <http://www.wikidata.org/entity/>
2 PREFIX wdt: <http://www.wikidata.org/prop/direct/>
3 PREFIX wdt: <http://wikidp.wiki.opencura.com/prop/direct/>
4 PREFIX wd: <http://wikidp.wiki.opencura.com/entity/>
5 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
6
7 SELECT DISTINCT ?item ?itemLabel ?wikidata ?article ?articleLabel WHERE {
8   ?item wdt:P1 wd:Q1.
9   ?item wdt:P6 ?wikidata.
10  ?item rdfs:label ?itemLabel .
11
12  SERVICE <https://query.wikidata.org/sparql> {
13    ?article wdt:P921 ?wikidata.
14    ?article rdfs:label ?articleLabel.
15  }
16  LIMIT 1000
```

Figure 5 Federated query on the WikiDP Wikibase SPARQL endpoint combining data from Wikidata with data from the WikiDP Wikibase. [Try it!](#)



```
1 PREFIX wd: <http://www.wikidata.org/entity/>
2 PREFIX wdt: <http://wikidp.wiki.opencura.com/prop/direct/>
3 PREFIX wd: <http://wikidp.wiki.opencura.com/entity/>
4 PREFIX wd: <http://www.wikidata.org/prop/direct/>
5
6 SELECT DISTINCT ?family ?familyLabel ?manual WHERE {
7
8   ?item wdt:P1 wd:Q3909.
9   ?item wdt:P23 ?family.
10  ?family wdt:P6 ?wikidata.
11
12  SERVICE <https://query.wikidata.org/sparql> {
13    ?wikidata wdt:P31 wd:Q7397.
14    ?wikidata wdt:P2878 ?manual.
15  }
16
17  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
18
19 }
```

Figure 6 Federated SPARQL query on the WikiDP end- point.

The NSRL collection contains software titles produced by Brøderbund, but does not contain any information about Brøderbund itself. If we consult Wikidata to see what information about Brøderbund

¹⁰ <https://creativecommons.org/choose/zero/>

¹¹ <https://github.com/SuLab/WikidataIntegrator>

has been added, we find a wide range of information. An archival collection related to the company is held by The Strong, as seen in Figure 7.

The Brøderbund item in Wikidata also contains information about a list of Brøderbund products from English Wikipedia as well as the category on English Wikipedia for Brøderbund games, as seen in Figure 8. Additionally, the item provides sitelinks to 21 articles in different language versions of Wikipedia about Brøderbund.



Figure 7 Information about archival collection on the Wikidata item for Brøderbund.

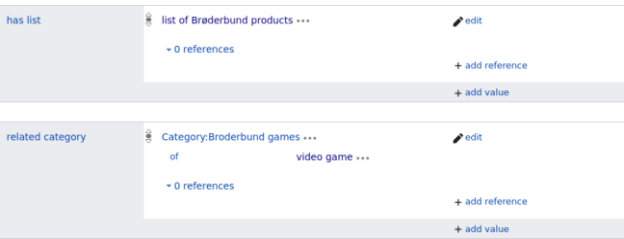


Figure 8 Statements on the Wikidata item for Brøderbund providing information about related information from English Wikipedia.

At the bottom of the page of the Wikidata item there are forty-three external identifiers listed. External identifier properties are used in Wikidata to provide links out to where a resource, in this case Brøderbund, is described by other sites. Wikidata has become a hub for storing and managing identifiers for items [21]. Rather than search for Brøderbund using the search options provided by these forty-three systems, this information is now stored in Wikidata, easing discovery. A sample of some of the external identifiers found on the Wikidata item for Brøderbund can be seen in Figure 9. Several national libraries have information about Brøderbund in their collections. Crunchbase, a database of technology companies has information about the corporate profile of Brøderbund. Justia Patents has information about patents filed or held by Brøderbund. General information about Brøderbund from Wikidata can be combined with information from the NSRL that

describes specific software titles that Brøderbund developed.

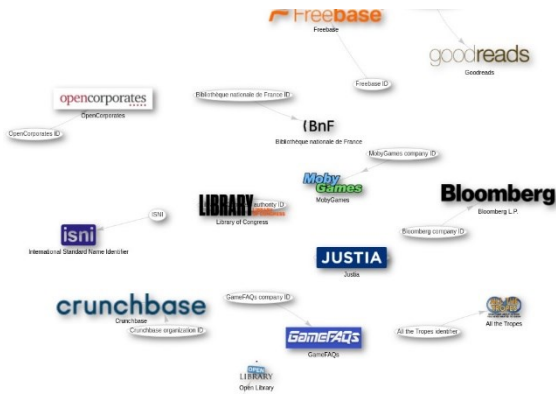


Figure 9 Logos of some organizations that operate repositories for which there is an external identifier related to Brøderbund in Wikidata.

After mapping items and classes from the WikiDP Wikibase to Wikidata we can contextualize information about the NSRL software within the larger sets of information about developers available in Wikidata. Depending on our use cases or our research needs, we can also quickly identify other resources on the web, like the Media Arts Database or the Justia Patents database, if we are interested in specific types of additional information.

Wikidata subsetting is an effective strategy for populating slices of data into a Wikibase. Establishing a property to store the Wikidata mapping for a corresponding item or property in Wikidata itself is useful for anyone who creates a Wikibase and plans to create mappings to Wikidata. As more Wikibases are created in this way we will see the ecosystem of Wikibases diversify in terms of content and data models. This will supplement Wikidata, and provide flexibility for organizations with specific use cases and modeling needs.

V. NATIONAL SOFTWARE REFERENCE LIBRARY

The National Software Reference Library (NSRL) is a collection of software and metadata about software created by the National Institute of Standards and Technology (NIST) of the United States. The purpose of the collection is to support research and investigation related to computer forensics [22].

NIST staff created the NSRL by collecting physical copies of software titles across distribution formats. They described the software using a set of metadata properties such as manufacturer, language,

compatible operating systems, etc. We compared the inventory of software titles in the NSRL with those described in Wiki- data and found only a small area of similarity. NIST donated copies of software titles and associated metadata from the NSRL to Yale University Library as part of the EaaSI program of work. These software titles are being used by EaaSI team members to create a broad range of pre-configured software environments that are available as part of EaaSI.

After reviewing the metadata in the NSRL collection, we designed a set of properties for the WikiDP Wikibase. We considered how we could align certain properties with Wikidata properties. We also considered the needs of the EaaSI system. The final set of properties that we created was influenced by these considerations.

VI. DATA MODELS

We created data models for software titles, software families, file formats, and configured software environments in the WikiDP Wikibase. We use these data models to communicate expectations about data structuring for these different classes of items in the knowledge base.

The EaaSI system provides a catalog of pre-configured software environments for users. These environments are configured by members of the EaaSI team from software available from the NSRL. The class of configured software environment items in the WikiDP Wikibase represents the set of software environments that have been described in the WikiDP Wikibase.

We first created a set of properties inspired by Wikidata. Some examples of these properties are: instance of P1, developer P2, version P3, and file extension P4. Each property also has a mapping to the corresponding Wikidata property as seen in Figure 10. We designed these properties to reflect their equivalent properties in Wikidata so that it would be simple to contribute the data back to Wikidata in the future.

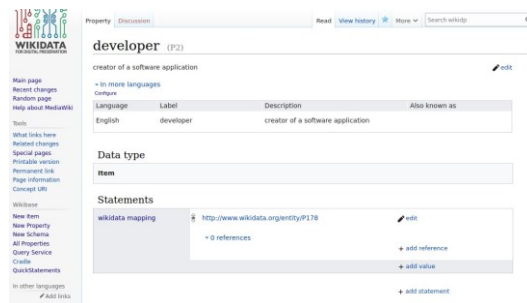


Figure 10 The property for 'developer' in the WikiDP Wikibase with a mapping to the corresponding Wikidata property.

We also created properties to model the NSRL metadata. Some examples are: NSRL manufacturer ID P8, etid P10, etidparent P9, Application ID P11, and NSRL application type P12. These properties reflect the meta- data model of the original NSRL corpus.

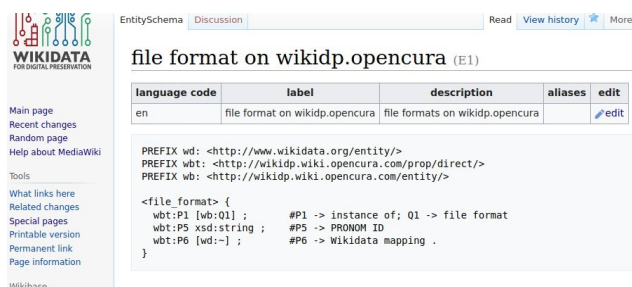
Specific properties we created for EaaSI include: Library of Congress copyright ID P16, base environment, P30, number of disks P38, and Internet Access Required P40. We designed these properties to reflect aspects of how a configured software environment are described.

The Wikibase data model includes references. The references data model makes it possible to reference individual statements. In this way, it is possible to source different statements on the same item to different sources, if needed. It is also possible to provide multiple references per statement. Applying references to each statement ensures that when results are returned via SPARQL, we can quickly identify the source of the information. The reference structure supported by the Wikibase software has been effective for Wikidata [23]. Building on our experiences with the Wikidata system, we work to add references to as many statements as a possible in WikiDP. People who reuse this data will be able to see, per statement, where the data originated and make decisions on whether or not it is relevant for their use case.

VII. SHAPE EXPRESSIONS

Shape Expressions (ShEx) is a formal modeling and validation language for RDF data [24]. ShEx is the schema language used in the Schema namespace (namespace E) of Wikidata and other Wikibase instances [25]. ShEx is the language we use to represent our data models. We write schemas in ShExC, the ShEx compact syntax. We publish our schemas in the E namespace of the WikiDP Wikibase.

The schemas describe the properties and references that are expected for a class of items as well as their expected values. Schemas are a concise way to communicate data models. People interested in contributing to the WikiDP Wikibase, or reusing data from the WikiDP Wikibase, can consult our schemas to gain understanding of our data models.



language code	label	description	aliases	edit
en	file format on wikidp.opencura	file formats on wikidp.opencura		edit

```

PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wbt: <http://wikidp.wiki.opencura.com/prop/direct/>
PREFIX wb: <http://wikidp.wiki.opencura.com/entity/>

<file format> {
  wbt:P1 [wb:Q1] ;          #P1 -> instance of; Q1 -> file format
  wbt:P5 xsd:string ;       #P5 -> PRONOM ID
  wbt:P6 [wd:-] ;          #P6 -> Wikidata mapping .
}

```

Figure 11 ShEx schema for file formats for the WikiDP Wikibase.

Once we have encoded a data model as a schema, we can then use these schemas to validate the entity data in our Wikibase. For example, if someone contributes an item describing a configured software environment, they can then validate that item against our schema for 'configured software environment' to test it for conformance. The ability to test entity data for conformance to a schema is useful in the open contribution model of the WikiDP Wikibase. Contributors from different institutional contexts, language backgrounds, and with different use cases for software emulation may want to describe configured environments in the WikiDP Wikibase. As they are becoming familiar with the system, testing the data they contribute for conformance to a schema provides a way to get automated feedback about where the data are not yet conformant, and what types of changes are needed to bring the data into conformance.

The schema for file formats in the WikiDP Wikibase is seen in Figure 11. This schema has a label and description to provide information about the content. Then there are prefix declarations that provide the namespaces from which the properties are derived. There is one shape in this schema and it is called "file format". The file format shape describes three triple patterns. First file formats should all have a statement that they are instances of (P1) file format (Q1). Then they may have a statement that provides their PUID in the form of a string. Lastly, they should

have a Wikidata mapping (P6) that provides a Wikidata URI for the corresponding file format in Wikidata.

Writing schemas to describe our data models allows us to communicate how our Wikibase connects to Wiki- data itself. This can be useful for people looking to reuse our data, or reuse our data in combination with data from Wikidata. It is also useful for indicating how our Wikibase fits into the network of Wikibases beyond Wikidata.

VIII. ECOSYSTEM OF WIKIBASES

While the breadth of Wikidata content spans many domains, not all data can be accommodated in the knowledge base. The German chapter of the Wikimedia Foundation, Wikimedia Deutschland (WMDE) promotes the concept of an ecosystem of Wikibases [16]. An ecosystem of Wikibases is a network of Wikibase instances each of which supports federated queries with Wikidata itself.

Wikidata was the only Wikibase instance for several years. The Docker image for Wikibase was created by Adam Shoreland and first made available in 2017¹². The Wikimedia Foundation has outlined a vision for how interconnected Wikibases will be created for many different uses¹³. The strategy describes how operators of Wikibase instances and developers of related tooling will work in concert to allow people to query multiple re- sources in order to bring together relevant data.

This ecosystem will encourage groups of people to explore setting up their own Wikibases to serve their own use cases. Some groups may be interested in data that is not appropriate for Wikidata, but can be usefully structured by reusing some properties from Wikidata. Some groups may be interested in creating a set of properties for their data that are not available in Wikidata. Some groups may reuse a subset of Wikidata proper- ties in combination with a set of properties not available in Wikidata. As each Wikibase instance has a SPARQL endpoint that supports federated queries with Wikidata, data can be more easily combined with data from Wiki- data.

Both Wikidata itself, as well as the ecosystem of Wikibases, represent the vision of the Semantic Web.

¹² <https://addshore.com/2017/12/wikibase-docker-images/>

¹³ https://meta.wikimedia.org/wiki/LinkedOpenData/Strategy_2021/Wikibase

"Semantic Web is the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration, and reuse of data across various applications" [26]. The Wikidata knowledge base fulfills the requirements outlined for the Semantic Web in that each resource has a unique identifier, is linked to other resources by properties, and that all of the data is machine actionable.

IX. CONCLUSION

Our work setting up this Wikibase instance and populating it with data has allowed us to interact with the metadata about software titles from the National Software Reference Library (NSRL) in new ways. The data is available on the web and can be searched via the search box in the interface as well as via SPARQL. The data can also now be combined with data from Wikidata.

Creating a Wikibase instance for a specific purpose allows you to establish your own set of properties. This is helpful if you need to represent data models that are not yet represented in Wikidata, or unlikely to be appropriate for Wikidata. For example, there are dozens of properties related to software in Wikidata, but there are many properties important to the data model for configured software environments that are not yet in Wikidata.

The SPARQL endpoint of the WikiDP Wikibase enables federated queries with other SPARQL endpoints.

This SPARQL endpoint allows us to leverage the benefits of combining multiple RDF data sets to ask questions of our data in the context of additional data. Effectively, this means we can ask questions of multiple databases with a single query.

We have contextualized the software described in the NSRL by strategically mapping parts of its data model to Wikidata. This means that we can now ask questions of the NSRL data that previously were impossible. For example, rather than asking about connections between a software developer and software titles that involve querying strings that represent entities, we can now ask questions that extend to the geographic locations of the headquarters locations of those software developers. Or we can ask questions that extend to the scholarly literature that describes research

involving those software titles. Mapping the NSRL data to Wiki- data yields URIs for the entities in the Semantic Web for those organizations. With those URIs we can tap into all of the structured data describing them that has been added to Wikidata.

As more people create Wikibases and populate them with relevant data sets, the ecosystem of repositories of structured data connected to Wikidata will grow and diversify. More people will map previously-siloed data sets to Wikidata, thus creating pathways to the linked open data (LOD) cloud [27]. These connections will unlock access to additional information sources that increase the value of these data sets. In this way, we can transform databases and information systems that were previously islands of data into linked clusters in the LOD cloud.

As an early member of the ecosystem of Wikibases, we expect that many additional Wikibases will be created in future years. As more organizations identify knowledge graphs they would like to have access to on the web that extend beyond the boundaries of Wiki- data, many will decide to manage their own Wikibase instances.

ACKNOWLEDGMENTS

We would like to thank Adam Shoreland for creating wbstack and Rhizome for funding wbstack. Thank you to the Su Lab at Scripps Research Institute for creating WikidataIntegrator and making it available under an open license. Thank you to Andra Waagmeester for maintaining and creating new features for WikidataIntegrator. We would like to thank the Andrew W. Mellon Foundation and the Alfred P. Sloan Foundation for generously supporting the EaaS program of work

REFERENCES

- [1] L.-A. Kaffee, K. M. Endris, and E. Simperl, "When humans and machines collaborate: Cross-lingual label editing in wikidata," in *Proceedings of the 15th International Symposium on Open Collaboration*, 2019, pp. 1–9.
- [2] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 34–43, 2001.
- [3] J. Hunter and S. Choudhury, "A semi-automated digital preservation system based on semantic web services," in *Proceedings of the 4th ACM/IEEE-CS joint conference on digital libraries*, 2004, pp. 269–278.
- [4] Y. Marketakis and Y. Tzitzikas, "Dependency management for digital preservation using semantic web technologies," *International Journal on Digital Libraries*, vol. 10, no. 4, pp. 159–177, 2009.

- [5] D. Tarrant, S. Hitchcock, and L. Carr, "Where the semantic web and web 2.0 meet format risk management: P2 registry," *International Journal of Digital Curation*, vol. 6, no. 1, pp. 165–182, 2011.
- [6] C. Schlieder, "Digital heritage: Semantic challenges of long-term preservation," *Semantic Web*, vol. 1, no. 1-2, pp. 143–147, 2010.
- [7] J. Hunter and S. Choudhury, "Panic: An integrated approach to the preservation of composite digital objects using semantic web services," *International Journal on Digital Libraries*, vol. 6, no. 2, pp. 174–183, 2006.
- [8] E. Cochrane, K. Rechert, S. Anderson, J. Meyerson, and E. Gates, "Towards a universal virtual interactor (uvi) for digital objects," 2019. [Online]. Available: <https://osf.io/xdehm/download>.
- [9] D. Vrandečić, "Wikidata: A new platform for collaborative data collection," in *Proceedings of the 21st International Conference Companion on World Wide Web, ACM*, 2012, pp. 1063–1064.
- [10] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, and D. Vrandečić, "Introducing wikidata to the linked data web," in *The Semantic Web—ISWC 2014*, Springer, 2014, pp. 50–65.
- [11] S. Harris, A. Seaborne, and E. Prud'hommeaux, *Sparql 1.1 query language, w3c recommendation*, 2013. [Online]. Available: <https://www.w3.org/TR/2013/REC-sparql11-query-20130321/>.
- [12] F. Manola and E. Miller, *Resource description framework: Primer*, 2004. [Online]. Available: <https://www.w3.org/TR/2004/REC-rdf-prime-20040210>.
- [13] K. Thornton, E. Cochrane, T. Ledoux, B. Caron, and C. Wilson, "Modeling the domain of digital preservation in wikidata," *iPRES 2017: 14th International Conference on Digital Preservation*, 2017.
- [14] S. Malyshev, M. Krötzsch, L. González, J. Gonsior, and A. Bielefeldt, "Getting the most out of wikidata: Semantic technology usage in wikipedia's knowledge graph," in *International Semantic Web Conference*, Springer, 2018, pp. 376–394.
- [15] L. Zhou, C. Shimizu, P. Hitzler, et al., "The enslaved dataset: A real-world complex ontology alignment benchmark using wikibase," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 3197–3204.
- [16] D. Diefenbach, M. D. Wilde, and S. Alipio, "Wikibase as an infrastructure for knowledge graphs: The eu knowledge graph," in *International Semantic Web Conference*, Springer, 2021, pp. 631–647.
- [17] L. Rossenova, D. Espenschied, and K. de Wild, "Provenance for internet art using the w3c prov data model," in *Proceedings of the 16th International Conference on Digital Preservation*, 2019. [Online]. Available: <https://osf.io/6xd4g/download>.
- [18] K. Thornton and K. Seals-Nutt, "Getting digital preservation data out of wikidata," in *Proceedings of the 16th International Conference on Digital Preservation*, 2019. [Online]. Available: <http://osf.io/guj3p>.
- [19] K. Thornton, K. Seals-Nutt, E. Cochrane, and C. Wilson, *Wikidata for digital preservation*, 2018. [Online]. Available: 10.5281/zenodo.1214319.
- [20] A. Waagmeester, G. Stupp, S. Burgstaller-Muehlbacher, et al., "Wikidata as a knowledge graph for the life sciences," *Elife*, vol. 9, e52614, 2020. [Online]. Available: <https://doi.org/10.7554/ELIFE.52614>.
- [21] J. Neubert, "Wikidata as a linking hub for knowledge organization systems? integrating an authority mapping into wikidata and learning lessons for KOS mappings," in *Proceedings of the 17th European Networked Knowledge Organization Systems Workshop co-located with the 21st International Conference on Theory and Practice of Digital Libraries 2017 (TPDL 2017)*, Thessaloniki, Greece, September 21st, 2017., 2017, pp. 14–25. [Online]. Available: <http://ceur-ws.org/Vol-1937/paper2.pdf>.
- [22] S. Mead, "Unique file identification in the national software reference library," *Digital Investigation*, vol. 3, no. 3, pp. 138–150, 2006.
- [23] A. Piscopo, L.-A. Kaffee, C. Phethean, and E. Simperl, "Provenance information in a collaborative knowledge graph: An evaluation of wikidata external references," in *International semantic web conference*, Springer, 2017, pp. 542–558.
- [24] I. Boneva, J. E. L. Gayo, S. Hym, E. G. Prud'hommeaux, H. R. Solbrig, and S. Staworko, "Validating RDF with shape expressions," *CoRR*, vol. abs/1404.1270, 2014. [Online]. Available: <http://arxiv.org/abs/1404.1270>.
- [25] K. Thornton, H. Solbrig, G. S. Stupp, et al., "Using shape expressions (shex) to share rdf data models and to guide curation with rigorous validation," in *European Semantic Web Conference*, Springer, 2019, pp. 606–620.
- [26] I. Cruz, S. Decker, J. Euzenat, and D. McGuinness, *The emerging semantic web*.
- [27] D. Abián, F. Guerra, J. Martínez-Romanos, and R. Trillo-Lado, "Wikidata and dbpedia: A comparative study," in *Semantic Keyword-based Search on Structured Data Sources*, Springer, 2017, pp. 142–154.

METADATA QUALITY IN DIGITAL LIBRARIES

An Analysis of Survey Response Data

Hannah Tarver

University of North Texas
USA

hannah.tarver@unt.edu

[0000-0003-2344-9268](tel:0000-0003-2344-9268)

Meredith L. Hale

University of Tennessee,
Knoxville

USA

mhale16@utk.edu

[0000-0001-9740-7437](tel:0000-0001-9740-7437)

Rachel White

Colgate University
USA

rwhite@colgate.edu

[0000-0001-6021-3268](tel:0000-0001-6021-3268)

Steven Gentry

University of Michigan
USA

gentrys@umich.edu

[0000-0003-0596-2439](tel:0000-0003-0596-2439)

Madison Chartier

Oklahoma State University
USA

madison.chartier@okstate.edu

[0000-0001-9883-4932](tel:0000-0001-9883-4932)

Rachel J. Wittmann

University of Utah
USA

rachel.wittmann@utah.edu

[0000-0002-7342-3907](tel:0000-0002-7342-3907)

Abstract – The Metadata Quality Benchmarks group of the Digital Library Federation's Metadata Working Group launched a survey in 2019 to gather preliminary data about metadata quality assessment and benchmarking practices used in digital libraries. Data gathered included information about the hosting organization; the size, scope, and technical aspects of the digital repositories; and quality assessment priorities and activities. Survey analysis revealed several trends and correlations, most noticeably in the overall usage frequency of elements, assessment of required versus optional elements, and prioritization of certain quality characteristics and evaluation methods. The authors hope conclusions drawn from this data will spur the broader creation and implementation of metadata benchmarks, enhancing the quality and overall impact of metadata in resource access, discovery, and preservation.

Keywords – Digital Library Federation, Digital libraries, Digital repositories, Metadata assessment, Survey response data

Conference Topics – Community; Resilience

I. INTRODUCTION

Metadata quality assessment in digital libraries is a challenging and complex subject. Successful long-term maintenance of digital resources requires high-quality metadata that documents detailed descriptions, administrative actions, preservation procedures, and other information. However, not only do opinions differ as to what constitutes

metadata assessment, but actual assessment practices may hinge upon available resources, or lack thereof, regardless of the perceived value of quality assessment.

The Digital Library Federation's (DLF) Assessment Interest Group (AIG) was formed in 2014 to address this need for clearer assessment practices in digital librarianship. Comprising numerous working groups that cover a broad field of topics, the AIG develops standards, tools, and practices to help institutions implement fundamental assessment practices. The DLF AIG Metadata Working Group (MWG) specifically targets metadata and strives to "build guidelines, best practices, tools, and workflows around the evaluation and assessment of metadata used by and for digital libraries and repositories" [1]. Since its inception, the MWG has headed several projects to help establish solid metadata practices for digital collections, including a clearinghouse of metadata documentation (e.g., metadata application profiles); a publicly accessible metadata assessment Zotero group [2]; and several metadata quality analysis workshops. More recently, the MWG has sought to gauge the efficacy of such tools and guidelines in promoting metadata quality by examining how institutions assess their metadata.

The Metadata Quality Benchmarks (MQB) group, a MWG sub-group, formed in 2018 to investigate current assessment practices and to suggest general

guidelines for measuring metadata quality. This paper discusses findings from a survey released by the MQB in 2019 that gathered preliminary data about organizations' metadata quality assessment and benchmarking practices. Institutions were asked to provide information about their organization; the size, scope, and technical aspects of their digital repositories; and their quality assessment priorities and activities.

II. LITERATURE REVIEW

Quality assessment is integral to establishing and enforcing good metadata practices for digital library collections. Reference [3] notes that "the quality of metadata can have significant impact in facilitating access, use, and long-term preservation to digital resources" (p. 2). However, metadata assessment may be nebulous work because the rubric underlying assessment depends on its specific context (noted by many, e.g., [4] and [5]). As stated in the DLF AIG Metadata Working Group's Metadata Assessment Framework and Guidance, "[M]etadata quality is subjective. How you define metadata quality will be unique to the core functions and mission of your institution or needs" [6] (see also [7]).

While the definition of quality may be unique to individual institutions' needs and values, the perceived impacts of metadata quality are universally recognized. Depending on the context, data quality may have serious consequences (e.g., medical data [8]). Even without such stakes, studies on the efficacy and deficiencies of metadata quality assessment tools, including a 2019 study of metadata creators' and managers' perspectives, confirm a shared concern of compromised user accessibility due to poor quality metadata [9]. However, quality assessment may be difficult to implement due to limited staffing, time, and resources, leading to manual evaluation or sampling, even in automated processes (e.g., [10]).

To offer some guidance, the MQB created and released a metadata assessment document based on a framework established by [11], which outlined seven quality aspects: accessibility, accuracy, completeness, conformance to expectations, logical consistency and coherence, provenance, and timeliness. This framework was updated to accommodate linked data in 2013 [12] and has been referenced by other authors exploring metadata assessment (e.g., [13], [14], [15], and [3]). For example, reference [16] found that administrative metadata criteria (e.g., provenance) have not been

studied as thoroughly as those concerning information retrieval, despite their value to institutional workflows and administrative audit trails for preservation purposes.

Metadata standards have been developed to promote consistency and shareability, including general schemas like Dublin Core (DC), as well as schemas for specific domains (e.g., Darwin Core [17]) or material types (e.g., VRA Core [18]). Dublin Core, established in 1995 [19], has frequently been employed by digital repositories due to its early creation and wide applicability. A 2002 study evaluated 100 Open Archives Initiative (OAI)-compliant repositories and analyzed the number of DC elements used in records and the frequency of usage [20]. A later 2014 study determined that DC was the most frequently used schema among 77 international repositories [21].

Documentation explaining how a particular institution or project implements an established schema may have varying names, such as guidelines, standards, practices, or Metadata Application Profiles (MAPs). These resources not only provide metadata practitioners with rules and directions when creating metadata, they also govern local metadata production and influence metadata quality. In a recent study of 24 MAPs from academic libraries in the United States, "[a] comparison of elements among the MAPs further revealed insights into the considerations and dilemmas that metadata creators face when attempting to describe disparate and unique materials" (p. 33) [22]. Although MAPs may not explicitly include metadata evaluation practices or procedures, such documentation is often closely connected to quality assessment. Consequently, one graduate-level library science class now combines assignments for these evaluation components to help library students better understand connections between metadata evaluation and MAPs [23].

In addition to local implementations, a number of initiatives have developed guidelines for sharing metadata, including regional or cooperative projects (e.g., [24]) and aggregations. The largest aggregation project MAPs include Europeana (for descriptive records in Europe) [25] and the Digital Public Library of America (DPLA), which uses a national network of hubs for cultural heritage materials in the United States [26]. Aggregations also provide options to test large-scale metadata evaluation. Variations in metadata quality can become more apparent once metadata is in an aggregated environment, housed

amongst collections from many institutions. Some research has focused on specific quality aspects (e.g., a study of completeness in Europeana [27]), specific elements such as *description* [28] or *subject* [29], or general element usage [30] in DPLA.

Major aggregation entities are trying to support quality among contributing organizations, such as the Europeana Publishing Framework [31]. Similarly, DPLA has embraced community-driven approaches to improve metadata quality with the development of task forces, training, and collaborative efforts to develop and review DPLA and network hub guidelines and MAPs [32]. Analysis of metadata aggregations can also highlight quality related to shareability and users' ability to make sense of local records outside their originating context [33], which further impacts the degree to which users may find relevant materials from different sources or understand information in simplified records [28].

Organizations may want to address quality issues in their digital repositories for a number of reasons, including findability, shareability, and long-term preservation. Using criteria in [11] as a basis to assess metadata quality, the MQB launched a survey in the summer of 2019 and invited metadata professionals to answer questions regarding their respective institutions' methods for measuring metadata quality. Initial results outlining aggregated data were released online as a white paper in 2020 [34]. This paper expands on these previous findings with additional discussion of selected survey results.

III. METHODS

The MQB collected data through a Qualtrics survey, which the group promoted across various library-domain listservs. The survey was active from May 23-July 10, 2019. Survey instructions asked that only one metadata professional from each institution provide responses. Only two questions in the survey were mandatory: 1) consent to take part in the survey, and 2) how many repositories are managed by the responding institution.

Overall, 240 respondents consented to take the survey; however, 89 (37%) did not answer any subsequent questions. Of the remaining respondents, 107 (45%) fully completed the survey, while 44 (18%) partially completed the survey, resulting in a total or partial completion rate of 63%.

Survey responses were exported as a .csv file and evaluated manually (using spreadsheets) by the researchers. This analysis compared data across multiple responses to find correlations, which was

not previously done. It did not re-evaluate data that was fully covered in the initial findings. The survey instrument and anonymized raw data are publicly available [35].

IV. RESULTS

Respondents came from a variety of backgrounds. Academic librarians were strongly represented (55%), followed by public librarians (9%) and librarians employed by museums, consortia, and aggregation projects (7%). Two-thirds of the responding institutions managed 1-2 digital repositories. Respondents were asked to describe their institutions' repositories, including whether a particular repository serves as an institutional repository (i.e., resources produced by the organization and/or constituent members), a digital collection (i.e., digitized or born-digital cultural heritage materials), or both (see Table 1). Each repository typically contained 10,000-100,000 records, although sizes ranged from less than 100 items to 10,000,000 or more.

Additionally, the survey asked if the metadata for each repository conforms to a Metadata Application Profile (MAP) and, if so, whether that MAP is an external document (e.g., a consortial MAP to participate in an aggregation, such as DPLA) or a locally-generated MAP. Among all repositories, local MAPs were most frequently used; however, "no MAP" was the most common response overall (see Table 1). For 13 repositories, respondents indicated a MAP is used but did not clarify whether the repository uses a consortial or local MAP.

Table 1
MAP Use by Repository Type

	Institutional Repository	Digital Collection	Both	Total
Local MAP	5	35	17	57
External MAP	1	5	6	12
Unknown MAP	3	7	3	13
No Map	13	28	19	60
Total	22	75	45	142

The survey asked about each repository's schema usage as a possible factor when applying standards or sharing data between institutions. Survey responses showed a marked preference for Dublin Core (DC)-based schemas—including simple DC, qualified DC, and locally-modified or supplemented DC—followed by Metadata Object Description Schema (MODS). Other responses demonstrated a variety of schema usage,

as well as a number of local schemas or combinations of multiple schemas. Similarly, there were no significant patterns in controlled vocabulary usage, although very few repositories (7 of 105) reported using no controlled vocabularies at all.

A. Overall Usage Frequency

The survey's main goal was to discover whether institutions evaluate metadata and, if so, what methods and procedures they implement (e.g., whether they evaluate every element in use or only select elements). The data was collected on a repository (rather than institutional) basis, so numbers do not correlate directly to a parent organization.

To establish a baseline, two questions provided a grid of 26 commonly-used metadata elements. The first question asked respondents to indicate for each individual repository which specific elements on the grid are required, recommended, or optional (with no answer meaning “not used”). The second question asked respondents to indicate whether each of the same elements is evaluated or not evaluated. Although the survey gave respondents the option of supplying local elements not represented on the grids in each question, these free-text responses were so varied that generalizations were difficult to derive from the data. Unless otherwise noted, this paper refers only to the 26 elements listed in the grids when discussing the frequency of element availability or evaluation.

Although the initial evaluation [34] reviewed aggregated data (i.e., total responses per question), the rate of responses across the two grids was not the same. For example, some respondents indicated whether repository elements were evaluated in one grid, but did not specify whether those elements were required, recommended, or optional in the other. When looking at total responses in each grid question, such discrepancies made the actual number of repositories difficult to determine. However, after adjusting the data to account for responses in either grid (assuming that “evaluated” elements are available), adjusted totals accounted for 123 individual repositories (see Table 4). Only the *subject* element is available in all 123 repositories, followed by *creator* and *date* in 122 repositories.

Aside from general responses, the frequency of individual elements being required, recommended, or optional can also be determined based on the

total availability (i.e., how often repositories prefer a particular usage for each element). For example, *subject* is required in 36 repositories, representing 29% of total *subject* usage; *physicalLocation* is also required in 36 repositories, but is available in 23 fewer repositories than *subject*. So required usage for *physicalLocation* is 36% of total frequency (see Table 2). Additionally, an overall difference of only 29 repositories exists between the element used most frequently—*subject* (in 123 repositories)—and the element used least frequently—*table of contents* (in 94 repositories).

Individual responses to the grids also provided more information about total element usage and distributions across repositories. Nearly half of the repositories (58) reported making all 26 possible elements available (see Fig. 1). The total number of available elements in the remaining 65 repositories ranges from 6 to 25, although a majority of those repositories (57) include at least 15 elements. Additionally, these elements can be broken down by level of usage. For example, most repositories tend to require either relatively few elements—17 repositories require 4 elements and 11 repositories require 3 elements—or a substantive number of elements—16 repositories require 7 elements and 14 repositories require 8 elements.

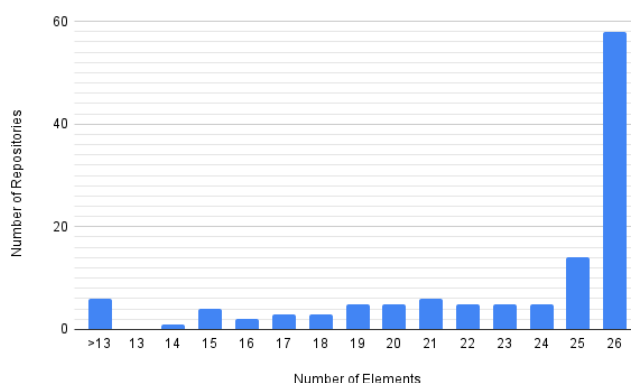


Figure 1 Number of Grid Elements Reported Per Repository

Similarly, most repositories tend to recommend or make optional 5 to 10 elements. No repositories require or recommend all elements, and only 1 repository does not require any elements and makes all 26 elements available optionally

B. Metadata Evaluation

In terms of evaluation, several correlations were identified when the data was broken down by individual responses (see Table 4). Required

elements are more frequently evaluated than not evaluated. However, when elements are only recommended, these elements are evaluated exactly half the time (13 of 26 elements), including one case—*identifier*—that is evaluated and not evaluated in an equal number of repositories (6). Optional elements are not evaluated more often than they are evaluated, except for *creator* (evaluated in 7 repositories and not evaluated in 4), as well as 2

elements that have equal numbers of repositories that do or do not evaluate them: *contributor* (23) and *title* (1). *Creator* is also the only element that is most often evaluated regardless of usage.

Only 12 repositories lacked any kind of metadata evaluation. One respondent clarified: all metadata records are evaluated prior to ingest; presumably, no evaluation occurs post-ingest.

Table 2
Total Element Frequency with Percentage of Element Use, Ordered by Frequency of Requirement

Element Name	Element Usage Frequencies		Percentage of Repositories in Which the Element is:				
	Required	Total	Required	Recommended	Optional	Unspecified	Not Used
<i>Title</i>	115	121	0.95	0.02	0.02	0.00	0.02
<i>Identifier</i>	93	118	0.79	0.11	0.08	0.03	0.04
<i>Rights</i>	75	119	0.63	0.22	0.13	0.02	0.03
<i>Type</i>	75	116	0.65	0.18	0.13	0.04	0.06
<i>Collection Title</i>	74	113	0.65	0.12	0.19	0.04	0.09
<i>Format</i>	55	117	0.47	0.22	0.26	0.05	0.05
<i>Date</i>	54	122	0.44	0.44	0.10	0.02	0.01
<i>Creator</i>	46	122	0.38	0.52	0.09	0.01	0.01
<i>Subject</i>	36	123	0.29	0.46	0.23	0.02	0.00
<i>physicalLocation</i>	36	100	0.36	0.18	0.39	0.07	0.23
<i>Description</i>	33	118	0.28	0.40	0.30	0.03	0.04
<i>Language</i>	30	114	0.26	0.32	0.35	0.06	0.08
<i>Extent</i>	27	115	0.23	0.36	0.37	0.04	0.07
<i>Publisher</i>	26	113	0.23	0.29	0.46	0.02	0.09
<i>Genre</i>	26	107	0.24	0.39	0.29	0.07	0.15
<i>isPartof</i>	17	103	0.17	0.28	0.51	0.04	0.19
<i>Contributor</i>	16	121	0.13	0.43	0.41	0.02	0.02
<i>Source</i>	16	107	0.15	0.32	0.50	0.04	0.15
<i>Abstract</i>	15	110	0.14	0.35	0.50	0.02	0.12
<i>Spatial</i>	11	111	0.10	0.37	0.46	0.07	0.11
<i>Coverage</i>	9	115	0.08	0.39	0.49	0.04	0.07
<i>Digitization Specs</i>	8	99	0.08	0.19	0.58	0.15	0.24
<i>Transcription</i>	8	98	0.08	0.28	0.59	0.05	0.26
<i>Relation</i>	4	103	0.04	0.23	0.68	0.05	0.19
<i>Table of Contents</i>	4	94	0.04	0.10	0.76	0.11	0.31
<i>Alternative Title</i>	1	115	0.01	0.14	0.81	0.04	0.07

For 3 repositories, respondents did not answer either grid question, but listed local recommended or optional elements, none of which are evaluated. This means, of 126 total repositories, 88% engage in evaluation of at least 1 element.

Although organizations are doing at least some quality control or assessment, they most commonly evaluate only a few of their total elements. Survey respondents for 5 repositories indicated they evaluate only 1 element (2 repositories) or 2 elements (3 repositories); most respondents (representing 52% of reported repositories) are evaluating 5 to 14 elements (64 repositories). However, these numbers are not reflective of “how much” available metadata is evaluated. Responses for the 123 repositories using grid elements indicated that, in 66 repositories—just over 50% of total repositories—less than half of the available elements are evaluated (see Fig. 2). Comparatively, the evaluation rate for 18 repositories is roughly half of their available elements (marked with a shaded bar in Fig. 2), and only 39 repositories reported evaluation of at least 60% of available elements. These numbers change when broken down by usage, since required elements are more likely to be evaluated than non-required elements. For example, 8 repositories evaluate all of their elements, and 57 repositories evaluate at least half of their total elements. However, 43 repositories evaluate all required elements, and 88 repositories (i.e., 71.5%) evaluate at least half of their required elements. Only 3 repositories conduct evaluation but assess none of the required elements.

One interesting finding is that repositories using the fewest grid elements tend to have the lowest percentage of evaluated elements. Original hypotheses anticipated a high assessment rate, as these repositories have less metadata to review as compared to repositories using more elements. However, there were 6 repositories making fewer than 13 of the 26 possible elements available; among those, half indicate that no elements are evaluated (see Table 3). Evaluation for repositories using 14-26 varied without any clear trends, including only 8 repositories that evaluate all available elements.

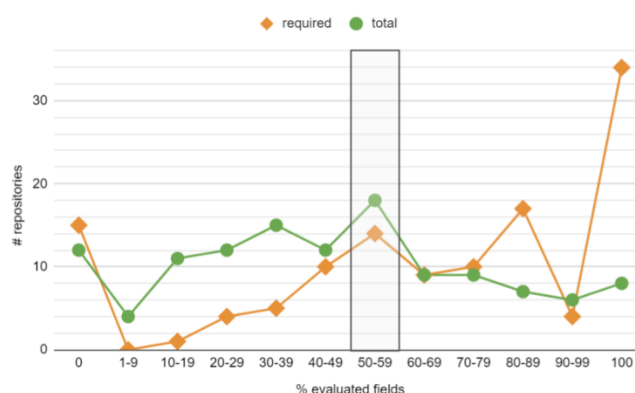


Figure 2 Percentage of Total (circles) and Required (diamonds) Elements Evaluated, by Repository

Significant overlap exists among the top ten respective elements that are most often available, required, or evaluated in repositories (see Fig. 3). Elements that are in the top ten for any of these facets account for half of the 26 grid elements. Seven DC elements—*creator*, *date*, *identifier*, *rights*, *subject*, *title*, and *type*—are in the top ten for all three areas. This degree of overlap suggests a number of informal shared expectations regarding preferred descriptive elements. There are also 3 elements that are frequently not evaluated, even though they are most often available or required: *description*, *format*, and *physicalLocation*. Although *description* may be more difficult to evaluate—as it is generally a complex, free-text element—why *format* or *physicalLocation* are not evaluated is less clear. *Format* is often managed by a controlled vocabulary, and *physicalLocation* may be important for managing materials, tracking digitization projects, and scheduling preservation measures. This overlap provides a potential starting point for identifying cross-organization expectations.

Table 3
Evaluation Rate for Repositories Using Fewer than 13 Grid Elements

Repo- sitory	Number of Elements				Elements Evaluated	
	Requ- ired	Recom- mended	Opt- ional	Total	Num- ber	Per- cent
A	1	1	4	6	1	0.167
B	3	4	0	7	0	0
C	1	0	8	9	0	0
D	7	3	0	10	6	0.600
E	1	4	6	11	0	0
F	7	4	1	12	5	0.417

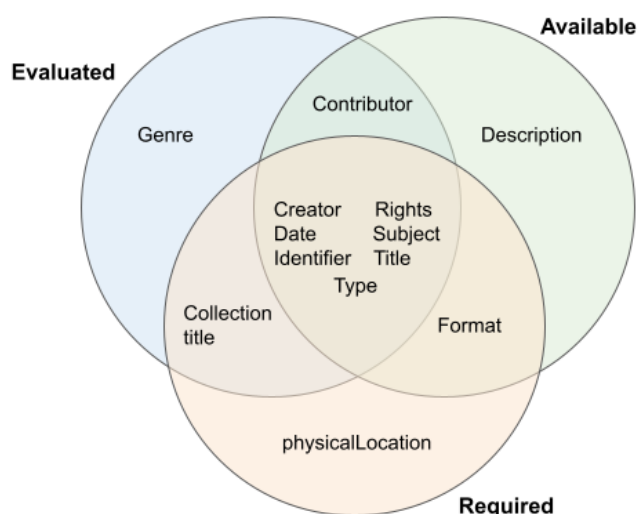


Figure 3 Venn Diagram of Overlap for the Most-Often Required, Evaluated, and Available Elements.

C. Characteristics of Metadata Quality

The survey's final questions asked about evaluation priorities and assessment methods for the seven quality aspects established by [11]: accuracy, accessibility, completeness, conformance to expectations, logical consistency and coherence, provenance, and timeliness. One question asked respondents to express the importance of these aspects at their institutions by ranking each from most (1) to least (7) important (see Table 5). Respondents often agreed on which rankings to give to individual aspects; however, among 87 individual responses, there were 73 unique combinations of rankings.

Table 4
Element Evaluation by Use (Required, Recommended, Optional, Unspecified), Ordered by Frequency of Requirement

Element Name	Required			Recommended			Optional			Unspecified	
	Eval-uated	Not Eval-uated	No Answer	Eval-uated	Not Eval-uated	No Answer	Eval-uated	Not Eval-uated	No Answer	Eval-uated	Not Eval-uated
<i>Title</i>	83	30	2	2	1	0	1	1	1	0	0
<i>Identifier</i>	57	30	6	6	6	1	4	5	0	1	2
<i>Rights</i>	58	13	4	16	9	1	3	10	3	0	2
<i>Type</i>	51	15	9	9	10	2	3	12	0	2	3
<i>Collection Title</i>	52	18	4	9	2	2	3	17	2	2	2
<i>Format</i>	35	14	6	7	14	5	8	22	0	2	4
<i>Date</i>	46	7	1	33	16	5	5	7	0	2	0
<i>Creator</i>	38	6	2	41	19	4	7	4	0	1	0
<i>physicalLocation</i>	20	14	2	8	5	5	8	27	4	2	5
<i>Subject</i>	32	1	3	42	11	3	10	16	2	1	1
<i>Description</i>	22	9	2	16	29	2	10	22	3	1	2
<i>Language</i>	19	10	1	14	18	5	8	25	7	2	5
<i>Extent</i>	13	12	2	15	20	6	6	31	5	3	2
<i>Genre</i>	19	3	4	27	11	4	7	20	4	4	4
<i>Publisher</i>	17	6	3	14	17	2	10	38	4	2	0
<i>isPartOf</i>	13	3	1	10	15	4	7	36	10	0	4
<i>Contributor</i>	15	0	1	32	17	3	23	23	4	2	1
<i>Source</i>	13	0	3	13	14	7	9	36	8	1	3
<i>Abstract</i>	8	7	0	19	18	1	12	30	13	1	1
<i>Spatial</i>	6	3	2	28	9	4	10	37	4	4	4
<i>Coverage</i>	6	1	2	28	13	4	12	39	5	4	1
<i>Digitization Specs</i>	4	3	1	7	9	3	4	39	14	4	11
<i>Transcription</i>	4	2	2	8	15	4	11	36	11	0	5
<i>Relation</i>	3	0	1	10	12	2	12	46	12	1	4
<i>Table of Contents</i>	3	1	0	3	6	0	6	47	18	2	8
<i>Alternative Title</i>	1	0	0	8	5	3	20	57	16	2	3

Table 5
Number and Percentage of Responses Ranking Quality Aspects,
Grouped by High, Medium, and Low

	Ranked 1 / 2		Ranked 3 / 4 / 5		Ranked 6 / 7	
	#	%	#	%	#	%
Accuracy	57	0.655	29	0.333	1	0.011
Completeness	52	0.598	29	0.333	6	0.069
Consistency	29	0.333	51	0.586	7	0.080
Conformance to Expectations	19	0.218	60	0.690	8	0.092
Accessibility	11	0.126	58	0.667	18	0.207
Provenance	4	0.046	21	0.241	62	0.713
Timeliness	2	0.023	13	0.149	72	0.828

This reflects that organizations tend to have different priorities and sometimes significant divergences in terms of which aspects are most important.

There were 3 quality aspects not ranked last by any institution: accuracy, completeness, and consistency. Additionally, those qualities were ranked 6th by only 7 respondents (for consistency), 6 respondents (for completeness), and 1 respondent (for accuracy). On the other end of the spectrum, neither provenance nor timeliness were ranked 1st by any organization. As these elements are largely representative of administrative metadata, crucial only to the repository and not the end user, the revelation that these aspects are not regularly evaluated is perhaps not surprising. It is also consistent with the findings from the literature. Two of the aspects ranked most important (consistency and accuracy) were also listed in subsequent questions as the aspects that organizations would most like to evaluate but currently cannot.

V. DISCUSSION

This survey gathered initial information regarding institutions' digital repositories and metadata evaluation practices. Despite the broad spectrum of repository types and metadata implementations, review of respondent-level data revealed insightful trends and correlations.

According to respondents, around 42% of repositories do not rely on any form of Metadata Application Profile (MAP). In the white paper documenting initial findings of the survey [34], the hypothesis was that the relatively low usage of MAPs (or the high number of repositories not using MAPs) may be due to terminology.

Some institutions may have guidance specifications or documents that effectively serve the same purpose as MAPs, even if they are not referred to as such. This consideration is relevant since a number of respondents said they use standards and documentation as a method of quality control. More investigation may be helpful in this area, both to determine the types of existing institutional documentation and to better understand how these are used as reference materials for validation or quality control.

Almost all repositories (93%) make at least 15 elements available in their metadata records. Of the 123 repositories that documented element usage in the grid questions, almost half (47%) reported using all 26 listed elements. This response rate suggests a fairly robust expectation for describing materials and content. Most repositories also tend to require at least 3 elements, with the bulk of repositories requiring either 3-4 elements (28 repositories) or 7-8 elements (30 repositories). If required elements can be understood to represent a "minimal-level" record, this average number of required elements is relatively sparse, but is still more than the 2 elements required by DPLA (which may serve as a baseline for institutions considering external aggregation). There was no significant correlation between the repository type and the number of available or required elements, so variations seem to be based on local preferences rather than the expected level of description for cultural heritage versus institutional materials.

As more than half of the repositories use a DC-based schema (basic or qualified), the top ten most-frequently-available elements (required, recommended, or optional) are all part of the DC set: *subject*, *creator*, *date*, *contributor*, *title*, *rights*, *description*, *identifier*, *format*, and *type*. The other 5 DC elements have varying levels of availability, from *coverage* (ranking 12th) to *relation*, which is 22nd out of 26 elements when ordered by frequency. In terms of total frequency, all but one of the elements in the grid are available in at least 80% of the repositories; the least-frequently available element (*table of contents*) is still available in 76% of the repositories. These frequencies suggest a high level of overlap regarding element usage, even among repositories of different types or those using different schemas. This may have implications for generalizing metadata quality benchmarks or recommendations for usage and value formatting. In fact, overlapping element usage may

be more useful than relying on schemas, considering the wide array of schema applications, including a variety of combinations and qualifications of the DC elements.

The survey data also establishes that most institutions evaluate at least some of the metadata in their digital library systems, although the extent (and some respondents' opinions on this topic) varied dramatically. In roughly 52% of repositories, less than half of the available elements are evaluated. Repository elements are more likely to be evaluated if they are required than if they are recommended or optional, although most repositories require relatively few elements. In fact, 17 of the 26 elements are required in only 30% or fewer of the repositories that make them available. These required elements may reflect a priority for findability and interface functionality (e.g., a *title* value that displays in search results for users).

Respondents also ranked completeness as one of the most important quality aspects at their institutions. This ranking could be related to a tendency to check required elements, as the use of required elements is sometimes considered a reasonable metric for a minimally-complete record, or may simply be required by the repository system to save the metadata. Limited personnel, unfamiliarity with evaluation tools, and other resources may also be factors, particularly given that a number of institutions reported a reliance on manual checks as their primary or only method of evaluation.

Although the data showed significant differences in quality aspect rankings in individual responses, some definite priorities can be generalized across institutions. For example, a marked preference for accuracy, completeness, and consistency is evident, as is a relative lack of interest in timeliness or provenance. Whether lower rankings represent a gap or need is unclear. Perhaps a quality aspect is not prioritized for evaluation because it is too difficult to assess or define internally. Maybe it requires less intentional review (e.g., provenance or machine-readability may be automatically generated, validated, or recorded and require no intervention). Certain quality aspects may also have less direct effect on user needs, which often assume a high priority in element selection (e.g., completeness could affect browsing functions across a collection or system versus timeliness, which may be more important on an individual record

level). These issues surrounding priorities and resource allocation tend to be complicated but may benefit from generalized benchmarks or guidance.

VI. CONCLUSION

This survey and its analyzed results provide insights into the broad range of current metadata implementation and evaluation practices. In the future, the authors hope to expand upon this research through efforts to identify and propose generalized metadata benchmarks, so as to provide a common standard to which all institutions may refer. One individual shared: “[O]ur metadata quality analysis ... [is] very ad hoc, irregular, and targeted to particular problems we experience... [W]e don't really make sure we're adhering to very many external guidelines” (p. 23) [34]. A common set of benchmarks may enable consistency when it comes to metadata quality, which will further enhance information institutions' ability to preserve digital repository metadata and to make this metadata, along with its affiliated resources, more shareable and findable for a variety of users. Based on the survey results, the MQB is currently drafting additional resources to support further benchmarking work, with plans to follow-up with respondent organizations and the digital library community.

As standards ensure successful, consistent, and shareable metadata upon creation, similar standards for quality assessment, established by a dedicated community of professionals, will prove influential in enhancing materials' usability, accessibility, and long-term preservation on digital platforms. Metadata evaluation practices are in need of standardization efforts similar to the community-based efforts to streamline institutional metadata creation through MAPs. The authors hope the results presented in this survey analysis will likewise inspire fellow metadata practitioners to come together to review and develop quality assessment procedures and make them an active part of their metadata workflows. Additionally, the authors hope efforts like the survey encourage more widespread community initiatives to identify and establish potential benchmarks that may prove useful to institutions implementing fundamental to advanced metadata.

REFERENCES

- [1] DLF Metadata Assessment Working Group. (2021) About. The DLF Metadata Assessment Working Group website. [Online]. Available: <https://dlfmetadataassessment.github.io/about>
- [2] DLF Metadata Assessment Working Group. (2022) Zotero Group: Metadata Assessment. [Online]. Available: https://www.zotero.org/groups/488224/metadata_assessment
- [3] O. L. Zavalina, P. Kizhakkethil, D. G. Alemneh, M. E. Phillips, and H. Tarver, "Building a Framework of Metadata Change to Support Knowledge Management," *Journal of Information and Knowledge Management*, 1st ed., vol. 14, pp. 16, 2015. Available: <https://digital.library.unt.edu/ark:/67531/metadc505014/>
- [4] K. Snow, "Defining, Assessing, and Rethinking Quality Cataloging," *Cataloging & Classification Quarterly*, vol. 55, no. 7-8, pp. 438-455, 2017. Available: <https://doi.org/10.1080/01639374.2017.1350774>
- [5] C. A. Reeves and D. A. Bednar, "Defining Quality: Alternatives and Implications," *The Academy of Management Review*, vol. 19, no. 3, pp. 419-445, 1994. Available: <https://doi.org/10.2307/258934>
- [6] DLF Metadata Assessment Working Group. (2017) Metadata Assessment Framework and Guidance. [Online]. Available: <https://dlfmetadataassessment.github.io/framework>
- [7] G. Gueguen, C. Harper, and C. Stanton. "Perspectives on Data and Quality (Slides)." DPLAFest 2016, Washington, DC, USA, (April 15, 2016). Available: tinyurl.com/2p9v97ud
- [8] L. M. Schriml et al., "COVID-19 Pandemic Reveals the Peril of Ignoring Metadata Standards," *Sci Data*, vol. 7, no. 188, 2020. Available: <https://doi.org/10.1038/s41597-020-0524-5>
- [9] N. T. Fox, H. Tarver, and M. E. Phillips (2019) Identifying Gaps in Tools and Interfaces for Assessing Metadata Quality [White paper]. [Online]. Available: <https://digital.library.unt.edu/ark:/67531/metadc1453742/>
- [10] M. Goovaerts and D. Leinders, "Metadata Quality Evaluation of a Repository Based on a Sample Technique," in *Metadata and Semantics Research. MTSR 2012*, J. M. Doderio, M. Palomo-Duarte, and P. Karampiperis, Eds. Communications in Computer and Information Science, Vol. 343, Berlin: Springer, 2012. Available: https://doi.org/10.1007/978-3-642-35233-1_19
- [11] T. Bruce and D. Hillmann, "The Continuum of Metadata Quality: Expressing, Exploiting," in *Metadata in Practice*. D. Hillmann and E. Westbrook, Eds. Chicago: ALA Editions, 2004, pp. 238-256. Available: <https://ecommons.cornell.edu/handle/1813/7895>
- [12] T. Bruce and D. Hillmann (2013, Jan.) Metadata Quality in a Linked Data Context. [Online]. Available: <https://blog.law.cornell.edu/voxpath/2013/01/24/metadata-quality-in-a-linked-data-context/>
- [13] X. Ochoa and E. Duval, "Automatic Evaluation of Metadata Quality in Digital Repositories," *International Journal on Digital Libraries*, Vol. 10, pp. 67-91, 2009. Available: <https://link.springer.com/article/10.1007/s00799-009-0054-4>
- [14] J.-R. Park. "Metadata Quality in Digital Repositories: a Survey of the Current State of the Art," *Cataloging & Classification Quarterly*, vol. 47, no. 3-4, pp. 213-228, 2009. Available: <https://doi.org/10.1080/01639370902737240>
- [15] J. Park, Y. Tosaka, S. Maszaros, and C. Lu, "From Metadata Creation to Metadata Quality Control: Continuing Education Needs Among Cataloging and Metadata Professionals," *Journal of Education for Library & Information Science*, vol. 51, no. 3, pp. 158-176, 2010.
- [16] H. Ulrich, et al., "Understanding the Nature of Metadata: Systematic Review," *Journal of Medical Internet Research*, vol. 24, no. 1, 2022. Available: <https://doi.org/10.2196/25440>
- [17] (2012) Dublin Core™ Metadata Element Set, Version 1.1: Reference Description website. [Online]. Available: <https://www.dublincore.org/specifications/dublin-core/dces/>
- [18] (2015) Darwin Core website. [Online]. Available: <https://dwc.tdwg.org/>
- [19] (2022) VRA Core Official Website. [Online]. Available: <https://www.loc.gov/standards/vracore/>
- [20] J. Ward, A Quantitative Analysis of Dublin Core Metadata Element Set (DCMES) Usage in Data Providers Registered with the Open Archives Initiative (OAI) [Master's Thesis], Chapel Hill, NC: University of North Carolina at Chapel Hill, 2002. Available: https://cdr.lib.unc.edu/concern/masters_papers/tq57nv59w
- [21] M. Curado Malta and A. A. Baptista, "A Panoramic View on Metadata Application Profiles of the Last Decade," *International Journal of Metadata, Semantics and Ontologies*, vol. 9, no. 1, 2014. Available: <https://doi.org/10.1504/IJMSO.2014.059124>
- [22] A. M. Green, "Metadata Application Profiles in U. S. Academic Libraries: A Document Analysis," *Journal of Library Metadata*, vol. 22, 2022. Available: <https://doi.org/10.1080/19386389.2022.2030172>
- [23] O. Zavalina, "Integrated Learning of Metadata Quality Evaluation and Metadata Application Profile Development in a Graduate Metadata Course," *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 2017. Available: <https://dcpapers.dublincore.org/pubs/article/view/3856.html>
- [24] C. Cronin, "Metadata Provision and Standards Development at the Collaborative Digitization Program (CDP): A History," *First Monday*, Vol. 13, No. 5, 2008. Available: <https://firstmonday.org/ojs/index.php/fm/article/view/2085/1957>
- [25] (2017) Europeana Data Model - Mapping Guidelines v2.4. [Online]. Available: <https://pro.europeana.eu/page/edm-documentation>
- [26] (2017) Metadata Application Profile. Digital Public Library of America website. [Online]. Available: <https://pro.dp.la/hubs/metadata-application-profile>
- [27] P. Király and M. Büchler, "Measuring Completeness as Metadata Quality Metric in Europeana," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 2711-2720. Available: <https://doi.org/10.1109/BigData.2018.8622487>
- [28] H. Tarver, O. L. Zavalina, and M. E. Phillips, "An Exploratory Study of the Description Field in the Digital Public Library of America," *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 2016, pp. 34-44. Available: <https://digital.library.unt.edu/ark:/67531/metadc910346/>
- [29] M. E. Phillips and H. Tarver, "Investigating the Use of Metadata Record Graphs to Analyze Subject Headings in the Digital Public Library of America," *The Electronic Library*, vol. 39, no. 3, 2021. Available: <https://doi.org/10.1108/EL-11-2020-0317>

- [30] C. A. Harper, "Metadata Analytics, Visualization, and Optimization: Experiments in statistical analysis of the Digital Public Library of America (DPLA)," Code4Lib, no. 33, 2016. Available: <https://journal.code4lib.org/articles/11752>
- [31] B. Daley, H. Scholz, and V. Charles, "Developing a Metadata Standard for Digital Culture: the Story of the Europeana Publishing Framework." [Online]. Available: <https://pro.europeana.eu/post/developing-a-metadata-standard-for-digital-culture-the-story-of-the-europeana-publishing-framework>
- [32] G. Guegen, "Metadata Quality at Scale: Metadata Quality Control at the Digital Public Library of America," Journal of Digital Media Management, vol. 7, no. 2, 2019, pp. 115-126. Available: <https://www.ingentaconnect.com/content/hsp/jdmm/2019/00000007/00000002/art00003>
- [33] S. L. Shreeves, et al., "Is Quality Metadata Shareable Metadata? The Implications of Local Metadata Practices for Federated Collections," in Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, H. A. Thompson, Ed. Chicago: Association of College and Research Libraries, 2005, pp. 223-237. Available: <https://www.ideals.illinois.edu/handle/2142/145>
- [34] S. Gentry, et al. (2020, May) Survey of Benchmarks in Metadata Quality: Initial Findings [White Paper]. [Online]. Available: <http://dlfmetadataassessment.github.io/assets/2020-dlf-mawg-mqb-white-paper.pdf>
- [35] DLF Metadata Assessment Working Group. (2020) Projects: Metadata Quality Benchmarks. The DLF Metadata Assessment Working Group website. [Online]. Available: <https://dlfmetadataassessment.github.io/benchmarks>

MAKING RISK MODELING ACCESSIBLE WITH DIAGRAM

Additional development of an online digital preservation tool to meet Web Content Accessibility Guidelines 2.1

David Underdown

*The National Archives
United Kingdom
david.underdown@nationalarc
hives.gov.uk
[0000-0002-8123-4655](tel:0000-0002-8123-4655)*

Alexandra Leigh

*City, University of London /
The National Archives
United Kingdom
alexandra.leigh@city.ac.uk*

Pauline Descheemaeker

*The National Archives
United Kingdom
pauline.descheemaeker@national
archives.gov.uk*

Abstract – this paper will examine the work undertaken on DiAGRAM (Digital Archiving Graphical Risk Assessment Model) to ensure it is fully compliant with the Web Content Accessibility Guidelines 2.1 (WCAG2.1) and the United Kingdom's Public Sector Bodies Access Regulations (PSBAR). This work also supports The National Archives' strategic goal of becoming the Inclusive Archive.

The initial development of DiAGRAM aimed to bring the power and insight of Bayesian Networks to the community of digital archivists. The prototype tool successfully demonstrated this, however, the relatively short project timeframe and adoption of a rapid prototyping approach (with R/Shiny) imposed constraints on accessibility. We offer some lessons learned from the project in how we could have approached this more effectively.

With The National Archives' own successful use of the prototype in support of an investment business case for improving our digital archive's resilience the tool's utility was sufficiently clear for a further phase of work. The contract was awarded to data analytics firm Jumping Rivers, who had also previously undertaken initial usability and accessibility improvements. A full external accessibility review was undertaken by TetraLogical.

The first part of this new project phase determined that the most appropriate way to proceed was to re-architect the DiAGRAM tool to separate the web front-end from the underlying model and connect the two via a new set of API endpoints. Opportunity was also taken to review the "Advanced customisation" modeling options to improve usability.

Redevelopment is now substantially complete although the formal retest to confirm full compliance with PSBAR took place in March 2022, the ongoing dialog with TetraLogical throughout development meant no new major issues were found, though some smaller issues did remain and are now being addressed.

Keywords – web accessibility, risk modeling,
Conference Topics – Community; Resilience.

I. INTRODUCTION

A. Background and Prior Project Work

DiAGRAM (Digital Archiving Graphical Risk Assessment Model) was an output of the project "Safeguarding the nation's digital memory" [1]. It is an Integrated Decision Support System (IDSS) designed to give digital archivists guided access to the underlying Bayesian Network (BN) representing the various risk factors (and the interactions between them) relevant to the preservation of digital materials. In order to undertake simple modeling, archivists answer a series of questions to fit DiAGRAM to the current situation of their archive by setting the input nodes of the BN. These reflect factors such as the balance between different types of digital material currently held by the archive and the broad classes of storage media used. DiAGRAM then returns scores for Renderability and Intellectual Control which reflect the probabilities that "The object is a sufficiently useful representation of the

original file” and that you “Hav[e] full knowledge of the material content, provenance and conditions of use” respectively [2].

Further discussion of the initial development of the model, and in particular of the structured elicitation process used to obtain rigorous probabilistic data where conventional quantitative data was not readily available can be found in the project team’s article in *Archives and Records* [3], and in a forthcoming chapter in the *Proceedings of the European Conference for Mathematics in Industry* [4].

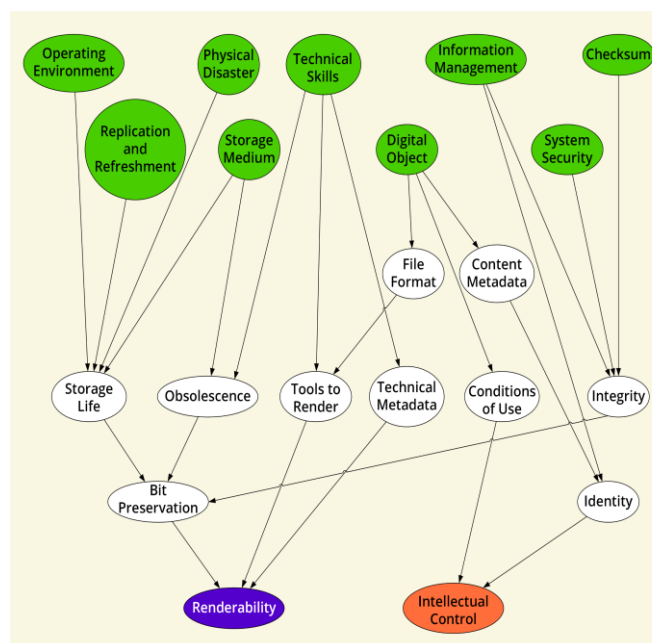


Figure 1 The network of risks behind DiAGRAM. Input nodes in green (top two rows). Each node has an associated conditional probability table which forms the full Bayesian Network.

Following the initial model development the project team presented a series of webinars to introduce DiAGRAM to the wider digital preservation community and to seek feedback on the tool. Formal user testing was also undertaken to assess usability. As the project had switched to engaging with stakeholders remotely due to the pandemic, we had unspent grant money which the National Lottery Heritage Fund approved to be repurposed on a further round of tool development to address the issues surfaced in the feedback and usability testing.

B. Initial Usability Improvements (2020)

Data analytics firm, Jumping Rivers [5], were appointed to make usability improvements to the tool initially developed by the project [6]. In addition to the usability issues that had been identified, we wanted to ensure that DiAGRAM was compliant with

the UK’s Public Sector Bodies (Websites and Mobile Applications) (No. 2) Accessibility Regulations 2018 [7] (PSBAR). This was the UK’s implementation of Directive (EU) 2016/2102 of the European Parliament and of the Council of 26 October 2016 on the accessibility of the websites and mobile applications of public sector bodies [8]. Essentially, this generally requires websites to comply with the World Wide Web Consortium (W3C) Web Content Accessibility Guidelines version 2.1 at the AA level [9] (WCAG 2.1).

At this point the remaining project budget gave us only 6 weeks’ development time. It was clear that this would not be sufficient to resolve all issues so it was decided to concentrate on the initial simple model building and scenario generation and to resolve as many general accessibility issues as possible.

Two sets of usability issues were addressed at this stage, to improve the ease of understanding and inputting percentages and to provide additional context to aid interpretation of the tool’s output.

The initial simple modeling process generally required archivists to enter percentage values for the input nodes. This is relatively straightforward for areas such as the breakdown of the types of digital objects in the archive or the types of storage media being used: but for technical skills, information management and system security users clearly found it much harder to give percentages in a consistent and meaningful way.

To guide users through these steps a series of structured inputs was developed drawing on existing digital preservation maturity models with weights in percentage terms assigned to the answers. For example, for the technical skills question the DigCurV skills framework [10] was used to select a subset of skills and levels of ability through the input screens of DiAGRAM which are then mapped to a percentage and fed into the underlying BN.

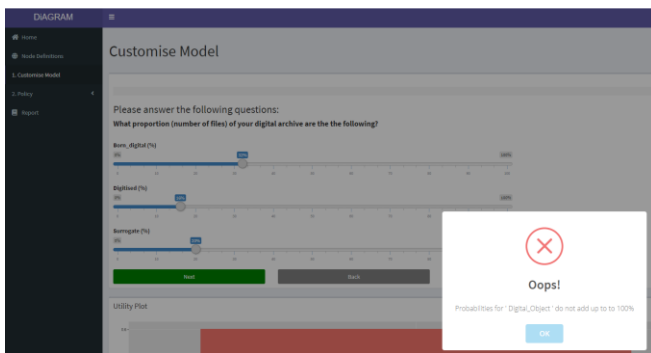


Figure 2 Original data entry screen for the Digital Object node showing sliders with no alternative data entry method and the warning produced if the percentages entered did not total 100%

Another general usability issue was that users were not prevented from attempting to submit values for a node which totaled more or less than 100%, although the DiGRAM prototype did then return an error telling users that this was a requirement. However, there was no assistance to users in calculating the percentages to ensure a total of 100% was reached. To resolve this input fields were linked and constrained so that changing the value in one linked field would lead to other linked fields being automatically adjusted to maintain an overall value of 100%. This was straightforward for pairs of values but is harder to implement where there are three choices. There was also no alternative data entry method other than controlling the sliders with the mouse which is also a failure of the accessibility criteria.

The other issue left unresolved at the end of the development period was giving users sufficient context to understand the scores generated for Renderability and Intellectual Control, although discussion of how to do this had taken place. Fortunately the project team were able to complete this in-house, introducing two built-in reference models. These represent a simple commercial backup service and a well-established digital preservation program at a generic national archives. While commercial backup gives some likelihood of being able to render a file later (at least in the short term), it has a very low Intellectual Control score as such services do not typically go into extracting technical metadata about files or determining descriptive metadata and conditions of use. The national archives model is more balanced (and scores much higher on both fronts), though it also demonstrates that there is likely to be remaining risk even in such an institution. As the answers to the input questions are also provided for these

reference models they can also be treated as a template for creating other models.

These changes were also evaluated against WCAG 2.1 using automated testing to improve accessibility. However, it was soon realized that the use of R/Shiny Dashboard meant that it would not be possible to make DiGRAM fully compliant within the time available. This was because the framework and underlying libraries generate much of the HTML and JavaScript sent to the browser and there were sometimes limited opportunities to intervene and rewrite this in an accessible way. In particular the plots used to show model scores were being generated using Plotly [11], which integrates easily within R/Shiny, but the interactive display element is not fully accessible and cannot easily be customized to resolve the issues. The R/Shiny Dashboard also generates a single page app which appears to have a tabbed navigational structure but does not have true URLs for each tab which is not compliant.

At the end of the development period the redeveloped DiGRAM [12] had much improved usability (as measured by feedback from further webinars) but still had a significant number of accessibility issues. This led to the decision to continue badging DiGRAM as a prototype rather than looking to formally make it a live service. It was felt that this mitigated the risks to The National Archives of having a non-compliant tool in the short term, but we knew we would have to make a further attempt to resolve the accessibility issues and bring DiGRAM into compliance with PSBAR once budget could be found.

C. Modeling With DiGRAM

In parallel with the development work, the project team had been applying DiGRAM to model the risks to The National Archives' (TNA) own digital holdings in order to support the organization's business case for investment in the digital archive. This was expected to set TNA's high level budget for the next three years.

The initial model was created using the simple modeling process. However, this was felt to not fully reflect the nuances of our situation, so we adjusted the model using the "Advanced customisation" options.

For this more advanced modeling archivists can directly edit the conditional probabilities associated with each node of the BN. The built-in probabilities

typically reflect median values, particularly for the data obtained via the expert elicitation process. In some cases archives may reasonably decide that the situation in their archive is better than the median and so substitute alternative probabilities (such as the 95th percentile value from the expert elicitation). In our case, due to factors such as the reliability of our tape storage system, we opted to use the relevant 95th percentile value for a number of nodes.

We then wished to model various scenarios to show the potential impacts of differing levels of funding, contrasted against the default position of no increase in investment. Flat cash means receiving exactly the same funding, with no allowance for inflation, so is in real terms a budget reduction. This was modeled as giving a compounded decay in areas such as Obsolescence and Technical Skills of 5% per annum.

From the project team's own use of "Advanced customisation" it appeared that it was broadly usable, albeit with a few bugs and issues to be understood for data entry. Of course the project team also had a higher familiarity with the statistical modeling concepts and the underlying model than would be the case for a typical archivist.

D. Further Developments (2021)

The team were able to use DiAGRAM to provide quantitative evidence to support the business case for investment in TNA's digital archive. The success of this application of modeling with DiAGRAM demonstrated that further development was warranted to enable the tool to be used more widely.

Following discussions with usability experts in the Digital Services Department at TNA, we began by commissioning a formal accessibility review of DiAGRAM. This was carried out by accessibility and inclusion specialists, TetraLogical [13]. As expected, a high number of issues (of varying severity) were reported following the audit, although it should be noted that a high proportion of these were effectively duplicates since some of the same components are used on multiple pages within DiAGRAM. Fixing the underlying component once would fix all occasions on which it was used.

With audit results available we again appointed Jumping Rivers to carry out the work, with ongoing assistance from the TetraLogical helpdesk.

The initial development task was defined as reviewing the current architecture and potential

alternatives to determine the most appropriate approach to resolving the accessibility issues. In addition we had to take into account the wider UK Government Digital Service (GDS) service standard [14] and determine the extent to which the different options would allow us to align with that.

While it was determined that the accessibility issues probably could be resolved within the existing R/Shiny Dashboard framework, the wider considerations of the service standard, in particular reducing the use of JavaScript in favor of HTML where possible in line with the GDS principle of progressive enhancement (that is, a service should work as far as possible with HTML only, without the use of CSS, JavaScript or other technologies except to provide a more refined user experience if they are available) [15], led to the decision to move to a new architecture.

The front end would be rewritten in native HTML as far as possible, some JavaScript is retained, particularly since the data structures required for input into the underlying model are too large to be handled via cookies. We also wanted to retain the original back-end feature that no data is stored on the server beyond the life of the user's session. This had proved popular in the original feedback as the data required for model building and the risk scores obtained are somewhat sensitive, and this means that neither The National Archives nor other partner has any access to the data used by an institution to model the risks to their digital archive. Users can however download their own model data for subsequent reuse if desired.

Much of the backend code (particularly that forming the BN itself) could remain essentially unchanged, but a new Application Programming Interface (API) would be required to enable communication between the front end and the back end code.

The API was generated using the R package plumber [16]. In line with further GDS guidance on the development of APIs [17] these have been documented with Swagger [18], although there is no intention currently to make the API publicly available as an alternative means of interacting with DiAGRAM.

Given the decision to redevelop the front end it also seemed appropriate to investigate options for some improvements to the "Advanced customisation" modeling since this had not been

done in the prior round of development. This was aided by the availability of one of TNA's Collaborative Doctoral Programme students on a professional placement supported by the Arts and Humanities Research Council's Additional Student Development Fund. As her PhD is in the area of Human Computer Interaction this provided the project with someone who could undertake user research while providing appropriate professional experience.

Initial user research for this phase was undertaken in mid-December 2021 with four users. This, in conjunction with the project team's own experience in using "Advanced customisation" modeling, was sufficient to determine the main issues and to propose improvements to the user journey and flow. Following the improvements, a further round of testing was undertaken, again with four participants (two of whom took part in the initial testing and two new participants) to validate the success of the changes.

This further development phase will conclude with a formal re-test of compliance with accessibility regulations carried out by TetraLogical.

II. USER TESTING

A. *Approach For Initial Usability Improvements*

Given the time constraint, with Jumping Rivers working on the project for a short period, and the limited capacity of the team, the goal was to identify the main areas of concern, prioritize them and implement simple and effective changes that would solve a maximum number of users' problems.

Before Jumping Rivers joined the project, the team performed a basic accessibility audit by manually testing the tool [19, 20]. These easy checks enabled us to detect a few accessibility problems. However we knew that this accessibility check was not sufficient and we needed to run detailed checks with automated tools and conduct testing with assistive technologies.

During a preliminary workshop, an inventory of the previous users' insights was drawn up. Usability issues were categorized and prioritized considering their impact, the number of users affected and their plausible repetitiveness. Interpreting the severity of these issues, the DiAGRAM team was able to list a dozen priorities to focus on during the following

sprints. The main areas of improvement were: content, navigation and accessibility. It was decided to divide the efforts in several working sessions, allowing the team to brainstorm and implement straightforward, yet effective changes to the different points of focus.

Two rounds of moderated remote user testing sessions [21] took place after recruiting research participants using a survey shared within The Digital Preservation Coalition (DPC) Network. Participants were recruited to reflect representative users of DiAGRAM: primarily archivists and record managers, people with a broad range of work experiences and skills; from new joiners to subject matter experts. Making DiAGRAM accessible and inclusive for many users as we could was key. We conducted the sessions with participants that had different levels of digital confidence. We tried to recruit participants with accessibility needs but it was difficult as our users are specialist target audiences. We realized that we needed help from agencies, charities and disability networks for the next iterations on the tool.

During testing, participants were asked to undertake a series of short and representative tasks with the tool while being observed by an interview and observer. Participants were asked to 'think-aloud' [22] as they completed the tasks, describing their actions as well as their thoughts and any reasoning behind their actions. Any actions requiring clarification or follow-up were probed by questions from both the interviewer and observer. The feedback gathered allowed the team to confirm whether or not the solutions put in place were efficient and uncovered new users' needs.

Jumping Rivers made significant alterations on the risk assessment tool using the outputs of the workshop, the design work sessions and the usability sessions. The DiAGRAM team received positive feedback on the content and the design changes made on the tool.

Due to tight deadlines, the team were unable to deal with the complex accessibility issues identified. However, the work done on content, which is in many aspects part of Accessibility benefited all users including users with accessibility needs.

B. Approach For “Advanced customisation” Improvements

When undertaking simple modeling, archivists answer a series of questions to fit DiAGRAM to their archives’ current situation. Conversely, “Advanced customisation” allows users to input probabilities directly and edit lower level conditional nodes which are not included in the basic modeling process. It is therefore aimed at larger or more specialized organizations which are more likely to differ from the median values which are incorporated in the default statistical data used in the BN.

During previous usability testing, the “Advanced customisation” feature in DiAGRAM was deprioritized and therefore not fully tested. As the original application in the R/Shiny framework remained live, it was possible to use this as a functioning prototype for testing.

As with previous usability testing, an iterative approach was taken, comprising two short rounds of testing. Adopting this approach made it possible to identify the majority of issues and allowed for iterative development of DiAGRAM as it was transferred to the new site [23]. We began with an evaluation of the “Advanced customisation” page, using a heuristic framework [24] which drew attention to potential problem areas where we would focus testing: namely, navigation and content.

Following this, we conducted two rounds of usability testing, each with three to four participants recruited from TNA’s pre-existing contacts. Participants included digital archivists; a digital analyst; a data lead; and a digital archives manager. The first round of testing sought to identify usability issues, focusing on 1) the user journey through the “Advanced customisation” page; and 2) page content, such as in the ‘Edit’ tables where changes to nodes are made.

Alterations were then made to the design of the page as it was implemented in the new site and a second round of usability testing was undertaken to validate the changes that had been made. Firstly, the elements were reordered so that the user journey flowed from the top to the bottom of the page. Secondly, an SVG image file was also added as an alternative means of navigation, though a dropdown box was retained to meet accessibility requirements. Several other minor

recommendations to improve usability were also made.

	Storage Medium	Technical Skills	Yes	No
1	A	Good	0.30	0.70
2	A	Poor	0.64	0.36
3	B	Good	0.14	0.86
4	B	Poor	0.50	0.50
5	C	Good	0.00	1.00
6	C	Poor	0.00	1.00

Node			
▼ Obsolescence			
Storage_Medium	Technical_Skills	Yes	No
B	Poor	0.5	0.5

Figure 3 Original “Advanced customisation” screen. The order of nodes in the “Choose” dropdown was unclear and the proximity of “Add change” and “Store Model” buttons confused users making it unclear when each should be used.

Editing: Obsolescence				
	Storage Medium	Technical Skills	Yes	No
1	A	Good	0.2985	0.7015
2	A	Poor	0.6422	0.3578
3	B	Good	0.1405	0.8595
4	B	Poor	0.5000	0.5000
5	C	Good	0.0010	0.9990
6	C	Poor	0.0010	0.9990

Log of changes				
▼ Obsolescence				
	Storage Medium	Technical Skills	Yes	No
1	A	Good	0.2985	0.7015
2	A	Poor	0.6422	0.3578
3	B	Good	0.1405	0.8595
4	B	Poor	0.5000 (0.5000)	0.5000 (0.5000)
5	C	Good	0.0010	0.9990
6	C	Poor	0.0010	0.9990

Figure 4 Revised “Advanced customization” screen. Initial setup moved to separate screen. Network diagram now used to choose node to edit (dropdown also remains available). “Store model” moved to bottom right to clarify order of operations.

Participants generally found the flow much improved by the reorientation of elements. All participants opted to use the SVG image to navigate the model. There was some indication that having the model to hand - in the form of the SVG image - helped with interpretation of some of the nodes, although participants continued to express some difficulty when interacting with lower level conditional nodes. This suggests that further consideration may be necessary to communicate the conditional nature of the model more effectively.

While qualitative data such as this holds high validity, testing with such small numbers naturally limits the generalizability of such findings. Therefore, to complement these findings a survey has been devised, incorporating System Usability Scale (SUS) [25, 26] to gather quantitative data on the usability of the website from a much larger number of users.

III. ACCESSIBILITY

Our experience in this project points to the importance of considering accessibility from the beginning of development of a new tool or service. However, given the short timeframe of the original project there were also valid reasons for developing the initial prototype in R/Shiny Dashboard.

Previous risk management frameworks within digital preservation have been more qualitative in nature [3], and there does not appear to have been a previous attempt to develop an Integrated Decision Support System in the field. Coupled with the fact that few archivists have a background which had given them much prior exposure to Bayesian Statistics and related concepts meant that we initially needed to develop a prototype very quickly (using a BN from a different domain) in order to help introduce the concepts to the archivists involved in the project and show what might be possible.

However, as the initial project was NLHF funded (with their current guidance on web accessibility dating to August 2020, after the start of the project [27, 28], there was essentially no digital guidance available at the start of 2020 [29]), and since the IDSS does not really fall within the GDS definition of a transactional service [14] “Your service is transactional if it allows users to either:

- exchange information, money, permission, goods or services

- submit personal information that results in a change to a government record”,

we did not initially focus strongly on accessibility (a weakness of the original project team was arguably the lack of a specialist UI/UX researcher who might have prompted greater focus on this area from the beginning).

This early work could also be considered as being in line with the Discovery and Alpha phases of an Agile project given the exploratory nature of the initial working together of archivists and statisticians and the uncertainty over what an IDSS might look like in this context.

With the opportunity to repurpose part of the project budget that had not been spent on in-person events as originally planned came the realization that accessibility needed to be improved, helped by the welcome addition of a user researcher to the project team.

We knew that given the very limited time period it was unlikely that we would be able to resolve all accessibility issues during this phase of development. Members of staff at The National Archives with some experience of accessibility undertook an informal review ahead of development starting which gave us a number of high priority work areas. Jumping Rivers also introduced the use of Koa11y, an automated assessment tool [30, 31]. However, it is well known that automated assessment and testing can only catch a fraction of the potential issues, particularly since aspects of accessibility such as a logical flow for users of assistive technologies (such as screenreaders) do not lend themselves to automated assessment [32].

Despite these constraints progress was made, and the wider usability improvements should also have contributed to accessibility, such as not requiring mental arithmetic from users to ensure that sets of values totaled 100%. Custom HTML and JavaScript was created to work around some issues arising from the R/Shiny Dashboard ecosystem, keyboard only navigation and data entry worked across much of the tool, but we knew significant obstacles to true accessibility remained, not least the single page app paradigm imposed by R/Shiny Dashboard which made it hard to navigate across the whole site in an accessible way.

There does also seem to be a lack of appreciation of the importance of accessibility within the open

source data visualization community. DiAGRAM originally used the plotly widget to provide visualization of the scoring of the two output nodes, Renderability and Intellectual Control. Again this was the natural choice from within the R/Shiny Dashboard ecosystem and for a user interacting with an online visualization via mouse offers an easy way to download an image of a chart, zoom in and out, select areas of the chart, and display further information about a data point, or compare data points. However, none of these actions have accessible alternatives by using the keyboard or other forms of interaction, nor is alt text or equivalent provided for to allow access to the data by non-visual means. An issue was raised in plotly's GitHub repository on 24 May 2016 asking for keyboard shortcuts to be added for accessibility (referencing the requirements of the US Section 508 rules, essentially the US Federal Government equivalent of PSBAR [33]), since then multiple others have supported the request, and other similar issues have been created but it does not appear to have been given much priority by maintainers, although the "needs sponsor" tag was added on 10 September 2020 [34]. Here of course we see another issue frequently raised in relation to open source coming into play, the lack of (financial) support for maintainers from those using an open source product, one problem here perhaps being the difficulty of fitting existing government procurement models to such sponsorship. Similar problems arguably exist within the ecosystem of open source digital preservation tools.

As we prepared to embark on a further phase of development we sought further advice from colleagues in the Digital Services Department at The National Archives who are more familiar with the challenges of accessible design, and undertook some introductory training [35]. As a result it was decided to commission a formal review of DiAGRAM from a specialized firm. After obtaining quotes from several, TetraLogical were appointed. In discussion with the project team they prepared a detailed test plan, ensuring coverage of all key features across the site (due to overlapping functionality and reuse of components it was not necessary to test every page). Following the completion of development they will also retest and sign off the work as compliant (assuming the work has been successful) and prepare an appropriate accessibility statement in compliance with the legislation.

We opted for what TetraLogical describe as a Lightning Report. For each area of the site under test the full set of WCAG 2.1 Success Criteria are listed and each criterion is marked as Fail, Pass, Not Applicable, and Not Tested. Fails are then graded on criticality as Low, Medium, High, or Critical. In total 16 areas of the site were assessed with 187 fails, 322 successes and 291 not applicable (there were no instances of criteria not being tested). Of the fails, 4 were deemed critical, 6 high, 52 medium and 125 low.

The critical issues related to the slider used on the Physical Disaster screen (within Create a model) not being keyboard operable, similarly it was not possible to select a model on which to base a scenario via the keyboard, nor to change the selection of models on View results or Download Results (these three all used the same underlying table component, the test also revealed what was actually a bug, using the mouse it was possible to double-click on some cells, such as Intellectual Control and then manually edit the value, such editing should not have been possible at all).

For each identified issue there was also a brief suggestion of how it might be possible to resolve it. TetraLogical also offer a more in depth service where they will raise detailed tickets in your system of choice (eg Jira, GitHub etc) with more comprehensive information on resolving the issues identified. However, they state that this is more appropriate when upskilling your own developers in developing accessible websites, rather than when contracting development to others (and is more expensive). They also provide a paid helpdesk service to provide advice during development. As part of their reappointment for the second round of redevelopment Jumping Rivers opted-in to this helpdesk service to strengthen their ability to deliver an accessible service.

This second period of redevelopment began with Jumping Rivers reviewing the existing architecture of DiAGRAM in order to make recommendations on the most appropriate route forward allowing for all accessibility issues to be resolved.

While they believed that it would be possible to resolve all issues within the existing R/Shiny Dashboard architecture, this would have required a large amount of custom JavaScript which The National Archives viewed as potentially risky, and in conflict with GDS principles of progressive

enhancement which provides that services should work even only HTML is available in the browser, but then additional styling and refinement of service can be added through the use of technologies such as CSS (Cascading Style Sheets) and JavaScript.

With this in mind, and taking into account the available time and budget, it was ultimately decided to take the opportunity for a more thorough reworking of DiAGRAM, moving to a relatively simple front end and retaining the main parts of the existing back-end (including, for example, the BN written in the R package gRain [36]) and introduce a new API layer using the R package plumber [16] to create the necessary endpoints to allow communication between the front- and back-ends.

As the data that comprises the model inputs exceeds the maximum permitted storage for cookies we have not been able to remove the need for JavaScript entirely, but the rebuild has moved us towards the principles of progressive enhancement with the quantity of JavaScript being greatly reduced across DiAGRAM, and avoiding using heavyweight frameworks. To eliminate JavaScript completely would probably have required us to move to having backend data storage of some sort, further complicating the rebuild. Also, feedback during the first round of development had been appreciative of the fact that there was no permanent storage of model data because this meant that other archives did not feel that TNA were “looking over their shoulder” and being able to see data on digital preservation risk levels that archives were not ready to share.

From November 2021 to March 2022 Jumping Rivers worked on the redevelopment. Work on the front- and back-ends was carried out in parallel, with the front-end using dummy data until the relevant API endpoints were available. Development was carried out in dialog with the TetraLogical helpdesk to ensure that accessibility targets were met. The redeveloped DiAGRAM [37] underwent a formal retest against the accessibility criteria at the end of March 2022. A few remaining small issues were identified which will be addressed before this version is made available. A slightly more problematic area is the PDF that is made available for download which still has several issues, investigations will continue to see how this can be made fully accessible. However, all information that is wrapped in the PDF is available in accessible form elsewhere in DiAGRAM. Following a further tender process Jumping Rivers were also

appointed to create the live environment for the redeveloped version of DiAGRAM and provide ongoing application support [38]. This new version of DiAGRAM is expected to be launched in late 2022.

IV. CONCLUSION

It is important to acknowledge the overall success of the Safeguarding the Nation's Digital Memory project which led to the creation of DiAGRAM, as evidenced in the DPC's evaluation report [39], the shortlisting of DiAGRAM for the Digital Preservation Awards 2020 [40], TNA's own use of modeling in support of making the case of investment (as described in this paper), and early evidence of wider usage such as described in a case study in the DPC's EDRMS preservation toolkit [41]. In October 2022 the project was announced as the winner of the 2022 Decision Analysis Practice Award by the Society of Decision Professionals and the Decision Analysis Society [42]. However, there are lessons for us all to learn in some areas of the project.

The development process could have been streamlined with an earlier focus on the issue of accessibility. This may not have changed the initial decision to prototype in R/Shiny Dashboard due to the ability this gave to easily connect to the back end model, while making initial rapid iteration of designs at the front end relatively straightforward. However, this would have given a clearer view of the trade-offs we were making, which would have allowed earlier planning of a route to achieving full accessibility of the tool.

Similarly a more explicit framing of what the initial funded project was aiming to deliver as effectively the Discovery and Alpha phases of an Agile project, producing a useful proof of concept, would have helped manage expectations around the amount of work that would still be required to create a true live product. What the project achieved was actually somewhat more (particularly after the initial round of work by Jumping Rivers) than the project team had really been anticipating.

As GLAM institutions around the world grapple with wider questions of diversity, equity and inclusion in relation to our collections and practice it is vital that we consider this in relation to the accessibility of our digital presence and tools as well: at The National Archives this is an integral part of our desire to become the Inclusive Archive. Many GLAM

organizations around the world are public bodies so will be operating under equivalent regulations to PSBAR (particularly those organizations within EU countries) so the considerations outlined in this paper should be broadly applicable.

Reflecting on the conference themes: the original project was very much centered around making our shared digital heritage more resilient, which was a key project outcome we needed to demonstrate to the National Lottery Heritage Fund[39] and an improved funding position for The National Archives is an early success in that direction. We also sought to broaden the community of practice which could participate in digital preservation by providing a tool which would help members of the community make an evidence-based argument for particular interventions in their context. If the tool we provide is not fully accessible we will be failing in that aim by excluding some members of the community.

REFERENCES

- [1] The National Archives, *Safeguarding the nation's digital memory*. [Online]: Available: <https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/research-collaboration/safeguarding-the-nations-digital-memory/> (accessed Mar. 8, 2022).
- [2] Jumping Rivers, The National Archives, Applied Statistics & Risk Unit, University of Warwick, "DiAGRAM", *Glossary: Definitions of nodes*. [Online]: Available <https://diagram.nationalarchives.gov.uk/glossary.html#definitions> (accessed Oct. 20 2022)
- [3] M. Barons, S. Bhatia, J. Double, T. Fonseca, A. Green, S. Krol, et al., (2021, May) "Safeguarding the nation's digital memory: towards a Bayesian model of digital preservation risk." *Archives and Records* [Online]. vol. 42, issue 1, pp. 58-78. Available: <https://doi.org/10.1080/23257962.2021.1873121> or [open access]: <https://www.nationalarchives.gov.uk/about/our-research-and-academic-collaboration/our-research-projects/open-access-research-from-our-staff/safeguarding-the-nations-digital-memory-towards-a-bayesian-model-of-digital-preservation-risk-article/>
- [4] M. J. Barons, T. C. O. Fonseca, H. Merwood and D. H. Underdown "Safeguarding the Nation's Digital Memory: Bayesian Network modelling of digital preservation risks," Selected topics from 21st ECMI Conference on Industrial and Applied Mathematics, in press.
- [5] Jumping Rivers (2021) [Online]. Available: <https://www.jumpingrivers.com/> (accessed Mar. 8, 2022).
- [6] Jumping Rivers, The National Archives, Applied Statistics & Risk Unit, University of Warwick, "DiAGRAM, Version 0.8.0" [Online]. Available: https://nationalarchives.shinyapps.io/tna_gui/ (accessed Oct. 20, 2022).
- [7] UK Statutory Instruments 2018 No. 952 *The Public Sector Bodies (Websites and Mobile Applications) (No. 2) Accessibility Regulations 2018*. [Online]. Available: <https://www.legislation.gov.uk/uksi/2018/952/contents>
- [8] Directive (EU) 2016/2102 of the European Parliament and of the Council of 26 October 2016 on the accessibility of the websites and mobile applications of public sector bodies (Text with EEA relevance). (2016, Oct. 26), EUR-Lex. [Online]. Available: <https://eur-lex.europa.eu/eli/dir/2016/2102/oj>
- [9] Web Content Accessibility Guidelines (WCAG) 2.1, W3C Recommendation, 5 June 2018. [Online]. Available: <https://www.w3.org/TR/WCAG21/>
- [10] DigCurV, (2013) *Skills in Digital Curation Curriculum Framework*. [Online]. Available: <https://digcurv.gla.ac.uk/skills.html> (accessed Mar. 8, 2022).
- [11] Plotly, *The front end for ML and data science models*. [Online]. Available: <https://plotly.com/> (accessed Mar. 8, 2022)
- [12] Jumping Rivers, The National Archives, Applied Statistics & Risk Unit, University of Warwick, "DiAGRAM Version 0.11.0 (Prototype)" [Online]. Available: <https://nationalarchives.shinyapps.io/DiAGRAM/> (accessed Oct. 20, 2022).
- [13] TetraLogical, *Hello, we're TetraLogical*. [Online]. Available: <https://tetralogical.com/> (accessed Mar. 8, 2022).
- [14] GOV.UK, *Service Standard*. [Online]. Available: <https://www.gov.uk/service-manual/service-standard> (accessed Mar. 8, 2022).
- [15] GOV.UK, *Service Manual-Technology-Building a resilient frontend using progressive enhancement*. [Online]. Available: <https://www.gov.uk/service-manual/technology/using-progressive-enhancement> (accessed Mar.8, 2022).
- [16] B. Schloerke, J. Allen, B. Tremblay, F. van Dunné, S. Vandewoude, RStudio. CRAN - Package *plumber* : An API Generator for R. [Online]. Available: <https://cran.r-project.org/package=plumber>
- [17] GOV.UK, *API technical and data standards (v2 - 2019)*. [Online]. Available: <https://www.gov.uk/guidance/gds-api-technical-and-data-standards> (accessed Mar. 8 2022).
- [18] SmartBear Software (2021), *Swagger - API development for everyone*. [Online]. Available: <https://swagger.io/> (accessed Mar. 8 2022).
- [19] W3C, Web Accessibility Initiative, *Easy Checks – A First Review of Web Accessibility*. [Online]. Available: <https://www.w3.org/WAI/test-evaluate/preliminary/> (accessed Mar. 8 2022).
- [20] GOV.UK - Central Digital & Data Office. (2019, Aug. 22). *Doing a basic accessibility check if you cannot do a detailed one*. [Online]. Available: <https://www.gov.uk/government/publications/doing-a-basic-accessibility-check-if-you-cant-do-a-detailed-one/doing-a-basic-accessibility-check-if-you-cant-do-a-detailed-one> (accessed Mar. 8, 2022).
- [21] K. Moran, K. Pernice, Nielsen Norman Group. (2020, Apr. 12). *Remote Moderated Usability Tests: Why to Do Them*. [Online]. Available: <https://www.nngroup.com/articles/moderated-remote-usability-test-why/> (accessed Mar. 8. 2022).
- [22] S. Makri, A. Blandford, A. L. Cox. (2011, May). "This is what I'm doing and why: Methodological reflections on a naturalistic think-aloud study of interactive information behaviour." *Information Processing & Management* [Online]. vol. 47, issue 3, pp. 336-348. Available:

<https://doi.org/10.1016/j.ipm.2010.08.001> (accessed Mar. 8, 2022).

- [23] J. Nielsen, Nielsen Norman Group. (2000, Mar. 18). *Why You Only Need to Test with 5 Users*. [Online]. Available: <https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/> (accessed Mar. 8, 2022).
- [24] J. Nielsen, Nielsen Norman Group. (1994, Nov. 1). *How to Conduct a Heuristic Evaluation*. [Online]. Available: <https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/> (accessed Mar. 8, 2022).
- [25] Brooke, J. (1996). SUS: A “quick and dirty” usability scale. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 189–194). London, UK: Taylor & Francis.
- [26] J. R. Lewis, (2018, Mar.). “The System Usability Scale: Past, Present, and Future”. *International Journal of Human-Computer Interaction* [Online]. vol. 34, issue 7, pp. 577–590. Available: <https://doi.org/10.1080/10447318.2018.1455307> (accessed Mar. 8, 2022).
- [27] A. McNaught, National Lottery Heritage Fund (2020). *Digital Skills for Heritage: Accessibility Online*. [Online]. Available: <https://www.heritagefund.org.uk/sites/default/files/media/attachments/Digital%20guide%20Introduction%20to%20online%20accessibility.pdf> (accessed Mar. 8, 2022).
- [28] National Lottery Heritage Fund (2020, Aug. 11). *Digital resources for heritage organisations*. [Online]. Archived version available in UK Government Web Archive: <https://webarchive.nationalarchives.gov.uk/ukgwa/20200817161839/https://www.heritagefund.org.uk/publications/digital-skills-heritage-digital-resources> (archived Aug. 17, 2020)
- [29] National Lottery Heritage Fund. *Digital guidance for applicants*. [Online]. Archived version available in UK Government Web Archive: <https://webarchive.nationalarchives.gov.uk/ukgwa/20200203113500/https://www.heritagefund.org.uk/publications/digital-guidance-applicants> (archived Feb. 3, 2020).
- [30] Open Indy Brigade's Civic UX Group. (2017, Oct.). *Koa11y*. [Online]. Available: <https://open-indy.github.io/Koa11y/> (accessed Mar. 8, 2022).
- [31] The Accessibility Project. *What does the term a11y mean?*. [Online]. Available: <https://www.a11yproject.com/about/#what-does-the-term-a11y-mean> (accessed Mar. 8, 2022).
- [32] University of Minnesota Duluth-Information Technology Systems and Services. *Strengths and Limitations of Automated Tools*. [Online]. Available: https://www.d.umn.edu/itss/training/online/wave/strengths_limits.html (accessed Mar. 8, 2022).
- [33] Architectural and Transportation Barriers Compliance Board. (2000, Dec. 21). Rule. *Electronic and Information Technology Accessibility Standards*. [Online]. Available: <https://www.federalregister.gov/documents/2000/12/21/00-32017/electronic-and-information-technology-accessibility-standards> (accessed Mar. 8, 2022).
- [34] J. P. Schmit. (2016, May 24), *Issue #562: Use keyboard to navigate plotly charts*, in GitHub repository plotly/plotly.js. [Online]. Available: <https://github.com/plotly/plotly.js/issues/562> (accessed Mar. 8, 2022).
- [35] University of the Arts, London: creative computing institute, Institute of Coding. *Introduction to UX and Accessible Design*. [Online]. Available: <https://www.futurelearn.com/courses/introduction-to-ux-and-accessible-design> (accessed Mar. 8, 2020)
- [36] S. Højsgaard. CRAN - Package *gRain: Graphical Independence Networks*. [Online]. Available: <https://cran.r-project.org/package=gRain> (accessed Mar. 8, 2022).
- [37] Jumping Rivers, The National Archives, Applied Statistics & Risk Unit, University of Warwick, DiAGRAM prototype version 0.11.0 in development. [Online]. Available (login required): <https://nata-dia2.impr.io/frontend/staging/> (accessed Mar. 8, 2022).
- [38] GOV.UK - Contracts Finder, The National Archives. (2022, Feb. 22). *Cloud Hosting Design, deployment and support for DiAGRAM (the digital archiving graphical risk assessment model)*. [Online]. Available: <https://www.contractsfinder.service.gov.uk/Notice/d7f128db-33e6-41d9-b7f1-c6cccd75df59> (accessed Oct. 20, 2022)
- [39] J. Mitcham, A. Currie, W. Kilbride, the Digital Preservation Coalition. (2021, February). *Safeguarding the Nation's Digital Memory-Project Evaluation Report for the National Archives-FINAL*. [Online]. Available: <http://doi.org/10.7207/op21-01> (accessed Mar. 8, 2022).
- [40] Digital Preservation Awards 2020. *DiAGRAM (the Digital Archiving Graphical Risk Assessment Model created by the Safeguarding the Nation's Digital Memory project)*. [Online]. Available: <https://www.dpconline.org/events/digital-preservation-awards/dpa2020-diagram> (accessed Mar. 8, 2022).
- [41] N. Steele, Grosvenor Estates. “Determining risk: a case study”. *EDRMS Preservation Toolkit*. [Online]. Available to members: <https://www.dpconline.org/digipres/implement-digipres/edrms-preservation-toolkit/edrms-toolkit-further-resources/edrms-toolkit-case-study-nsteele> (accessed Mar. 8, 2020).
- [42] Society of Decision Professionals (2022, Oct. 19) [Online]. Available: <https://www.linkedin.com/posts/society-of-decision-professionals-2022-das-sdp-practice-awards-finalist-presentations-activity-6988134669561081856-eRWt> (accessed Oct. 20, 2022).

E-ARK, TEN YEARS AND STILL GOING STRONG:

Results, Use Cases And Benefits

Janet Anderson

Highbury R&D
Ireland

Janet.andersoneabb@gmail.com
[0000-0003-2673-4830](tel:0000-0003-2673-4830)

David Anderson

Highbury R&D
Ireland

cdpa@btinternet.com
[0000-0001-7643-4866](tel:0000-0001-7643-4866)

István Alföldi

Poliphon Kft
Hungary
alfi@poliphon.hu

Jaime Kaminski

Highbury R&D
Ireland

drjkaminski@gmail.com
[0000-0003-2907-0128](tel:0000-0003-2907-0128)

Carl Wilson

OPF
UK/Netherlands

carl@openpreservation.org

Diogo Proença

INESC-ID
Portugal

diogo.proenca@tecnico.ulisboa.pt
[0000-0002-3671-9637](tel:0000-0002-3671-9637)

Abstract – The E-ARK Consortium has been working steadily over the last ten years to provide specifications, tools, and best practices for digital archiving across Europe and beyond. The E-ARK Consortium has grown over the years and the work has taken place under different auspices: first as an EC-

¹. This paper comprises the eArchiving Building Block results, benefits, and Consortium member use cases as presented at the DLM Forum meeting in October 2021.

Keywords – E-ARK, Digital Europe Programme, eArchiving

Conference Topics – Innovation; Resilience.

I. INTRODUCTION

The impetus for the E-ARK effort came in 2011 when the Slovenian and Estonian National Archives were faced with archiving their e-government records. They could not manage this new and daunting task alone and so they got together with DLM (Digital Lifecycle Management) members to lobby the European Commission for help to produce a pan-European, usable digital archiving suite of specifications and tools. The first E-ARK project ran from 2014 to 2017 ²to deliver this. In E-ARK, a broad consortium of members: archivists, researchers, software developers and membership organisations,

funded PSP CIP pilot B project, then as the eArchiving Building Block under the Connecting Europe Facility (CEF) banner. The CEF Programme has just finished and eArchiving will be continuing under the Digital Europe Programme (DEP) as a procurement

pooled their expertise and products and produced the first tranche of specifications and software tools that could be used across Europe and beyond, for national, regional, local or cross-border digital archiving tasks. The E-ARK project focussed on archiving ERMS records, geospatial data, databases and also carrying out Big Data analysis on cross-border datasets.

The next instantiation of E-ARK from mid 2018 to late 2021 was as a Connecting Europe Facility (CEF) eArchiving Building Block³, whose Owner was the European Commission's (EC's) Directorate General (DG) CNECT in Luxembourg, with stakeholder management provided by the EC's DG DIGIT Stakeholder Management Office (SMO) in Brussels and with the E-ARK Consortium as the Solution Provider. eArchiving functioned alongside other building blocks including eSignature, eDelivery and Blockchain, together supporting the EC's Digital Single Market. From 1st April 2021 to 31st October

¹ The E-ARK Consortium were successful with this procurement and the kick-off was held on 14th October 2022.

² Grant Agreement No 620998, <https://eark-project.com/>

³ This was in the form of two EC grants: E-ARK4ALL,

Agreement number: LC-00921441 CEF-TC-2018-15 eArchiving (2018-2019), and E-ARK3, AGREEMENT No. LC-01390244 CEF-TC-2019-3 E-ARK3, (2019-2021).

2021 the newly-formed Health and Digital Executive Agency (HaDEA) in Brussels was responsible for the day-to-day administration of the eArchiving Building Block, together with the four eArchiving-dependent Generic Services projects which began in the autumn of 2021.

The specifications and tools were further developed and harmonised as part of the eArchiving Building Block work, and substantial training and outreach was conducted. The scope of the work increased to encompass eHealth records, cultural heritage and research data, and a comprehensive Reference Architecture was created, together with a Maturity Assessment model. The CEF Programme has now concluded, and eArchiving will continue under the Digital Europe Programme⁴ (DEP).

This paper will first delineate the various stages of E-ARK activity, then outline the E-ARK results: the specifications and tools etc. We will then move on to present a range of case studies from adopting organisations, followed by use cases from and benefits reported by the E-ARK Consortium member organisations, showing how it is possible to share the digital archiving burden.

II. E-ARK RESULTS

Overall reflections

Overall, the eArchiving Building Block was highly successful. In 2021 there were several high-profile, very well attended events where eArchiving use was strongly recommended to EU businesses, agencies, and institutions. The events included “eArchiving in Action” in January 2021, the CEF’s “Trust Café” in March 2021 and the CEF’s “DigitAll” meetings in April 2021. Following this outreach there was a step change: a very noticeable increase in our onboarding contacts, with several key EC departments getting in touch for long-term help, for example.

In terms of strategic collaboration, in December 2020, the eArchiving and eSignature building blocks got together to discuss close collaboration in order to provide a joint offering to users. eArchiving also worked with the Archiver project [1] in the research infrastructure domain where the two winning tender organisations are both using E-ARK specifications to build a pan-European European Open Science Cloud

(EOSC) research infrastructure. Other significant collaborations include working with experts in Digital Cultural Heritage sector and the Engineering Data Space.

Specifications

Turning to the eArchiving outputs, the specifications are now the core foundation of all the E-ARK undertakings as everything is built upon these key standards, which are now mature and used in production environments.

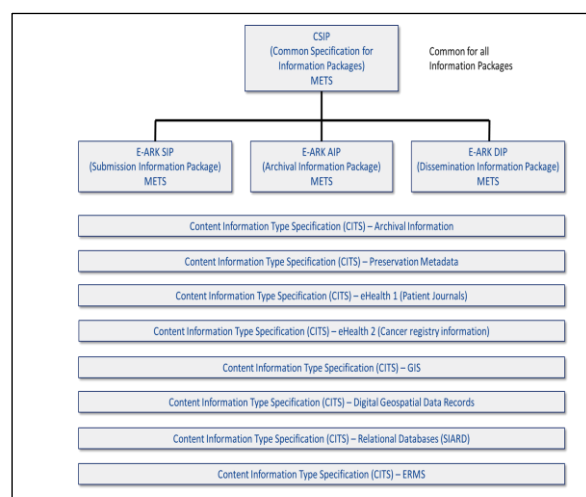


FIGURE 1

The specifications are maintained and developed by the Digital Information Lifecycle Interoperability Standards (DILCIS) Board [2] and the full specifications are available there, together with requirements, XML-schemas, Schematron code, guidelines and examples. In figure 1, the top four Information Package (IP) specifications are based on the Open Archival Information Standard (OAIS) [3] and include the Common Specification for Information Packages (CSIP); the Submission Information Package (SIP); the Archival Information Package (AIP); and the Dissemination Information Package (DIP). These are complemented by accompanying Content Information Type Specifications (CITS) which cover databases and the SIARD format (Software Independent Archival of Relational Databases); geospatial data; ERMSs; eHealth data; archival information for digitised material and lastly preservation data based on the PREMIS standard [4].

ITAL/Technical+Specifications There will be a new eArchiving website under the EC’s DEP.

⁴ Please note that the EC’s CEF eArchiving website is no longer live. An archived version is available at <https://wayback.archive-it.org/12090/3/https://ec.europa.eu/cefdigital/wiki/display/CEFDIG>

The E-ARK Sample Software Portfolio includes two mature, end-to-end, Open-Source digital archiving systems (RODA from KEEP Solutions and ESS Arch from ES Solutions (see below for more on the two SMEs)) plus many different Open-Source components and software libraries etc., all developed on a modular basis so that a mix-and-match approach can be adopted⁵. In terms of software, there was noticeably improved specification support: previously available sample software was extensively tested and improved to meet the v2.0.4 of the E-ARK IP specifications. All components are now on a next level of conformance. New software had also been created to support the specifications: a brand-new eHealth SIP Creator and an E-ARK IP Viewer. The new software components were available from the end of October 2021⁶. The sample software portfolio, and indeed every aspect of eArchiving, was covered by a dedicated Service desk which proved increasingly useful under the CEF Programme, even though eArchiving was not a hosted service. The Service desk provision ceased at the end of October 2021, but provision was made for the E-ARK Consortium to be contacted with any queries.

The eArchiving Building Block manages a large and diverse portfolio of components, including various specifications, tools, and services delivering support, training, and dissemination. The objective of release management is to sustain theoretical conformity and technical compatibility between the versions and revisions of the components of the eArchiving service portfolio. Release Management is an internal activity for the eArchiving team and end-users do not really have to see what guides the version numbering and release dependencies of the different eArchiving components. Here “no news” is really “good news”.

The Reference Architecture work on the other hand has been producing a lot of news. This was a new initiative at the beginning of the E-ARK3 project and we are immensely proud to have announced the first full version of the eArchiving Reference Architecture by the end of October 2021. In order to make the model easier to understand and use, a lot of example business scenarios and component layout views have been added to the pure ArchiMate diagrams, and an html based online version was created as well. The web-based application gives us

the opportunity to add introductory sections about the background and scope of the model, about ArchiMate, a glossary, and a Download area. The online model is available at the DLM Forum Knowledge Centre [5].

In terms of validation, a new validator and validation REST API were released mid-October 2021 [6], and the validation input was for version 2.1 of the specifications. A validation strategy report was released in June 2021. The final test corpus was reviewed to ensure that the validation rules were functioning as expected, and the SIARD CITS was assessed to see how practical it was to validate it.

D. Training

E-ARK training provides content to support existing E-ARK specifications, software, and tools. It has been delivered in the form of webinars, YouTube videos, modules on the Moodle platform, and supplementary online workshops and is an essential part of E-ARK's support for onboarding. E-ARK developed an integrated approach to training, where the Webinars and videos have been closely linked to Moodle training with cohesive branding. In total, over 3,000 delegates have booked on the 17 E-ARK training webinars since 2020. The online YouTube training videos have seen an additional 2,900 views.

The training has been driven in part by the results of the ‘user needs’ survey which has proved to be an important tool for understanding user requirements, providing a conduit to the user communities, thereby encouraging open communication, and raising awareness of the eArchiving Building Block and its services.

E. Outreach

Finally, Onboarding and outreach was one of the main priorities for the eArchiving Building block. This effort built on all the specifications and software components and systems to gather traction on the

⁵ <https://github.com/E-ARK-Software>

⁶ <https://github.com/eark-project>

adoption of the outputs E-ARK created, enhanced, and maintained over the last decade.

To put this effort in perspective, at the beginning of the Archiving Building block (Q3 2018) there were 14 organisations known to be reusing E-ARK specifications and software components many of whom are ‘multiplier’ organisations that are responsible for determining the standards that other organisations follow. Fast-forward to Q3 2021 we had 29 organisations reusing, six committed to reuse and 18 committed to analyse the use of E-ARK in their organisations. The evolution of these metrics is detailed in Figure 1. Moreover, there is a list of 300+ leads in our CRM ready to be onboarded (details on these leads are depicted in Figure 2). All these facts really demonstrate the interest in E-ARK for years to come.

These figures are also explained by the effort of the Consortium to broaden the horizons and expand beyond our original Archiving community to focus on other sectors, such as, Research Data, Digital Signature, Finance and Healthcare. This new avenue showed that the need and benefits of Digital Preservation are understood, but other communities are still taking their first steps in appreciating that data backups are not preservation and that there is a whole community ready to share the knowledge acquired over the last decades.

To better explain the benefits of digital preservation and guide new onboarding leads to adopt E-ARK, the Consortium focused on developing tools to aid in this effort, such as, the eArchiving Maturity assessment tool and the Reference Architecture⁷. The first tool can help organisations identify their level of maturity on the subject while the second tool can guide organisations on what tools and specifications to adopt in order to reach their target maturity level. Both tools were developed symbiotically to guarantee an integrated framework for assessment and improvement for organisations.

A good indicator of community interest in the eArchiving Building Block generated by our outreach activities was the great success of the first eArchiving Generic Services funding call, which received 13 applications, from 36 organisations. From these, four projects were eventually chosen for funding:

- *eArchiving of Engineering and Science Library*: The main objective of which is to adapt existing digital archives and repositories in the engineering and science domains to apply the Common Specifications for Information Packages established under the eArchiving building block;
- *J-Ark – European Jewish Community Archive*: This aims to deploy a community-driven approach to

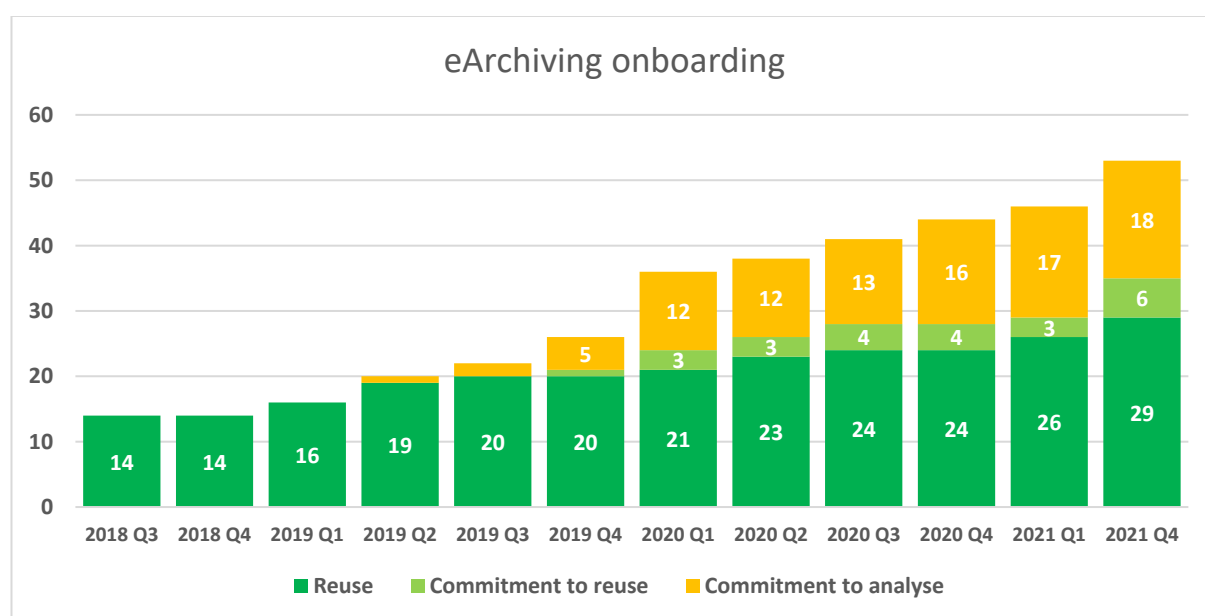


Figure 2 – eArchiving onboarding evolution (2018 to 2021)

⁷ <http://kc.dlmforum.eu/eark-products>

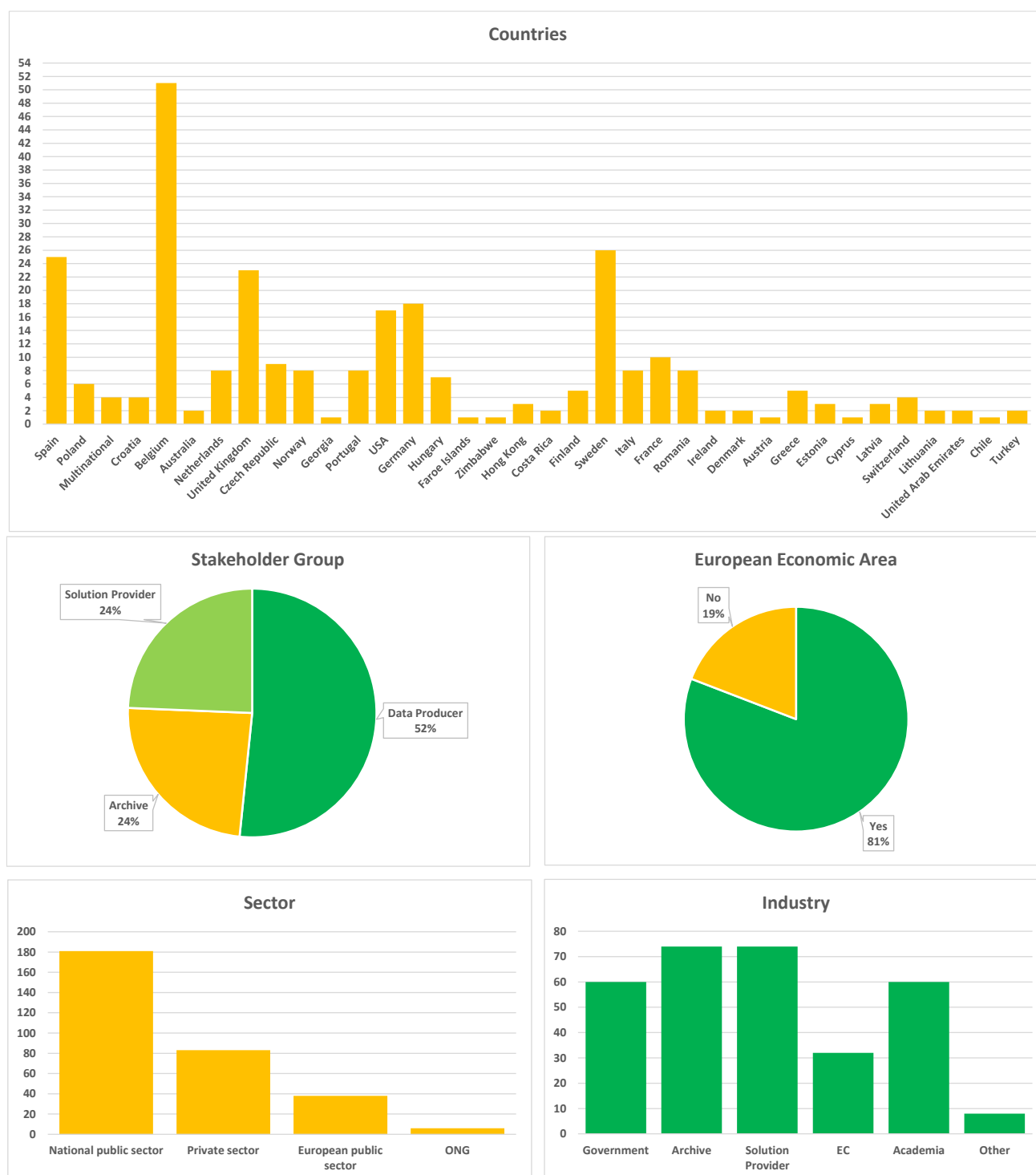


Figure 3 – Details on the CRM leads

the long-term integrity and accessibility of heritage materials in line with the specifications provided by the eArchiving Building Block. The European Jewish Community Archive (EJCA) will bridge the gap between Jewish community archives and technology providers with the means to provide (digital) infrastructural service, support, and advice;

- *Protecting Oral Histories Using Blockchain*: PROHUB will make use of the European Blockchain Services Infrastructure (EBSI) to implement a trusted data sharing approach together with the data archiving principles and standards provided by the eArchiving Building Block and the relevant E-ARK specification for information packages (CSIP);

- *One click eArchiving*: The goal of which is to build a simple to use one click solution which allows users to convert their existing CRM, CMS and ERMS exports into (an) E-ARK compatible SIP format(s) and to enrich/connect the metadata with existing data sources such as Europeana, Finna or other open data sources.

Finally, our outreach efforts also focused on providing case studies about organisations that adopted E-ARK. Five success stories were created during the CEF eArchiving period each focusing on a different sector:

- *The EU Publications Office (EUPO), (Luxembourg)*: EUPO has the mandate to ensure the long-term preservation of all official publications produced by the European Institutions. The number of such documents that need archiving is ever growing. As of August 2019, their system EUDOR v3 stores over 90 million files with texts dating back to 1951 – years before the European Commission existed. The E-ARK products helped to (1) prevent vendor lock-in and facilitate future migrations, which are needed in order to preserve the content and avoid format or support obsolescence, (2) establish a common language to communicate within the archival community, and (3) enhance interoperability;
- *Saint John Hospital (Portugal)*: Officials were acutely aware that clinical records in different formats and not always available in real-time made it harder to diagnose diseases and treat patients accordingly. The Clinical Records Repository project enabled healthcare professionals to access their 'patients' clinical history in a digital format;
- *Rotterdam Local Authorities (Netherlands)*: endow their existing digital archiving system with support for archiving databases by (1) setting a preservation plan for the new content, (2) selecting a preservation format, and (3) creating a service that would enable 'Rotterdam's citizens to easily access the 'city's archives';
- *State Archives (Italy)*: have been looking after the 'country's most important documents for the last 150 years. To make this precious heritage accessible to future generations, they were tasked with building a digital national platform. After looking for solutions both in and outside Italy, they found in eArchiving the answers they were looking for. They based their new system on

the E-ARK specifications because of the strength of the E-ARK underpinning models, which were precursors to the Reference Architecture. In fact, the Italian State Archives joined in the work on the Reference Architecture;

- *National Customs (Sweden)*: This agency oversees collecting custom duties and monitoring international traffic across the Swedish border. It is also responsible for facilitating commercial links between Sweden and non-EU countries, while stopping criminals from smuggling illegal goods in and out of the country. The goal was to digitally archive all its records.

III. E-ARK USE CASES AND BENEFITS

In this section we present the use cases and benefits reported from each of the E-ARK Consortium partners. Several organisations participated in E-ARK via the DLM Forum (DLM).

The Danish National Archives (DNA), the E-ARK Coordinator during the CEF phase, reported widespread benefits from E-ARK:

- a) cooperation with the Consortium and the EC, helping to promote eArchiving as EC policy;
- b) adopting the specifications for the IPs – CSIP, with the advantages of scalability (segmentation), and the CITS for databases and the CITS for geospatial data;
- c) using the Reference Architecture for eArchiving;
- d) using the sample tools for creation, validation and presentation.

The DNA have been a consortium member and field leader in database archiving from the beginning of the E-ARK effort.

It is important to note that for a National Archive it can take many years from the start of the collaboration until the adoption of new standards and deployment of new software tools. This is due to the impact on public administration, and especially in the case of the DNA which, according to Danish law, can mandate that all public institutions (ministries, agencies, courts, counties etc) must use the DNA specifications, thereby covering 99% of all public institutions.

Poliphon is a Hungarian consulting and solution development company, participating mostly in business process and document management projects. They are used to modelling and automating business processes specific to one enterprise. In the E-ARK and eArchiving projects where they were

involved via DLM, they gained another view on processes. Here they looked for the common not the specific and learnt that interoperability between organisations really means interoperability between the processes run by these organisations. But organisations cannot be compelled to adopt your processes, and organisations remain different in many ways even in the area of interoperability. This, of course, is why you need the specifications. They are the common fixed points where different processes can meet and around which processes should be built in order to achieve interoperability. This is why specifications are E-ARK's main assets, and processes are the means to harness the interoperability opportunities of the specifications.

The Archives of the Republic of Slovenia (ARS) stated that eArchiving was needed to make their service better and they achieved this by working with the Slovenian Cancer Registry to develop an eHealth SIP which could then be used across the European Union (EU). ARS have been using the database tools for several years now, and the geospatial CITS and the Reference Architecture work also meet the needs of producers and stakeholders that ARS is responsible for, so it helped ARS to develop a better service for them. E-ARK was also one of the forms which ARS could use during the pandemic to exchange information and also answer some questions.

The National Archives of Finland (NAF) joined the E-ARK consortium in 2019 via DLM and they were able to benchmark their practices in the areas of Database preservation and Geoinformation where they carried out case studies. They benefitted from the Reference architecture work and found that E-ARK participation had broadened their ways of thinking and forged a transition from a local to a European way to manage information. Networking with professionals, NAF gained a huge amount of information and top-level skills around digital preservation possibilities, including to develop their next generation of services.

NAF experienced a quick start with open-source tools, support testing, and adapting practices without having to carry out their own software development. The E-ARK training and learning provided NAF with food for thought. They received encouragement to work with multinational teams and to change ideas for the common good. This led to opening eArchiving possibilities for a broader audience and a better understanding of the CEF Building blocks.

KEEP Solutions, a Portuguese SME, participated in E-ARK since its early stages and were able to cross all stages of evolution from research and innovation, up to commercial exploitation. They stated that their clients could be ensured that KEEP software is compliant with international specifications and best practices as these are backed up and validated by an EU Building Block. Working with large international organisations allowed KEEP to expand their knowledge, understand different realities and develop their tools to cope with more advanced use cases. The Building Block gave them international visibility, which was key in reaching certain markets and certain types of international clients.

In support of the wider DLM Forum mission, DLM's involvement with E-ARK gave them an opportunity to play a leading role developing Open-Source tools and services for the Digital Preservation Community. They facilitated the participation of many of their member organisations, who would not otherwise have been able to take part. They were able to contribute to community knowledge by supplementing the training programme initiated within E-ARK with their own series of related webinars. As a result of their participation in E-ARK, DLM has increased its membership, broadening the community for the benefit of all.

Kommunalförbundet Sydarkivera, a Southern Sweden municipality association acting as the archive authority for its members, contributed to E-ARK via DLM, and through their participation in the E-ARK webinars, their knowledge and outreach have been extended. The tools developed within E-ARK have proved to be valuable to their members in helping them create SIPs. Two concrete examples of this are the eHealth1 specification used for the creation of SIPs with patient medical journals, and the facilities provided by the update of SIARD and the creation of CITS.

The Swiss Federal Archives (SFA) participated in E-ARK via DLM and their collaboration with the E-ARK projects has contributed substantially to the development of the SIARD Format and Suite 2.2. It has provided valuable opportunities for knowledge exchange, mutual support, commitment to success, and inspiration.

Gabinete UMBUS SL, a Spanish independent consultancy in information and records management stated that E-ARK has given them the chance to work in a multinational environment and has shown them a holistic approach to digital

preservation that can be applied to their projects. A clear benefit in working with people from different “archival environments” is that they have been able to share different approaches with their client base. Their work on the eArchiving Building Block has opened up for them a new range of opportunities and has helped them develop their business.

Georh, a Slovenian SME said that E-ARK had a significant effect on their company, enabling them to identify and exploit a whole new market. Through their collaboration, they have learned the needs of a new customer base, and they have gained both the experience and the knowledge to allow them to win new business.

Highbury R&D, stated that coordinating E-ARK with the Danish National Archives allowed them to network with many people across the Digital Preservation (DP) domain. As they are now based in Ireland they can continue to play a leading role in E-ARK. Concrete examples of the benefits they have enjoyed are being able to broaden their base of operations into areas the company was not previously working in, such as eHealth. They were also able to bid for the Generic Services project and were successful – this brought them into the new area of using eArchiving and Blockchain to preserve sensitive oral testimonies. For Highbury, leading training has been an important experience, and will continue to be a central part of their portfolio in the future.

The South-Eastern Finland University of Applied Sciences (XAMK) said their involvement with E-ARK has expanded significantly not only their EU-wide connections but has also opened up contacts in the UK and USA. Directing following from exposure gained on E-ARK3, XAMK now has an elected representative on the Executive Committee of the DLM Forum. Before their involvement in E-ARK3, they were well known within Finland, but now they have established themselves on the wider European stage. Finally, they have followed up their work in E-ARK3 and are now leading the HADEA-funded OneClick eArchiving Generic Services project.

The National Archives of Estonia (NAE) has been continuously involved with E-ARK since the establishment of the very first E-ARK consortium. They clearly recognise the relationship between the number of participants in developing digital preservation solutions and the quality of the final output. E-ARK has enabled them to work as part of a much larger and broadly-based group and achieve

more than would have been possible otherwise. E-ARK has developed specifications and software that can be deployed in Estonia, without requiring these to be developed in-house. For them, the most useful output of E-ARK is the standardisation of database archiving: the solution currently deployed by NAE is around 90% dependent on E-ARK knowledge, and connections.

The National Archives of Norway (NAN) stated that working as part of the E-ARK consortium has given them access to different groups where they have been able to discuss archive challenges for the future, and how to solve them. They have found, in particular, that the Reference Architecture is a good framework for international cooperation and common understanding across organisations, and that SIARD has proven very useful for us. The principles developed within E-ARK represent very useful guidance for archives working in the digital age.

Easy Lean OÜ, an Estonian SME worked on E-ARK via DLM and gained knowledge and professional experience of the pre-ingest and IP specifications. As a direct result of this, ERMS content is now better organised and prepared. E-ARK exposed them to a wide range of contacts and ideas which led them to bid for a new project OneClick eArchiving which combines the benefit from existing E-ARK specifications and tools. E-ARK also provided inspiration to develop PhD work, in particular on how technological development and changed processes influence the creation and evolution of preserved digital information.

The Technical University of Lisbon, Portugal (INESC-ID), said that E-ARK has given them a deep understanding of the Digital archiving community. It has helped their research data community (RDA and EOSC) engage with the Digital Archiving community. It has enabled them to participate in ISO committees on topics addressed within the project (Enterprise Architecture and Business Process Management). Furthermore, it has opened up further funding opportunities.

The Austrian Institute of Technology, Vienna (AIT) stated that working on E-ARK has helped them learn more about the archiving needs of companies and organisations. As a result, they now have a better understanding of how the CEF building blocks can help them to work more closely together at a European level. They bid successfully for a Generic Services project “PROHUB” through which they are making use of the eArchiving and Blockchain building

blocks. They have expanded their network to now include the archiving community.

PIQL, a Norwegian SME, said that in E-ARK they learned 'best practice' in writing (eArchiving) specifications, which they regard as an important expansion to their skill set in producing a well-documented specification. They also gained greater depth of knowledge in eHealth and gained experience of running a very focused software development project with very specific outcomes for low cost and in limited time. Their experience in E-ARK put them in a position to bid successfully for two follow-on projects: Science and Engineering and OneClick.

The Open Preservation Foundation (OPF) stated that participation in E-ARK has enabled them to develop a deeper understanding of the needs of their archive members. It has allowed them to forge connections with a community focused on addressing interoperability between archival systems and organisations, and it has given them the chance to work on Free and Open-Source Software (FOSS) [7] for information package validation. This, in turn, has allowed them to improve METS validation in general. The E-ARK validator has proven flexible enough to enforce other Metadata Encoding and Transmission Standard (METS) profiles [8] with minimal development, and they see this as a major advantage.

ES Solutions (ESS), a Swedish SME has increased their knowledge of international conditions regarding digital preservation and gained valuable experience through various collaborations with expertise within the E-ARK projects. They have connected to a very valuable network with expertise in various areas within digital information management. They have attracted the attention of others internationally, which has led to a developing exposure to other markets. Their software portfolio has matured through the E-ARK projects and now provides through its comprehensive functionality an overall E2E solution for digital preservation.

IV. CONCLUSIONS

The E-ARK offerings have now come to maturity and are being deployed across Europe in many different ways and extents. There are many more use cases and success stories than are reported here where we have just concentrated on how the E-ARK Consortium members have used and benefitted from the E-ARK outcomes. Working as part of an EC

CEF Building Block has brought real benefits to the E-ARK Consortium members but what is noteworthy is just how long it can take for large organisations to scope out, design, implement then deploy a digital archive (see the DNA experience above). Another hindrance is the lack of EU legislation on digital archiving: even though E-ARK products can be used across different national legislatures, there is still the tendency for countries to go it alone.

ACKNOWLEDGEMENT

The authors would like to thank all the E-ARK consortium members who provided the material for this paper: Anders Bo Nielsen (DNA); Kuldar Aas (NAE); Karin Bredenberg (Sydarkivera); Miguel Umlauff and Carlota Bustelo (Gabinete); Jože Škofljanec and Anja Paulič (ARS); Audun Lund and Krystyna Ohnesorge (SFA); Gregor Završnik (Geoarh); Markus Merenmies (NAF); Miguel Ferreira (KEEP); Anssi Jääskeläinen (XAMK); Kristin Jacobsen (NAN); Karin Oolu (Easy Lean); Sven Schlarb (AIT); Stephen Mackey and Bendik Bryde (PIQL); Becky McGuinness (OPF), and Björn Skog (ESS). We also acknowledge the contribution of our EC colleagues Fulgencio Sanmartín and Adelina Dinu from DG CNECT and Pawel Stech and Tom Fillis from DIGIT: it has been a huge privilege, a pleasure and great experience working with all the folks in the SMO and in DG CNECT.

REFERENCES

- [1] Archiver project. <https://www.archiver-project.eu/>, accessed 3rd March 2022.
- [2] DILCIS Board. <https://dilcis.eu/>, accessed 3rd March 2022.
- [3] OAIS. <https://public.ccsds.org/pubs/650x0m2.pdf>, accessed 4th March 2022.
- [4] PREMIS. <https://www.loc.gov/standards/premis/>, accessed 4th March 2022.
- [5] DLM Knowledge Centre. <http://kc.dlmforum.eu/home>, accessed 4th March 2022.
- [6] E-ARK Validator. <https://www.itb.ec.europa.eu/cef/itb>, accessed 4th March 2022.
- [7] FOSS. <https://itsfoss.com/what-is-foss/>, accessed 4th March 2022.
- [8] METS. <https://www.loc.gov/standards/mets/METSOverview.v2.html>, accessed 4th March 2022.

CONSTRUCTION OF A BENCHMARK MODEL FOR EVALUATING ACADEMIC INFORMATION ON SOCIAL MEDIA WORTHY OF LONG-TERM PRESERVATION

Liu Hui

*Department of Information
Management of Peking University
National Science Library of CAS
China
liuhui_22@stu.pku.edu.cn
[0000-0003-3182-8372](tel:0000-0003-3182-8372)*

Zhang Dongrong¹

*National Science Library of CAS
University of Chinese Academy of
Sciences
China
zhangdr@mail.las.ac.cn
[0000-0002-0745-3681](tel:0000-0002-0745-3681)*

Abstract – There is a wealth of academic information accumulated on social media, which has not been discussed in long-term preservation practices and research at home and abroad. In order to provide forethought for the future development of long-term preservation, this paper attempts to discuss how to evaluate academic information on social media worthy of long-term preservation. Based on the theoretically analysis of the characteristics and preservation value of academic information on social media, this paper proposes a evaluation index system based on meta-synthesis method. The next step is to invite experts to make judgments and propose amendments to this evaluation index system.

Keywords – academic information; social media; evaluation indicators; long-term preservation; preservation value

Conference Topics – Resilience; Exchange

I. INTRODUCTION

With the advent of the new media era, social media has become an important way for many scholars to access academic materials, share academic achievements, conduct academic exchanges and innovate academic research, as well as a platform for the dissemination of scientific knowledge.^[1] As a result, scientific researchers in every discipline spread a wealth of information across social media platforms. Academics and the public should be able to access such information data not only in the present but also in the future. But the vulnerability of social media data hinders

long-term sustainable access to this type of information. Preserving academic information on social media is urgent and necessary. Relevant parties should take action as soon as possible.^[2] In addition, the rapid growth of academic information on social media has been accompanied by information overload, information noise, misinformation, and other problems, making it difficult to guarantee the accuracy and reliability of information. This not only challenges the ability of scientific researchers to judge and use information, but also makes the long-term preservation scope difficult to define. Therefore, scientific evaluation of academic information on social media is of research significance. In order to achieve this goal, this paper puts forward a comprehensive preservation value evaluation index system by using meta-synthesis method, thus to help the preservation organization to judge the value of information and determine the scope of preservation, and also provide a reference for the social media platform to strengthen information management.

II. LITERATURE RESEARCH

At present, more research has been done on how to evaluate the value or credibility of social media information. For instance, P. André et al.^[3] evaluated the value of Twitter's content through a web survey and analyzed what kind of information was usually valuable. With the flourishing of academic exchanges

¹ Corresponding Author

on social media, especially the widespread use of academic social networks, academics have also conducted research about how to evaluate the credibility or quality of academic information on academic social networks. E.g. Wang Jie^[4] built a quality evaluation system for academic public account information on WeChat. In fact, initially the relevant research focused on the evaluating from information internal attributes, and with the further development of research, the dimensions of information external attributes and platform functions become important components of the evaluation. E.g. Bi Liping et al. ^[5] constructed an evaluation system of the WeChat public platform of academic journals based on the three evaluation dimensions of form, content and utility in the "full evaluation" analysis framework. In addition, more and more researchers took user needs, experience and behavior as the key focus of evaluation. E.g. Zhang Ning and Yuan Qinjian ^[6] constructed a CPUC model of the influencing factors of academic social network information quality from the perspective of user perception. In summary, (1) The purpose of most researches is to optimize the operation of social media and improve users' information judgment skills. No studies have discussed evaluation issues around the delineation of long-term preservation based on an perspective of information resource management. (2) Most of the studies are aimed at a certain social media platform, lacking a macroscopic and comprehensive evaluation perspective. And it has become a trend to expand social media information evaluation to a

multi-angle and multi-dimensional generalized evaluation. Therefore, we plan to conduct theoretical research on the systematic understanding and integration of existing research results, and build a scientific and comprehensive evaluation benchmark model to guide the development of long-term preservation of social media academic information.

III. CHARACTERISTICS OF SOCIAL MEDIA ACADEMIC INFORMATION

Social media academic information refers to the information content produced by the academic community by the symbol system of scientific context. Social media platforms have changed the mode and blurred the boundaries of academic communication. The information generated by traditional formal academic communication ways also circulates on social media (uploading, creating, disseminating, using, etc.). So compared with information generated by traditional academic exchanges, social media academic information performs many unique features in many aspects, as shown in Table I. In addition, social media platforms themselves have a large impact on the characteristics of academic information. Due to the various types and different functions of social media, as well as problems such as ease of control, preference, strong domain, strong regionality, and fragmentation of communication in use, social media academic information also appears strong source, structure, tenure, type, privacy, and quality complexity.^[7]

Table I
Traditional VS Social Media Academic Information Characteristics

Difference	Traditional academic information	Social media academic information
Production aim	Expand academic communication and influence, and promote scientific research cooperation, etc.	Expand academic communication and influence, and promote scientific research cooperation, etc.
Academic value	More systematic, logical and repeatable	More inspiring, divergent and fragmented
Credibility	(To be) Peer-reviewed, high credibility	Not peer-reviewed, credibility unstable
Stability	Stable, not easy to fade and change	Unstable, easy to fade and change
Originality	Overall higher	Overall lower
Spread effect	Low efficiency, narrow range	High efficiency, wide range
Audience	Academia, elite	Academia and society, democratization
Publication cycle	Long period, fixed frequency	Short period, variable frequency, strong timeliness
Presentation form	Single, structured,	flexible form, better user experience
Copyright Protection	perfect	imperfect
Interaction	Authors, publishers and readers cannot directly interact and communicate, and feedback is poor.	Authors, publishers and readers can directly interact with readers, and the feedback is flexible.
Release channel	Specialized publishing institutions, academic conferences, institutional knowledge bases, etc.	Rely on social media platforms
Acquisition cost	Higher	Lower
Organization	There is a systematic classification system and organization method	There is no systematic classification system and organization method
Storage method	Mature	Immature

IV. PRESERVATION VALUE OF SOCIAL MEDIA ACADEMIC INFORMATION

4.1 Theoretical Analysis of Value

The process of preserving content selection is essentially a process of value selection.^[8] To fully understand and explore the preservation value of social media academic information, this paper drew on a mature theory of value discovery in Chinese Archives Science-"archive dual value theory"^[9], and analyzed the preservation value of social media academic information from both "content value" and "tool value", shown as Fig. 1.

(1) *Content value*. As an object entity, social media academic information is a kind of information, and also an asset. It has rich content value, including academic value, social value, economic value, and cultural value.

(2) *Tool value*. It refers to that social media academic information resources, as content carriers, can continue to exert their content value for a long time in the digital preservation process. Although tool value is determined by the purpose of different subjects, it generally includes: 1) Realize resource enhancement and enrich information assets. 2) Enhance service capabilities and develop business areas. 3) Benefit future research. 4) Facilitate academic evaluation.

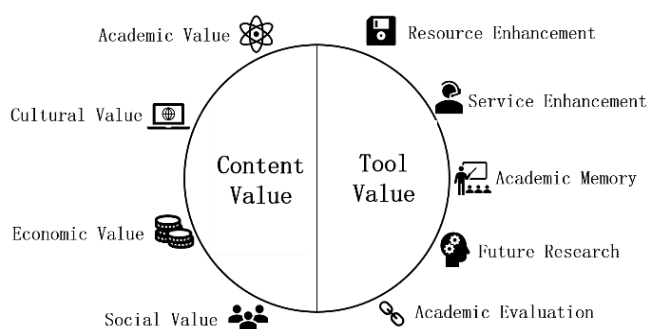


Figure 1 Framework of dual preservation value of social media academic information.

4.2 Realistic Problem of Value Appraisal

Value appraisal is an indispensable link and operational basis in the digital preservation.^[10] Social media academic information obtains rich preservation value, but when the theoretical analysis is applied to practical activities, it faces many problems: (1) The complexity of academic information on social media makes it more difficult to identify the value of information and delineate the scope of preservation. (2) Preservation of value is based on the relationship between subject and

object. The diverse demands and vague demands of different subjects for objects make it more complicated to delineate a reasonable range of preservation.^[11] (3) Time range of preservation is very long, and the present value of information is not completely equal to the importance of the future. The potential of value changing over time poses challenges to determine the current selection range.

In summary, it is not easy to identify the preservation value of academic information on social media, and there are problems such as poor pertinence and applicability by relying on the existing identification theories and principles with a certain ambiguity, subjectivity, and weak operability. Therefore, before considering the implementation of preservation, the determination of information value must be achieved in a more efficient, operational, and standardized way. According to literature research, many papers have researched how to evaluate the quality, credibility, importance, and user satisfaction of social media information.

V. METHODS

In order to comprehensively synthesize the achievements accumulated and discover the potential consensus on the design of indicators in related research fields, this study uses the meta-synthesis method.^[4] This method provides a content analysis system method, which identifies new metaphors and constructs new theories or models by comparing, explaining and combining various existing frameworks, and possesses great potential applications in scientific evaluation.^[12] Due to the complex characteristics of social media academic information, its evaluation involves multi-dimensional concepts, therefore the meta-synthesis method is an appropriate method to fully integrate the existing evaluation index system with the requirements of preservation practice, providing a macro picture of the research object and ensuring a higher promotion in evidence-based research. There are four steps:

Step 1: Select a Collection of Papers

The key to this step is to ensure a systematic and comprehensive literature search, as well as standardized and relevant literature selection. In this study, CNKI and Web of Science were selected as the main search sources, and Google Scholar was used as the supplementary search source. A series of combined Chinese and English keywords were

considered. Then three rounds of retrieval were conducted. The first round was to retrieve literature related to “social network academic information evaluation/ assessment”, focused on 2017-2021. The second round was to retrieve literature related to “digital preservation value evaluation”. Because the search results are too few, there is no time limit for publication. In the third round, the references of literature found in the above two rounds were searched, making sure no relevant literature was missed. The specific screening procedure is shown in Fig. 2. In the end, 30 pieces of related literature were obtained, including 23 Chinese literature and 7 English literature; 24 journal papers, 5 dissertations and 1 conference paper. The following principles were obeyed when browsing and judging whether a document is selected and reviewed: (1) The subject is relevant; (2) The proposed indicator system has at least two layers of structure, and each indicator is clearly explained. (3) There is a certain theoretical basis. (4) Not a single research method was used.

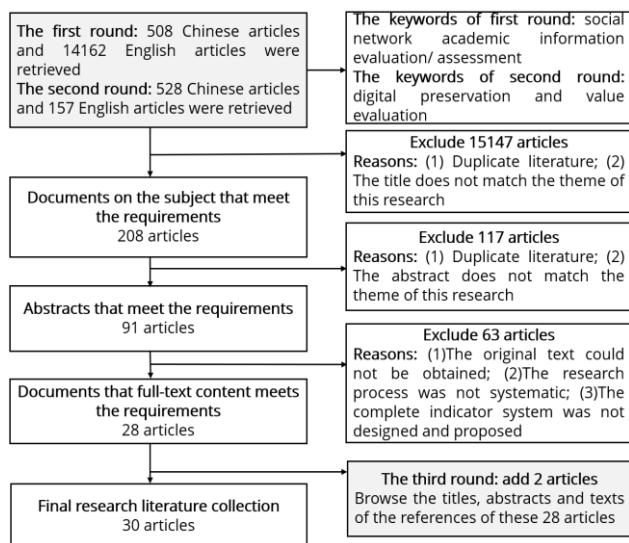


Figure 2 The selection process of the literature collection

Step 2: Extract indicators

The 30 papers are divided into four categories according to their themes: online academic information evaluation (9), social media academic information evaluation (9), social media information evaluation (6) and long-term preservation value judgment (6). The four topics are represented by code 'WAI', 'SMAI', 'SMI', and 'DPV', encoding each document by topic. Taking 'social media information evaluation' as an example, the label of each document under this topic is SMI1-6. Then the research team read each literature and extracted its designed evaluation system, including all dimensions

and specific indicators, while retaining the original text. As shown in Table II, a total of 113 dimensions, 405 first-level indicators and 188 second-level indicators were finally extracted.

Table II
Examples of Metric Extraction

Code	Author	Publication time	Indicator
WAI 9	Soung, Sereywa thna ^[13]	2017	<i>Relevance of information</i> : the level of information, the adequacy, the graphic support, the methodology indicated <i>Source reliability</i> : the physical existence of the source, the appearance of the site <i>Reputation of the author</i> : affiliation, expertise, frequency of citation, biographical information <i>Content quality</i> : objectivity, accuracy, timeliness
WAI 6	Keshavarz H, Givi M E, Norouzi Y ^[14]	2018	<i>Credibility</i> : personal information, objectivity, morality, writing style, website appearance, website management, website identification; <i>Professional knowledge</i> : professional information, coverage, resource availability, interaction, accuracy
WAI 2	Liu Bing, Jiang Xiaohan ^[15]	2019	<i>Content characteristics</i> : information comprehensiveness, accuracy, novelty, etc. <i>External characteristics</i> : authority of information source, form of information expression, timeliness, etc.
SMAI 2	Bi Liping et al ^[5]	2020	<i>Form</i> : recognizability, trustworthiness, interactivity; <i>Content</i> : usefulness, ease of use, friendliness <i>Utility</i> : overall communication power, average communication power, headline communication power, peak communication power
SMI 8	Keshavarz, Hamid ^[16]	2020	<i>Information source</i> : User profile, Authority <i>Information presentation</i> : Content, Links, Layout, Writing

			Information credibility: Objectivity, Currency, Accuracy, Usability Decision related: Risks, Benefits, Trust, Organizational issues
--	--	--	--

Step 3: Code and Integrate

This step integrated and encoded the obtained 113 dimensions and 593 indicators to prepare for the next step to extract the key dimensions and indicators required for this study. After discussing the division of labor within the team, Liu independently integrated and encoded the dimensions and indicators according to the original expression in the literature. Zhang checked and raised objections. The team discussed these issues, consulted other experts, and then got the final results. When integrating and coding, we found that there are three notable relationships between the indicators in each literature: (1) Different expressions but similar meaning; (2) Similar expressions but different meanings; (3) Different expressions and meanings but belong to the same dimension. Therefore, three meta-synthesis methods, namely Reciprocal translational analysis, Refutational synthesis and Lines of argument synthesis, were used to combine, select, classify or delete dimensions indicators. In the above process, the hierarchical relationship between dimensions and indicators were preserved. Finally, 8 dimensions and 81 indicators were integrated, and the frequency of them was counted, as shown in table III & IV.

Table III
Dimension Integration and Coding

Code	Dimension	Source	Frequency	Number of Indicators
D1	Content	SMAI 2, 4, 6; WAI 1, 3, 4, 5, 7, 8; SMI 1, 2, 4, 5, 6, 7; DPV 1, 2, 3, 4, 5, 6	21	20
D2	Function	SMAI 4, 5; WAI 1, 3, 4, 7; SMI 1, 4, 5, 6; DPV 4, 6	13	16
D3	Usability	SMAI 2, 4, 5, 6; WAI 2, 5; SMI 2; DPV 5, 6	9	13
D4	Source	SMAI 3; WAI 6, 7, 8; SMI 2, 3, 5, 6, 8, 9; DPV 1, 4, 6	20	10

D5	Form	SMAI 2, 4; SMI 8, 9; DPV 1, 2, 3; WAI 1, 2, 3, 4; DPV 5	12	9
D6	User	SMAI 1, 4, 5, 6; WAI 1, 3, 4, 5; SMI 1, 3, 4, 5, 6, 7; DPV 1, 2	18	7
D7	Service	SMAI 1; WAI 2, 4	3	6
D8	Environment	SMI 4, WAI 3	2	

Each dimension appears ≥ 1 times in a single article.

Table IV
Examples of Indicator after Coding and Integrating

Code	Indicator	Source	Frequency
C1	Timeline	SMAI 3, 4, 5; WAI 1, 2, 5, 7, 8, 9; SMI 2, 4, 5, 6, 8; DPV 1, 2, 4	18
C2	Accuracy	SMAI 6; WAI 1, 2, 3, 4, 5, 6, 7, 8, 9; SMI 2, 5, 7, 8	15
C3	Relevance	SMAI 3, 4; WAI 1, 4, 7; SMI 1, 2, 5, 7; DPV 3, 4, 6	13
C4	Objectivity	SMAI 3; WAI 1, 6, 8, 9; SMI 1, 2, 7, 8	13
C5	Innovation	SMAI 3, 4; WAI 1, 2, 4, 5; DPV 2	11
S1	Functional ease of use	SMAI 2, 4; WAI 1, 2, 4, 5, 7; SMI 5	10
S2	Interface friendliness	SMAI 2, 3, 4; DPV 4; WAI 1, 3, 4, 5; SMI 4	9
S3	Page integrity	SMAI 3; WAI 1, 3; SMI 5; DPV 1, 2, 3, 4	8
S4	Information security	WAI 1, 3, 4; SMI 5, 6; DPV 2, 4	7
S5	Interface	WAI 1, 2, 8, 4; SMAI 2, 4	6

	interaction		
F1	Presentation form	DPV 1、2、6; SMI 2、4、6; WAI 2	7
F2	Writing style	WAI 1、6、8; SMI 8	4
F3	Content classification	WAI 3; DPV1、4	3
F4	Format specification	WAI 1、3	2

C means content, S means system, F means form.

Step 4 : Focus Key Indicators

According to the results obtained in the third step, it was found that the evaluation dimensions of most literature involve information attributes, information sources, system platforms, user experience and information environment. According to the information ecosystem theory, information attributes point to the information object; information source, media platform and user experience belong to producers, managers and consumers of information subject; information

environment refers to the environmental factors in the ecosystem. Therefore, this study used elements in the information ecosystem to determine the evaluation dimension.^[17] The design of indicators often depends on the purpose of evaluation. Therefore, when focusing key indicators according to the key dimensions, we not only selected them according to their frequency, but also combined the work requirements of long-term preservation. Finally, 5 dimensions, 12 first-level indicators and 44 second-level indicators were extracted, as shown in table v.

VI. RESULTS

As shown in Table v, this paper constructed a social media academic information evaluation index system with a hierarchical structure, which better unifies the public evaluation, expert evaluation and market evaluation. In addition, the layers of this system are named as dimension layer, object layer and measure layer. The first and the second layer are the core layers, and the third layer is the optional layer, that is to say, before the application of this evaluation system, the preservation subject must analyze its applicability and operability based on the specific situation, and adjust the third layer index.^[18] The connotation of each dimension and indicator is explained below.

Table V
Interpretation of Key Indicators

Dimension	Primary indicator	Secondary indicator	Interpretation
Information object	Content feature	Academic	The Strictness, precision and standardization of theoretical knowledge and method application. It focuses on measuring the significance and role of information for the development of human academic careers.
		Timeliness	The half-life of information value.
		Objectivity	Whether it is an objective statement.
		Innovation	Originality, inspiration.
		Authenticity	Whether it is based on facts, whether it cites literatures, and whether the data source is supported.
		Originality	Whether it is not reproduced, whether it possesses copyright.
		Sensitivity	Whether it involves user privacy.
		Organization	Integrity, logicity and clarity.
		Frontier	Whether it reflects research progress in frontier areas.
		Professionalism	The depth and pertinence of related professional fields.
		Digital native	Whether it is digitally native, and whether there is corresponding paper data.
	Formal feature	Presentation friendliness	Whether it is graphic and clear layout.
		Language specification	Whether there are spelling and grammar errors.
		Writing style	Expression tendency and wording characteristics.

		Content classification	Information topics and categories. To select classification method according to demand, such as information scene, communication channel, information nature, information form, property ownership, etc.
		File format	Format for file types such as text, images, videos, hyperlinks, etc.
Information publisher	Publisher property	Publisher type	-Divided into individual and institutional accounts. -Divided into academic creators, academic publishers, academic service providers and academic media organizations by role and function
		publisher identity	Background resume, reflecting its affiliation, status, ability, etc.
	Publisher influence	number of fans	Size of audience.
		communication power	Attention, including the number of clicks, forwards, comments, etc.
	Publisher credibility	account level	Operation time-periods of accounts
		official certification	audit certification by the platform or the third-party organization
		profile completeness	The completeness of the account profile, including functions, positioning, etc.
	Publisher activity	Publishing frequency	The number of releases in a certain period.
		Publishing amount	Total number of releases since account opening.
		Interaction degree	The frequency of the publisher's response to comments.
Target user	User utility	Absolute utility metrics	reads/plays, retweets, comments, likes, etc.
		Relative utility indicators	Praise click ratio, forward click ratio, comment and click ratio, etc.
	User characteristic	User attributes	Personal traits (gender, age, education, occupation, etc.), habit preferences (retrieval, use of information, etc.), knowledge background (professional field, information literacy, media literacy, etc.)
		User motivation	he urgency and pertinence of users' information needs.
Media platform	System function	Information security	Whether it contains unsafe or illegal links.
		System stability	The reliability of the hardware and software system. To ensure that the information on it can be accessed normally at any time.
		Functional adequacy	Whether the platform function module design is comprehensive and fast, level-clear, concise and clear.
		Interface interactivity	Whether the page can be displayed stably after clicking the link, whether there are empty links, dead links, etc.
		Response timeliness	The degree of interaction between user and platform.
	Platform Policy	Intellectual property protection policy	The principle of intellectual property rights in the process of data utilization and preservation formulated by the platform.
		Quality control policy	Include editorial review system, qualification review system and peer review system.
		User privacy protection policy	Rules on obligations, rights and responsibilities for user privacy protection.
		Information security policy	A series of rules for deleting, auditing, hiding, and preventing user-generated information
		Data open access right	Regulations on the authority and scope of third parties to obtain and use data.
Information environment	External environment	Laws and regulations	Specifications and constraints on the preservation subject and preservation behavior; limitations and exceptions for the utilization of preservation objects.
		Policy system	The requirements and support of governments to ensure that long-term preservation activities are carried out in a standardized and orderly manner.
	Internal environment	Storage condition	Whether sufficient storage space and storage equipment is realized.
		Technical condition	Whether mature tools are mastered, which possess capture, save and exploit capabilities.

6.1 Information object The internal and external attributes of information itself are the most important factors in judging its value: (1) *Content features*. It goes deep into the core of the evaluation object, and often relies on peer experts to judge the value and quality of information content through several qualitative indicators. (2) *Formal features*. It refers to the external representation of information, including language style, format, and theme. It directly or indirectly affects the efficiency of user use, and it also affects the sustainability and feasibility of long-term preservation.^[19]

6.2 Information publisher In the network environment, information publishers have a significant impact on information credibility, and it is the directly available heuristic clues for the preservation subject to judge the credibility of the information and decide whether or not to choose to preserve it. Generally speaking, if the reliability is high, so is its preservation probability.^[20]

6.3 Target user This dimension is to evaluate the value of information from the user's information needs or expectations^[21]: (1) *Perceived utility*: It measures users' subjective perception of the inherent characteristics of the acquired information, that is, to reflect the social, economic and cultural benefits of the information through user satisfaction. (2) *User characteristics*: The needs of users are differentiated and divergent, and preservation institutions should fully consider the characteristics of target service groups when selecting digital resources to be preserved.^[10]

6.4 Media platform The system platform on which information depends is inseparable from the information itself, and it is a technical and physical factor affecting the long-term preservation value of information. Whether the media platform is safe and stable is the basis for ensuring the authenticity, reliability and integrity of social media academic information.^[22] In addition, the management policies of the platform itself also affect the feasibility and scope of preservation, meanwhile will limit consumer application.

6.5 Information environment This dimension measures the underlying support conditions for long-term preservation activities and evaluates the external and internal environmental factors that affect the storage, reading and utilization of information. The external environment mainly includes relevant laws, regulations and policy

systems, requiring preservation work to be carried out within the framework of relevant laws and policies of the country; the internal environment mainly includes storage conditions and technical conditions, requiring the preservation subject to assess whether they have the conditions to achieve the sustainability of preservation activities.

VII. DISCUSSION

To achieve a scientific, comprehensive and systematic evaluation of long-term preservation value, this paper used meta-synthesis method to construct a multi-dimensional, multi-level and systematic evaluation system, which presents the following characteristics: (1) It focuses on measuring the academic value and attributes of preserved objects; (2) It considers the characteristics of social media platforms, and the source, structure, ownership, type, privacy and quality complexity of preserved objects; (3) It has strong applicability for different preservation subjects and purposes. However, this method is a heuristic qualitative analysis method, and the research conclusions are limited by the quality of the analysis text and the limitations of the researcher's knowledge. Therefore, we plan to invite 6 experts in relevant fields to identify problems and rank indicators through scoring. According to the evaluation system, this paper uses the Likert scale to design the expert review table, which includes two aspects: (1) *To judge the necessity of each indicator, and design a 5-level rating scale* (5 is very necessary; 4 is necessary; 3 is general; 2 is not necessary; 1 is not necessary); (2) *To put forward suggestions for modifying this system*. The Likert scale is used to evaluate the necessity of each indicator, so the indicators with an average value of more than 3.5 and a standard deviation of less than 1 are regarded as consensus standards.

ACKNOWLEDGMENT

In the writing process, I am very thankful to the young scholar Duan Meizhen for her suggestions and pointers on the framework, content and research method design of this paper. Thanks to Mr. Qi Zheng for her modification suggestions on indicators integration and system design. In addition, I would like to thank the experts who participated in reviewing indicators. With the joint efforts of everyone, this paper has been being continuously improved.

REFERENCES

- [1] Li Yujia, Zhang Keyong. "Research on the characteristics and process of academic new media information dissemination in the context of mobile Internet," *Library Work and Study*, no. 6, pp. 81-86, June 2017.
- [2] Huang Xiaoyu, Qian Hongmei. "Is Your Online Memory Safe: Reflections on the Ownership of Social Media Documents," *China Archives*, no. 4, pp. 68-69, April 2014.
- [3] André, Paul & Bernstein, Michael & Luther, Kurt., "Who Gives A Tweet? Evaluating Microblog Content Value," *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 2012, pp. 471-474.
- [4] Wang Jie. *Study of Assessment System of Academic WeChat Public Number Information Quality*. 2018. Hebei University, MA thesis.
- [5] Bi Li-ping, LIAO Shu-yu, LI Zhan, et al. "The Evaluation System of Academic Journal WeChat Public Platform Based on the All-round Evaluation System," *Information Science*, vol. 38, no. 2, pp. 156-162, February 2020.
- [6] Zhang Ning, Yuan Qinjian. "Governance and Improvement of Information Quality on Academic Social Networking Sites," *Documentation, Information & Knowledge*, no. 6, pp. 105-113, October 2018.
- [7] Wang Yuefen, Jia Xinlu, Li Dongqiong. "A Study on the Influencing Factors of WeChat Academic Information Sharing Intention," *Library & Information*, no. 3, pp. 9-18, June 2020.
- [8] Xing Bianbian, Sun Dadong. "Research on Modes and Methods of Network Information Archiving," *Beijing Archives*, no. 2, pp. 13-16, February 2016.
- [9] Nie Hua. "Development of Scholar Information Assets System as the Core Components of University Research Infrastructure and Research Capacities," *Journal of Academic Libraries*, vol. 36, no. 5, pp. 80-86, September 2018.
- [10] Cai Cuimeng. "Research on the Influence Factors of the Value of Digital Resources' Long-term Preservation in the Library," *Journal of the Library Science Society of Sichuan*, no. 5, pp. 65-69, December 2017.
- [11] Xiong Yushan. *Research On Problems And Countermeasures of Social Media Document Archiving in China*. 2019. Hubei University, MA thesis.
- [12] Noblit, George W., and R. Dwight Hare. "Chapter 5: Meta-Ethnography: Synthesizing Qualitative Studies." *Counterpoints*, vol. 44, Peter Lang AG, 1999, pp. 93-123, <http://www.jstor.org/stable/42975557>.
- [13] SOUNG S. Evaluation Criteria for Scientific Information in the Numeric Era: The Case of Graduate Students in Education in Quebec Universities [J]. *Documentation Et Bibliothèques*, 2017, 63(3): 36-49.
- [14] KESHAVERZ H, GIMI M E, NOROUZI Y. "Credibility evaluation of scientific information on websites: Designing and evaluating an exploratory model," *Journal of Librarianship and Information Science*, vol. 52, no. 4, pp. 1086-1101, February 2020.
- [15] Liu Bing, Jiang Xiaohan. "Research on the Construction of Relationship Model of Network Academic Information Selecting and Judging of Scientific Researchers," *Library and Information Service*, vol. 63, no. 12, pp. 77-85, June 2019.
- [16] KESHAVERZ H. *Evaluating Credibility of Social Media Information: Current Challenges, Research Directions and Practical Criteria* [J]. *Information Discovery and Delivery*, 2020,
- [17] Liu Hui, Zhang Dongrong. "Research on the Structure and Optimization of Social Media Academic Information Ecosystem," *Information Studies: Theory & Application*, pp. 1-10, September 2021.
- [18] Zhang Xiaojuan, Tang Changle. "Overviews on Management of Long-term Preservation Metadata for Digital Information Resources," *Documentation, Information & Knowledge*, no. 3, pp. 43-52, May 2019.
- [19] Xu Kuan, Ren He. "Evaluation Basis Research on Content Value of the Long-term Preservation of Digital Resources," *Library and Information Service*, vol. 57, no. 13, pp. 72-75+100, July 2013.
- [20] Zhang Ziliang, Dong Hongbin, Tan Chengyua, et al. "Evaluation of Weibo credibility based on data provenance," *Application Research of Computers*, vol. 35, no. 11, pp. 3330-3334, November 2018.
- [21] Bi Qiang. "Exploitation and Innovation of the Study on the Evaluation of Information Quality in the Network Environment—Book Review on the Study on the Evaluation of Information Quality from Users' Perspective in the Network Environment," *Library and Information Service*, vol. 61, no. 4, pp. 138-142, February 2017.
- [22] Ye Fengyun, Shao Yanli, Zhang Hong. "An empirical study on the evaluation of information quality of mobile social media users based on behavioral process," *Information Studies: Theory & Application*, vol. 39, no. 4, pp. 71-77, April 2016.

CULTIVATING THE SCIENTIFIC DATA OF THE MORROW PLOTS

Visualization and Data Curation for a Long-term Agricultural Experiment

**Bethany G.
Anderson**

University of Illinois
United States
bgandrns@illinois.edu
[0000-0001-6602-1312](https://orcid.org/0000-0001-6602-1312)

**Sandi L.
Caldron**

University of Illinois
United States
caldron2@illinois.edu
[0000-0001-6392-5279](https://orcid.org/0000-0001-6392-5279)

Joshua Henry

University of Illinois
United States
jkhenry@illinois.edu
[0000-0002-7826-5960](https://orcid.org/0000-0002-7826-5960)

Heidi J. Imker

University of Illinois
United States
imker@illinois.edu
[0000-0003-4748-7453](https://orcid.org/0000-0003-4748-7453)

Hoa Luong

University of Illinois
United States
hluong2@illinois.edu
[0000-0001-6758-5419](https://orcid.org/0000-0001-6758-5419)

Kelli Trei

University of Illinois
United States
ktrei2@illinois.edu
[0000-0002-6436-1011](https://orcid.org/0000-0002-6436-1011)

**Sarah C.
Williams**

University of Illinois
United States
scwillms@illinois.edu
[0000-0001-7968-1870](https://orcid.org/0000-0001-7968-1870)

Abstract – The Morrow Plots at the University of Illinois at Urbana-Champaign are the longest-running continuous experimental agricultural fields in the Americas. This paper discusses efforts to identify, curate, and preserve data from the Morrow Plots and visualization tools to enhance understanding of the historical and scientific context for the data. This ongoing effort to draw attention to the greater scientific value of the Morrow Plots and to test data curation and visualization methods underscores the importance of interdisciplinary collaborations to curate longitudinal scientific data sets.

Keywords – data, agriculture, archives, curation, visualization

Conference Topics – Community; Exchange

I. INTRODUCTION

The Morrow Plots at the University of Illinois at Urbana-Champaign (UIUC) are the longest-running continuous agricultural fields in the Americas. Established in 1876 by the College of Agriculture and professors Manly Miles and George E. Morrow, the plots were created to facilitate a long-term experiment with crop rotations and fertilization. In

1968, at UIUC's centennial, the plots were designated a National Historic Landmark. The duration and uniqueness of the experiment garnered the plots historical significance. At the plots' designation, Congressman William L. Springer noted that "...through scientifically proven practices, the productive capacity of an acre of land can be multiplied fourfold" [1]. Despite recognition of the plots' scientific value, it is the duration of the experiment that scholars typically cite [2]. Assessing and understanding the Morrow Plots' scientific value has been difficult due to the scattered nature of the plots' data across various archival sources, which have been published in a piecemeal manner over time. Additionally, sources offer different information about the maintenance of the plots, the factors that effected their yields, and the ways the experiment evolved over time. The distributed nature of the data—and information about the data and the plots more generally—thus poses challenges for understanding the greater impact of the Morrow Plots and for accessing the data of this significant longitudinal agricultural experiment.

II. BACKGROUND

A. History of the Morrow Plots

To make the data set from the plots publicly available, and to celebrate the plots' sesquicentennial in 2026, the Morrow Plots Data Curation Working Group was established in 2018. Comprising agriculture and life sciences librarians, data management and curation specialists, an IT professional, and a science archivist, the working group seeks to identify archival records in digital and analog formats and aggregate, curate, and preserve the data in a usable and accessible format that can be broadly shared and preserved through the Illinois Data Bank [3].¹ Apart from curating the data set and identifying extant records to be transferred to the University of Illinois Archives, the working group seeks to create best practices and share lessons learned for curating and preserving a longitudinal agricultural data set. One of the challenges of curating the Morrow Plots data is displaying the data set in a format that can account for and illustrate variations and the ways the plots themselves evolved over time. The working group is also testing the ways that visualization can complement data aggregation and curation to provide a deeper understanding of the factors that influenced the plots and its scientific and historical context.

In this paper, we discuss the efforts of the Morrow Plots Data Curation Working Group in identifying and preserving relevant materials from the Plots, challenges in converting the data into reusable format to support open science, and the creation of a visualization that provides historical and scientific context for the data set to tell the story of the Morrow Plots. This case study of an ongoing effort to draw attention to the greater scientific significance of the Morrow Plots and test data curation and visualization methods and tools demonstrates the importance of interdisciplinary collaborations to curate longitudinal data sets. At the same time, we hope to demonstrate the value of visualization in complementing data curation efforts to facilitate historical understanding and scientific engagement.

In 1876, ten half-acre plots of land for corn, oats and clover hay were planted by Manly Miles, a professor of agriculture at the Illinois Industrial University. Initially known as "Experiment 23," the plots continued to be developed by the first dean of the College of Agriculture, George E. Morrow (1878-1894). It was Morrow who asked the university's Board of Trustees in 1880 for a "... formal commencement of what is designed to be a long continued experiment to show the effect of rotation of crops, contrasted with continuous corn growing with and without manuring, and also the effect of clover and grass in a rotation" [4]. While Experiment 23 was one of several agricultural experiments at the university, it specifically focused on the study of crop rotation. In 1895, ten years after the Illinois Industrial University was renamed the University of Illinois, an astronomical observatory was built on plots 1 and 2. An expanding university further reduced the experimental fields in 1903 to three remaining plots which were subdivided (3, 4, and 5). Despite this reduction, faculty continued work with the plots, such as Professor Cyril G. Hopkins, head of the Department of Agronomy (1900-1919), who focused his research on soil fertility [5].

Data of yields from the plots were not recorded between 1876 and 1887 [6]. At that time, six plots contained corn, two of oats, and two of clover (the latter being introduced in 1881). The introduction of fertilizer enabled study of not only crop rotation, but also the ways that fertilizers could enhance yield. The UIUC's student newspaper in 1927 noted, "Soil receiving no treatment in the three year rotation averaged 50 bushels of corn, 45 bushels of oats, and two tons of clover per acre. The portion of the plot receiving treatment aver 67 bushels of corn, 63 bushels of oats and 3.6 tons of clover per acre during this period" [7]. Over time, commercial fertilizers began to be used (1955) and oats were eventually replaced with soybeans (1967). The latter coincided with the university's growing interest in soybean research, including the establishment of an international soybean program in 1966 [8]. Today, faculty, students, and staff continue to study crop rotation and factors that affect the Morrow Plots' yields.

¹ For more information about the preservation architecture of the Illinois Data Bank, see <https://journal.code4lib.org/articles/15821>.

B. *Morrow Plots Working Group*

The Morrow Plots Working Group was formed in 2018 by the College of Agriculture, Consumer, and Environmental Sciences (ACES) at UIUC. Given the significance of the plots, and their sesquicentennial in 2026, the working group's mission is to identify and make publicly available data from the plots to facilitate use of the data and engagement with the plots' scientific legacy. The working group comprises an interdisciplinary team from both ACES and the University of Illinois Library that includes agriculture and life sciences librarians, data management and curation specialists, an IT professional, and a science archivist. The working group has engaged in several activities, including oral history interviews with faculty and staff on the history of the plots; data curation efforts; identification of relevant archival records and creation of a topic guide; and digitization of materials for public access. These efforts aim to broadly promote the history and scientific value of the Morrow Plots, and ensure its data is preserved and made accessible.

III. MORROW PLOTS DATA

Historical records tell us that we should expect to find crop rotation schedules for every year dating back to 1876, as well as yield and soil treatment data going back to 1888 [2]. One of the working group's aims is to clean and compile the data for all available years with the ultimate goal of creating learning objects for use in data science education. The data were originally recorded in ledgers but have been partially compiled by scholars who have previously published on the plots [2], [9], [10]. These were prepared for print and one of the chief challenges is creating a comprehensive machine-readable data set.

A. *Legacy Data*

Two farm/field managers from the Department of Crop Sciences at UIUC compiled existing data in two Excel files (in XLS format) that correspond with different phases of the experiment. The first file, which appears to be formatted for print, contains three parallel tables in the same sheet, one for each plot. All three tables are almost entirely complete and track year, plot, and soil treatment data from 1888 through 1954. The three-plot format with two layers of headings make it very easy for humans to

interpret immediately, but the file needs to be completely reformatted, with new variables added, to make it machine readable.

The second file, which tracks planting and yield data from 1955 through 2021,² is formatted for analysis in Excel and takes advantage of some of that software's many special features, like embedded charts and color coding, which provide a richer context for the data, but create challenges for both machine readability and digital preservation. Like the first file, the second file tracks year, plot, and soil treatment data, and includes additional variables for hybrid/variety, planting date, removed stover amounts and population (plants/acre), some of which are rather sparse. The second file also includes some ambiguities common to data sets that have not yet been curated, such as duplicate copies of the table in additional sheets, columns containing more than one data format or unit of measurement, and color coding of both cells and text.

B. *Data Wrangling*

We decided to employ the tidy data model for tabular data in which every column is a variable and every row an observation with one value per cell [11]. This format is supported by the tidyverse, a coordinated collection of R packages for data cleaning, visualization, and modeling. The tidy data format makes it easy to connect multiple data sets and pivot between different visualizations, which is useful for exploratory analysis. It can, however, require quite a bit of data wrangling up front. The tidy data format also addresses issues with data cleanliness and consistency that often do not arise until the publication and preservation stage. This is in a way a preemptive data curation strategy.

Although we plan to use R for the bulk of our work with the data, we started in Excel. Using Excel is risky because any transformations are made directly to the data. Excel also does not keep change logs or allow for easy backtracking. It is, however, expedient. We found it useful for quickly exploring the data and experimenting with different arrangements. One member of the working group, Heidi Imker, used Excel to transform a subset of the data into the tidy data format. To mitigate risk, she saved a series of versions at key points in the process. She was then able to import that data into RStudio, and link it to

² Data collection is ongoing, and we intend to expand our data set over time.

weather and crop price data, providing essential context for interpreting the data. Now that we have that end result as a guide, we can reproduce most of those steps in a safer tool like RStudio that allows for data manipulations in the software without changing the underlying data file.

This data cleaning is forcing us to grapple with difficult questions such as, how should we deal with notes made in cells that should contain numerical data? Text notes in the soil treatment variables are particularly challenging. Most treatments are measured in pounds with separate amounts recorded for the specific treatments applied. However, many cells contain complex text statements, which are specific enough to be valuable, but not structured enough to be machine readable. Some of these notes also record treatments measured in gallons instead of pounds, compounding the problem. To strike a balance between retaining the original information and cleaning the data, we shifted text comments to a Notes column, and flagged the plots as treated even though no specific amounts are recorded in the treatment columns. This at least preserves the original information in case we need it later. It also may be useful in keyword searches of the data set.

Overly complex notes can appear in any data set, but longitudinal data sets like this one face a particularly thorny challenge—how do we represent change to the experiment design over time? At first glance, the Morrow Plots themselves may look like they have not changed much while the campus expanded and grew around them, but the data tells a much more complicated story, especially when it comes to plot divisions.

Over time, the original plots were subdivided again and again as new variables were introduced. Each original plot eventually became eight subplots. One option would be to impose those subdivisions backwards in time and split old data into eighths. That would allow us to make clearer comparisons over time, but at a cost. It elongates the data for early years eight-fold. It also has a way of flattening time and presenting all the complexities of history at once. We could also widen the data set and create separate columns for plot and subplot. Whichever route we take, we will be sure to document our

reasoning and include an explanation in the data documentation. Perhaps we will incorporate decisions like this one into the learning objects we eventually produce, giving students the opportunity to explore the pros and cons of various data wrangling strategies.

C. Communicating Data Context

Data cleaning decisions like these require a firm grasp on the experiment design and its history. We can look to publications about the plots for context, but it takes work to translate narrative text into a mental map, and even more to keep track of how that map changes at key points in time. Visualizations and visual aids are much better suited to communicating spatial concepts like plot divisions, and current design tools make it easy to layer in all kinds of symbols that communicate much more quickly and easily than words.

To aid our own understanding of the history of the experiment, we created a Morrow Plots infographic (Fig. 1) that visualizes the plot divisions, how and when they changed, the crops grown, the rotation schedule, and the key phases of the experiment. We also included a timeline of historic markers to provide additional context and emphasize the experiment's longevity. The visual medium allows us to communicate not just with words and labels but with icons and colors. We used Canva, an online visual design tool with libraries of drag and drop graphic elements. Although these additional layers make the infographic more complex than the charts and maps typically found in academic publications, they paradoxically make it much easier to understand.

The visuals make it more engaging as well. Although we have not yet published it, drafts have been shared with several stakeholders across campus because it is eye-catching and fun. When people see it, they want to learn more and share it with others. It is still in draft form, but once it is finalized, we expect to use it as part of the upcoming anniversary celebrations and as context for any learning objects we produce from the data. We also hope it will help us reach beyond the typical audience for academic publications and engage the broader community.

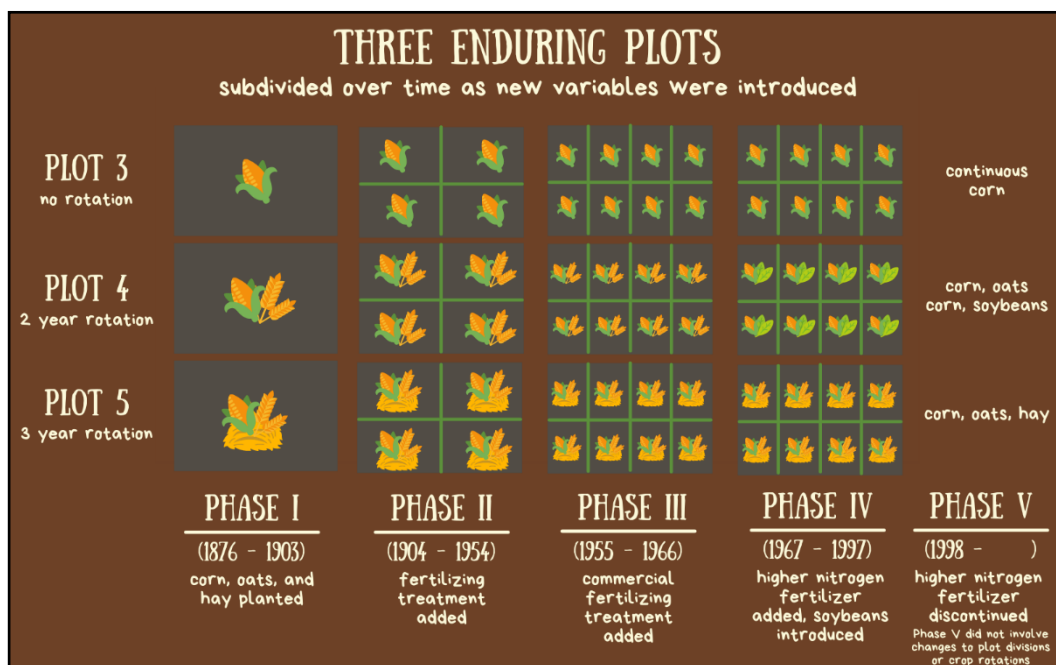


Figure 1 One section of draft infographic communicating plot divisions and crop rotations over time. See Appendix for full graphic.

IV. CONCLUSION

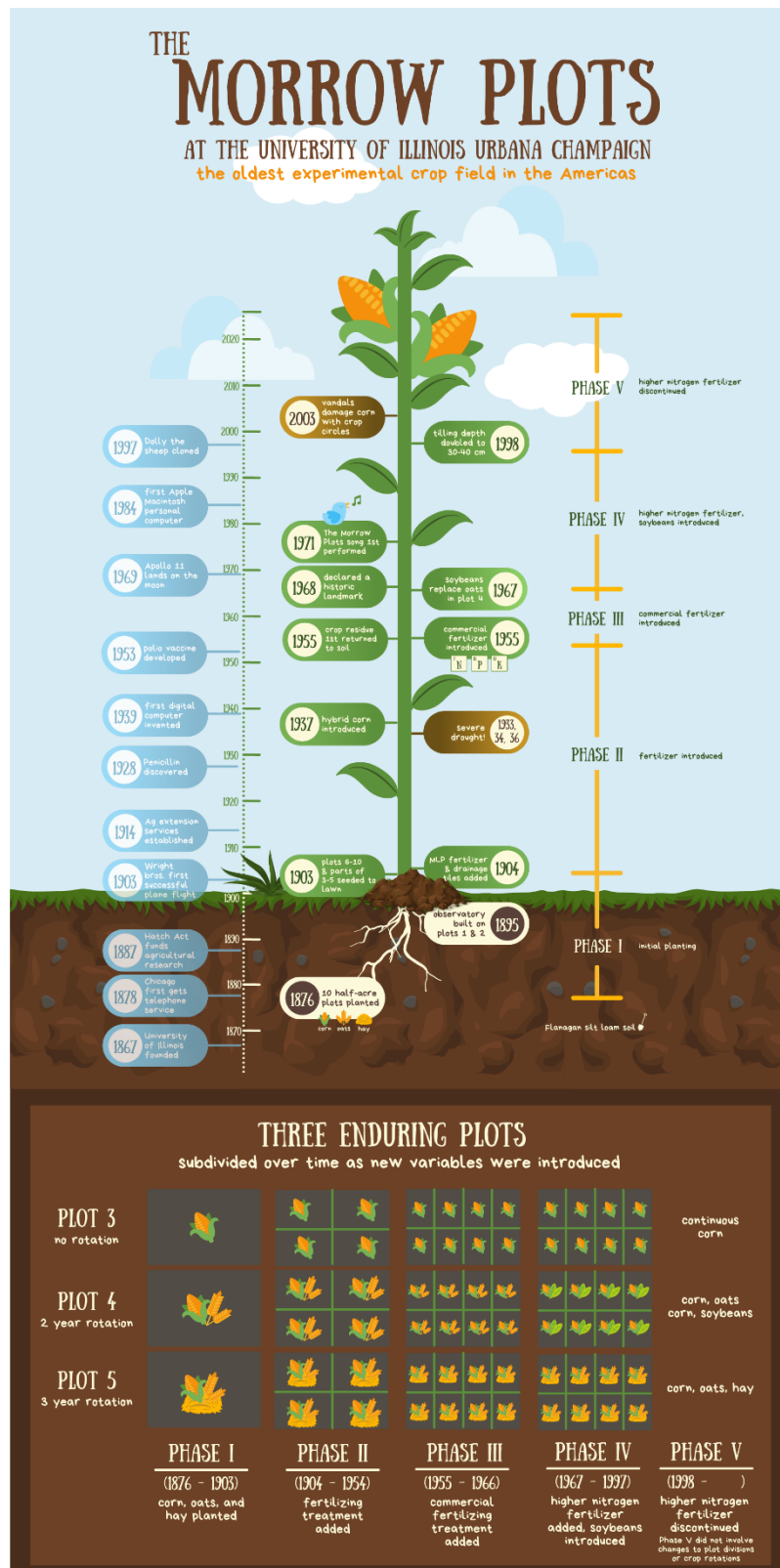
While the efforts of the Morrow Plots Working Group have made progress in discovering previously hidden data, curating the data, and enriching the connected context around it, there is still much work to do. Visualizations like the infographic included with this paper help to move the work forward by making both the richness and complexity of the data set more accessible and engaging for a broader audience. The tidy format produces preservation-friendly CSV files that distill the complexities of the data for broader accessibility. Efforts to fully curate the available yield data in the tidy data format for Plot 3 of the experiment are nearly complete, but this work still needs to be completed for the remaining plots. Additionally, work to identify and acquire additional archival sources continues, primarily through the work of the informational interviews with researchers and staff who through the years have worked with the Morrow Plots in some way. Finally, we hope for this work to culminate in 2026 with an interdisciplinary celebration and symposium of the historical and scientific contributions of the Morrow Plots.

REFERENCES

- [1] "Board of Trustees Minutes -1970," *Board of Trustees Biennial Reports*, Record Series 1/1/802, p. 118, University of Illinois Archives.
- [2] S. Aref and M. M. Wander, "Long-Term Trends of Corn Yield and Soil Organic Matter in Different Crop Sequences and Soil Fertility Treatments on the Morrow Plots," *Adv. Agron.*, 1997, doi: 10.1016/S0065-2113(08)60568-4.
- [3] "Illinois Data Bank." <https://databank.illinois.edu/> (accessed Mar. 07, 2022).
- [4] "Board of Trustees Minutes -1878-1880," *Board of Trustees Biennial Reports*, Record Series 1/1/802, p. 232, University of Illinois Archives.
- [5] *Cyril G. Hopkins Papers*, Record Series 8/6/21, University of Illinois Archives.
- [6] *Morrow Plots Notebook*, Record Series 8/6/16, University of Illinois Archives.
- [7] C. Stonberg, "Benefits of Soil Treatment, Crop Rotation Shown," *Daily Illini*, Jul. 08, 1927.
- [8] "Board of Trustees Minutes -54th Report," September 21, 1966, p. 84., University of Illinois Archives.
- [9] R. T. Odell, S. W. Melsted, and W. M. Walker, "Changes in organic carbon and nitrogen of morrow plot soils under different treatments, 1904-1973," *Soil Sci.*, 1984, doi: 10.1097/00010694-198403000-00005.
- [10] E. D. Nafziger and R. E. Dunker, "Soil Organic Carbon Trends Over 100 Years in the Morrow Plots," *Agron. J.*, vol. 103, no. 1, pp. 261-267, 2011, doi: 10.2134/agronj2010.0213s.
- [11] H. Wickham, "Tidy Data," *J. Stat. Softw.*, vol. 59, pp. 1-23, Sep. 2014, doi: 10.18637/jss.v059.i10.

APPENDIX

Morrow Plots Infographic (draft)



IT TAKES A VILLAGE IN PRACTICE

Growing Communities During a Pandemic

Megan Forbes

LYRASIS
United States
megan.forbes@lyrasis.org
[0000-0002-2611-1441](tel:0000-0002-2611-1441)

Laurie Arp

LYRASIS
United States
laurie.arp@lyrasis.org
[0000-0002-6929-4209](tel:0000-0002-6929-4209)

Abstract – What does it take to ensure open-source software (OSS) programs serving cultural and scientific heritage are sustainable and enduring? In 2017, the It Takes a Village (ITAV) project produced a Guidebook that serves as a reference source to help OSS programs plan for long-term sustainability. In 2020, the Institute of Museum and Library Services in the United States funded a new phase of It Takes a Village work, *ITAV in Practice*, to create and pilot an adaptable set of tools for practical use in planning and managing sustainability for OSS initiatives. It Takes a Village in Practice was proposed pre-COVID, but work did not begin until after pandemic-related restrictions had spread across our participating programs and organizations. This paper shares the challenges, successes, and lessons learned as the project team worked to grow and build the ITAV community and resource during the pandemic.

Keywords – open source, community engagement, sustainability

Conference Topics – Community; Resilience

I. INTRODUCTION

The It Takes a Village (ITAV) project [1], funded in 2017 by the Institute of Museum and Library Services (IMLS) in the United States, was designed to bring together open-source programs serving cultural and scientific heritage to develop shared sustainability strategies, and to provide our communities with the information needed to assess and contribute to the sustainability of the programs they depend on.

A growing body of open-source software (OSS) supports cultural and scientific heritage organizations, and while some initiatives have been successful at creating sustainable programs, others have struggled to determine what strategies will work once grant funding ends or other major pivots are required.

This paper describes the work that went into turning the ITAV Guidebook, published in 2018, into a practical toolkit that could be used by OSS programs to plan and manage for sustainability. We also discuss the challenges, successes, and lessons learned when a project predicated on extensive community participation and feedback was planned, proposed, and funded pre-COVID, and then forced to pivot to an all-virtual undertaking.

II. PROJECT HISTORY

The ITAV team assumed that while there is no single approach to sustainability, there might be common threads among programs that would lead to mutual needs and strategies for meeting those needs. In collaboration with the ITAV Advisory Group - Rob Cartolano, Columbia University; Tom Cramer, Stanford University; Michele Kimpton, LYRASIS; Katherine Skinner, Educopia Institute, and Ann Baird Whiteside, Harvard University Graduate School of Design - we developed a survey and conducted a two-day forum in Baltimore in the fall of 2017 to test this idea. During the Forum, representatives of 27 OSS programs discussed project lifecycles, governance, financing, resources, community building, outreach, and bumps in the road. Several digital preservation or related platforms were among the 27 programs represented, including DSpace, Archivematica, BitCurator, Fedora, and LOCKSS.

In looking at their own OSS programs, Forum participants articulated that sustainability is not a linear process with specific starting and end points. Instead, they defined OSS sustainability as an iterative process evolving across facets and phases. The four facets describe the different, but intertwined components of OSS sustainability:

governance, technology, resources, and community engagement. The three phases speak to a program's place within a facet: Getting Started, Growing, or Assessing and Evolving. Each facet is equally critical but may be in a different phase and have different timelines and needs.

These findings were combined with other resources and shaped into a Guidebook [2] freely shared with the larger community in 2018. The Guidebook serves as a practical reference source to help plan for long term sustainability, ensuring that commitment and resources are available at levels sufficient for a program to remain viable and effective as long as it is needed, by: 1) creating a framework for evaluating sustainability – using a combination of lifecycle phases and sustainability facets; 2) identifying goals, characteristics and common roadblocks for each phase in each facet; 3) providing guidance for moving programs to the next phase; 4) highlighting case studies and additional resources to help programs; and 5) including the full survey results as a reference source and benchmark.

In 2020, IMLS funded a second phase of *It Takes a Village* work, *It Takes a Village in Practice* (ITAViP) [3], in which we are working to create and pilot an adaptable set of tools (e.g. templates, checklists, discussion and process guides, etc.) for practical use based on the framework laid out in the ITAV Guidebook. The results of the project will strengthen the ability of libraries, archives, and museums to sustain community supported OSS programs, which are critical to managing and growing local and national digital infrastructures. *ITAV in Practice* will enable all stakeholders in an OSS program to participate in an assessment of each facet of sustainability at current phases, develop balanced strategies to advance sustainability goals, and integrate sustainability plans into other organizational planning efforts.

Now in its second year, the ITAViP program team has collaborated with stakeholders from a wide range of different types and sizes of OSS programs through a series of four workshops to address each of the four facets of sustainability. Several OSS programs, including Quire, ePADD, Folio, and VuFind, have beta tested the tools for each facet to identify gaps, challenges, and unaddressed needs. In the summer of 2022, *ITAV in Practice* as a whole will be piloted with two additional OSS programs: Samvera and Mukurtu [4].

III. CHALLENGES

We have faced several challenges during the ITAViP project. Chief among them was the need to pivot from a series of in-person information gathering and tool development events to a set of virtual meetings due to COVID-19. The planning and agenda development process needed to be completely re-thought; for example, we intended that the workshop attendees would do a good bit of brainstorming around identifying and developing new tools for each facet. It became clear as we planned the virtual events that instead, creating a set of strawman tools for workshop attendees to critique would gather more effective feedback. This change shifted the burden on to the ITAViP project team to identify, and in many cases create, tools for workshop attendees to evaluate in advance of each meeting.

We also have several silver linings from the pivot to virtual events. First, was our ability to increase the number of participants, as budgets for travel and accommodations were no longer a concern. We were also able to engage participants who may have a hard time traveling or attending in-person events, including primary caregivers, international attendees, and people with disabilities. In our follow-up surveys after each workshop, some attendees lamented the loss of the serendipitous conversations that spring up at in-person gatherings, while others noted that it was easier for them to say yes to virtual events.

Early in the grant period, we sent out a call for interested programs to volunteer to beta test or pilot the toolkit. We had an excellent response rate, with many programs volunteering to beta test multiple facets. As with our in-person workshops, the realities of living and working with COVID struck in the eighteen months between programs' expressing interest and then being asked to test. During that time, program leadership and priorities shifted, leading to several of our beta testers having to withdraw from testing. The ITAViP program team also counted on programs beta testing more of the toolkit than they were eventually able to, leaving many tools untested at the end of our beta period. Of course, changes in leadership, priorities, funding streams, etc., are events that should lead to a renewed focus on sustainability planning. It can be very difficult, however, to advocate for long-range planning when the short-term feels as though it is in crisis.

IV. LESSONS LEARNED

Despite these challenges and setbacks, the ITAV Toolkit is on track to launch successfully later in 2022. The core lessons we have learned about building, testing, and sharing community-developed resources during this difficult time include:

- Pivoting to virtual does not mean just holding the same planned in-person event online. The ITAViP team worked to assess and select virtual collaboration tools; create an agenda that gave all participants equal opportunity for participation and engagement while also accommodating shorter online attention spans; and provide advance readings, Q&A sessions, and tools to ensure that time spent during the meeting was not spent getting everyone up to speed on the agenda and process.
- A pivot to virtual does not have to mean an event is less than; rather, it may open participation to new audiences, including staff that are unable to travel, junior staff without travel funding, and others.
- It is easy to overestimate what the conversion rate from interested parties to participating stakeholders will be, especially over long-time spans. Continual engagement, low barriers to participation, and support from the program team can all help mitigate attrition rates.

Through ITAV in Practice, we expect that libraries, archives, museums, and academic institutions will be able to take a long-term view of the OSS they use, so software is created not just to fix a problem, but rather to endure and provide functionality for as long as it is needed.

REFERENCES

- [1] <https://www.lyrasis.org/itav>
- [2] https://www.lyrasis.org/programs/Documents/ITAV_Interactive_Guidebook.pdf
- [3] <https://wiki.lyrasis.org/display/ITAV/It+Takes+a+Village+in+Practice>
- [4] <https://wiki.lyrasis.org/display/ITAV/ITAViP+Toolkit%3A+Home>

VAULT: BUILDING AN EXTENSIBLE, AFFORDABLE DIGITAL PRESERVATION & REPOSITORY SERVICE

Jefferson Bailey

Internet Archive

USA

jefferson@archive.org

[0000-0002-0830-6325](tel:0000-0002-0830-6325)

Abstract – Since 1996, the Internet Archive (IA) has provided storage, preservation, and access infrastructure and services to over 1,000 cultural heritage organisations around the world. It has also provided customized digital preservation services on a contractual basis to a handful of large institutions. In 2020, IA began building a more generalized digital preservation and repository service in response to the needs of a broader range of institutions and to leverage IA's self-owned data centers, non-profit cloud services, and demonstrated expertise in both small and petabyte-scale digital stewardship. This system is being developed in direct dialogue with 30+ organizations, including universities, public libraries, arts organizations, and cultural heritage organizations, over the course of the 2021 - 2022 year. This paper shares key takeaways from the information collected from this pilot phase and early launch of the service and also positions the service, Vault, within the digital preservation landscape, particularly as it relates to the distinct needs and goals of nonprofits, libraries, and cultural heritage organizations, that this service aims to address.

Keywords – digital preservation, product development, archiving, open infrastructure, sustainability

Conference Topics – Scanning the New Development; Building the Capacity.

I. INTRODUCTION

Since its inception as a non-profit digital library in 1996, Internet Archive has focused on ensuring the continued availability and accessibility of human knowledge by creating a digital library to permanently store digital content. The Internet Archive is the world's largest public web archive, with hundreds of petabytes of data stored within its independently owned and operated, not-for-profit data centers. Currently 1000+ partners, including national libraries, universities, and cultural heritage

organizations, collaborate with the Internet Archive on various archiving, access, open source technology development, and digital library projects with the shared mission of ensuring perpetual preservation and access to diverse, cultural, and historically-relevant digital collections from around the world.

Internet Archive has built a new general purpose digital preservation service to complement and extend its existing suite of free, paid, and subsidized non-profit services for digitization, web archiving, general data storage, and web and access services. The new Digital Preservation Service, called Vault, is built on existing Internet Archive infrastructure and open-source software and has incorporated the feedback of dozens of pilot partners and peer stakeholders who are using the service as it is developed and progresses through the product life cycle. The pilot phase and early rollout has featured iterative development cycles informed by pilot partner usage and by broader input from the community of users of IA services. One of the goals is to build a service that prioritises simplicity, extensibility, and a cost-model that makes it available to organizations of any type and size and embeds the principles of the original NDSA Levels of Digital Preservation in its design. [1][2] One of the co-creators of the Levels of Preservation is the Director of the service and the service aims to take the guidance and principles of the Levels of Preservation and translate them into a best-of-class service for the cultural heritage, non-profit, and social impact sectors. Engaging users at all stages of development will help ensure the service's fidelity to the goals and needs of mission-aligned organizations and, in turn, further the capacity of these non-profit organizations to preserve and protect valuable materials for the public good.

II. CENTRAL FEATURES OF THE SERVICE

At its core, Vault allows users to deposit any digital content of any size, specify what geographical location their data will be stored (across multiple locations in, currently, 3 different countries), set how many copies of the data will be replicated and their distribution across various data centers in various regions, select if they want their collections stored in different technical architectures, and select the frequency of audit and repair operations such as fixity checking and digital object correction. Vault also features a variety of standard collection management tools. In line with IA's user-centric design philosophy, a key success metric for Vault is the ability to accommodate the diverse preservation goals of organizations of various sizes, locations, and expertise in digital preservation management. Vault is responsive and customizable to various use cases, with partners able to select custom numbers of copies, specify desired storage locations, and schedule multiple fixity occurrences with service levels from basic storage services to highly-replicated, full-features distributed digital preservation services.

The service has an interactive dashboard to view the real-time status of all preserved data, including storage location, fixity reporting, manifests, analytics, and other transactional metrics, so that partners will be able to actively monitor their data and make timely decisions about its organization or the what various service features should be implemented for specific collections within their overall account. Reports and metadata will also be available through APIs, with additional plans for integration with peer services, repositories, and preservation systems in progress.

Mindful of the resource constraints of nonprofits, Vault also benefits from the Internet Archive's efficiencies of scale to offer storage and preservation solutions at minimal cost so that mission-aligned organizations, particularly those who have heretofore been unable to participate in digital preservation practice, no longer have cost or technology as a barrier for entry, a common finding in IA's regular "State of WARC" survey amongst IA partner organizations. [3]

III. STAKEHOLDER NEEDS ASSESSMENT

From continued conversations with stakeholders, the Internet Archive product team has

learned a great deal about the current digital preservation service gaps experienced by libraries, universities, and non-profit organizations. Many organizations have also used the Levels of Preservation guidance as an assessment tool for analyzing or planning for their own digital preservation activities. [4] An early, welcome surprise to the team's call for participants was the high amount of enthusiasm and demand for this service, suggesting that current service providers are not meeting the variety of needs of many heritage organizations, especially smaller or more unique libraries and archives. Our intended 3 group calls more than doubled to 8 to accommodate the growing number of organizations wanting to participate in these initial conversations. Several organizations were keen to incorporate the features of the service in their long term organizational preservation planning and plan to develop their digital preservation strategy alongside our service's product development. Additionally, several organizations voiced dissatisfaction with current commercial solutions (detailed further below). The data amassed from our needs assessment form and early conversations guide our belief that Vault taps into a high need area for mission-aligned, memory organizations. In all, over 50+ organizations engaged with us, including college or university libraries, public libraries, religious, specialty, or research libraries and archives, arts and museum institutions, multiple consortia entities, and international organizations. As potential users, many organizations had distinct ideas for how they would like to use Vault and shared how their current process or solutions are in need of improvement. These findings include:

Priority Features: We asked each potential pilot partner to indicate which feature they were most interested in for their organization. Features deemed most desirable were 1) geolocation options that would allow partners to select between 3 countries in 2 continents for storage, 2) dashboard tools that provide clear data monitoring, provide simple visual representation of the content, status, and activity related to preserved data organized in various collections, and ready access to audit and repair reports, and 3) replication functions that allow partners a flexible means to manage content replication according to various criteria related to types of digital objects and to initiate more or less copies for different subsets of their content. As several organizations desired to store audio-video

files within the service, preview and appraisal tools were of higher need than initially anticipated, and highlighted an aspect of digital collections management that suggest a need for temporary basic storage that can be easily connected with preservation storage with more flexibility for administering collection status.

Nascent Digital Preservation Practice: Roughly a third of the pilot participants indicated a need for more support for building their digital preservation strategy. In addition, these organizations lacked external digital preservation tools or service providers. These organizations were in the early stages of developing a digital preservation strategy, had been attempting to DIY solutions at a level that were deemed unsatisfactory, or, in one case, had their existing digital preservation service decommissioned within the last few years. Amongst this particular group of organizations, there was a large spectrum of technical proficiency, some had attempted to build a patchwork of services in house while most had generally kept digital records on a server without a comprehensive digital preservation plan. All organizations within this category came to the product team hoping for ready-made solutions for active monitoring of their data, for replicating digital materials as needed, and ensuring perpetual access.

Large-Scale Grants and Acquisitions: Multiple organizations viewed Vault as a solution for an anticipated influx of digital materials in conjunction with recent grants for digitization efforts or for new acquisitions to their collection. One potential partner reported that they will be acquiring an additional 100 TB of data from grant-funded projects within the year and another institution shared that they were in need of 500 TB of storage for a new film and media archive. The large scale of these new acquisitions warranted a digital preservation solution that can accommodate the size of these collections and provide the adequate tools for organizational oversight, including large-scale fixity checks and comprehensible reports.

Consortial Considerations: Large consortial organizations, many of which engaged in early conversations with the product team, described the necessity of nimble solutions that can sufficiently address the needs of their various participating member organizations. Responding to the divergent needs of many affiliated organizations tests the strength of the service's customizable controls and

options. Such organizations present valuable operational opportunities to apply both consortial and individual organizational digital preservation strategies to diverse use cases.

Dissatisfaction/Cost Constraints with Commercial Services: Many of our potential partners shared difficulties and limitations with current service providers. The overriding difficulties related primarily to 1) unintuitive interfaces not mapping to desired workflows, including a lack of options with fixity checks, 2) high expense associated with commercial services, with maintenance worries if organizations experience lapses in funding, 3) cost constraints associated with commercial services relegating such options to one of many patchwork services that do not add to a comprehensive, end-to-end solution for organizations. Most organizations within our pilot partner group are not able to afford more holistic service offerings which results in considerable operational investment from staff and additional difficulties when managing expanding collections.

IV. NEXT STEPS

With the wealth of information provided by the initial cohort group, the product team has prioritized features into the initial product design of the service. Mindful of the expressed need for a comprehensive service with clear and accessible controls, the team will continually validate feature development and service offerings in direct collaboration and dialogue with the pilot cohort.

The pilot is currently underway and intended to run into early 2022 with the goal of providing early, no-cost access to the service's core features for 30+ selected organizational partners in exchange for their use, input, and feedback on ongoing technical development. Pilot partners have been depositing multiple terabytes of data are guaranteed perpetual preservation and access to the data and ongoing access to the service. In return, pilot partners are providing their feedback to the product team in quarterly check-in calls, meetings, survey forms, and other communication instruments.

The iterative, co-creation design principles of the service's development lifecycle bolsters the Internet Archive's capacity to build relevant, accessible, and sustainable preservation solutions for mission-aligned organizations. The success of such efforts will be measured through stakeholder feedback

sessions, user testing, and, eventually, in the uptake of the service's offerings within the digital preservation ecosystem. In line with the open source ethos of the Internet Archive, findings and lessons learned from the development and launch of the digital preservation service will be shared with the larger research community to further the field. In addition, the Internet Archive will continue to pursue collaborations, integrations, and/or future testing opportunities with diverse, mission-aligned organizations to ensure services developed are as inclusive as possible of various cultural and technical contexts

REFERENCES

- [1] Bailey, et al, "The NDSA Levels of Digital Preservation: An Explanation and Uses," Proceedings of the Archiving (IS&T) Conference, 2013, https://web.archive.org/web/20190207222334/http://www.digitalpreservation.gov/documents/NDSA_Levels_Archiving_2013.pdf.
- [2] Levels of Digital Preservation, <https://ndsa.org/publications/levels-of-digital-preservation/>.
- [3] Archive-It, "State of the WARC," 2020, <https://archive-it.org/blog/category/state-of-the-warcs-reports/>.
- [4] Baucom, et al, "Using the Levels of Digital Preservation as an Assessment Tool," NDSA, 2019, <https://osf.io/m6j4q/>.

REPOSITORY SPEED DATING

A methodology for narrowing the field

Karin Bredenberg

*Kommunalförbundet Sydarkivera
Sweden*

*karin.bredenberg@sydarkivera.se
[0000-0003-1627-2361](tel:0000-0003-1627-2361)*

Sven Schlarb

*Austrian Institute of Technology
Austria*

*Sven.Schlarb@ait.ac.at
[0000-0003-3717-0014](tel:0000-0003-3717-0014)*

Carl Wilson

*Open Preservation Foundation
United Kingdom*

*carl.wilson@openpreservation.org
[0000-0003-1535-1770](tel:0000-0003-1535-1770)*

Abstract – Selecting a repository system is a task many collecting institutions have to carry out at least once. There are many challenges, while the variety of alternative systems available is a good thing, making sense of the marketplace can be difficult. Assessing potential candidates is time-consuming and it's difficult to reuse the work of others as every organization has unique requirements. Here we present a simple methodology intended to help organizations to narrow the field by putting together a high-level set of requirements based upon the OAIS Reference Model, placed within the context of the OAIS Reference Model. This can help organizations evaluate solutions to create a shortlist of suppliers.

Keywords – OAIS, selecting, matrix, evaluation, working-together

Conference Topics – Community.

I. INTRODUCTION

Have you faced the same issues I've encountered at my institution when thinking about updating your digital repository and wanting to align with the OAIS Reference Model [REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS), ISO 14721:2012]? If you haven't yet, it's likely that you will have to soon given digital preservation's constant evolution. Once you reach that point, where do you start? How can you assess which of the variety of architectures and software systems match your criteria and requirements without immediately starting detailed and time-consuming discussions with vendors? Your organization's procurement policies might even mean contacting vendors isn't initially an available option.

Ideally, there would be a place where it's easy to compare available repository systems so you just can pick one, similar to choosing an air fryer after reading reviews by "Råd och Rön" or another consumer organization. In reality, these resources aren't available and they would likely be of limited use since

organizational requirements are unique regarding the types of content preserved, institutional policies and local legislation.

In this paper, we present a methodology for carrying out a high-level evaluation of potential repository solutions or an assessment of existing repository systems against a set of requirements, we call it speed dating for repository systems. This uses the OAIS model as a device for classifying and aggregating requirements and repository features to produce an easy-to-use evaluation matrix.

II. BACKGROUND

One of the leading resources for working with digital repositories is the OAIS Reference Model. As the name suggests, this presents a reference model, which outlines the components for an ideal archival information system. The model's scope extends beyond software products, covering the organization and staff that administer and manage the system. The reference model is ubiquitous in digital preservation disciplines, providing key terms and functional definitions. However, there are still very few good educational resources for the model, meaning practitioners must often educate themselves. In combination with the OAIS Reference Model, you can use the Audit and Certification of Trustworthy Digital Repositories (TDR), ISO 16363:2012 to ensure that the whole archival information system, including administrative and management functions, align with and follow OAIS. The certification standard contains a huge number of requirements, with TDR certification coming at a financial and resource cost. With the caveat that this methodology is not certifying the technical system, instead, it is certifying the whole system, including the organization that hosts and administers the repository. Other certification and self-evaluation

models are available, but they still do not evaluate the system itself. Despite resources, we are still somewhat in the dark regarding the technical aspects of the digital repository itself, and what we should compare when procuring new systems or refreshing existing ones. How can we be sure that the digital system storing our information packages adheres to the OAIS Reference Model? We often have to simply trust that vendors and solution providers are implementing OAIS properly and that their system follows the model. The authors aren't saying you shouldn't trust vendors and solution providers. Indeed, trust in your supplier is essential. However, when it comes to choosing a product, I want to ensure it satisfies key requirements, important to my organization to make informed decisions without contacting vendors. This means that I need to be able to trust the information available online when at the information gathering stage. System descriptions, manuals and fact sheets should be open for all and easily accessible. We need sufficient accurate information to start an initial evaluation before carrying out a Request For Information (RFI) or full-blown procurement process.

The OAIS reference model is split into a number of functional areas. I want to be able to evaluate the available systems that align to these areas. I also need to find the criteria that are important to my organization, since our digital preservation mission will not be exactly the same as yours. Organizations will see different criteria as more or less important than another, meaning there are few shortcuts when evaluating digital preservation systems. It's not possible to simply copy someone else's approach regardless of how much you like it. You need to put in the time and resources to first figure out what you want the system to do for you and then set up a matrix to assist you in making your evaluation. The goal is to be able to use detailed requirements as a guide to assist you in narrowing down the choice of available options. The labour involved can be organized in different ways, but you need to recognize that it is a necessary step in planning for a new repository platform.

III. USING OAIS TO ESTABLISH AN EVALUATION FRAMEWORK

The OAIS model provides a conceptual framework for a digital archive, consisting of an organization of people and systems with responsibilities to preserve and provide access to

information. The reference model includes concepts and terminology that can be used to describe and compare architectures and operations of digital archives as well as preservation strategies and techniques. The methodology described in this paper uses this framework to provide a foundation for categorizing requirements to produce an evaluation matrix for comparing archival systems. While we assume some familiarity with OAIS terminology this section provides definitions of key concepts used in the methodology.

A. *The OAIS Environment*

The OAIS environment defines four interacting entities, producers of information, consumers of information, a management entity which sets policies for archival content, and the archive itself. The term "consumers of information" describes a broad population of users wishing to access content held in the archive. OAIS also defines a specific term for groups of consumers identified by archives, Designated Communities.

Designated Community: A group of consumers defined by an Archival organization by some criteria, e.g., occupation or location, who require access to particular information sets. The Designated Community may be composed of multiple user communities and its composition may change over time.

B. *The OAIS Information Model*

The OAIS model defines an information model for digital items, known as data objects and any metadata needed to interpret data objects. These are components of Information Packages which consist of data objects and any metadata required to support long-term preservation and access bound into a logical package. OAIS identifies three types of Information Package described below.

Submission Information Package (SIP): An Information Package that is delivered by the Producer to the OAIS for use in the construction or update of one or more AIPs and/or the associated Descriptive Information.

Archival Information Package (AIP): An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS.

Dissemination Information Package (DIP): An Information Package, derived from one or more AIPs, and sent by Archives to the Consumer in response to a request to the OAIS.

Designated Community: A group of consumers defined by an Archival organization using some criteria, e.g., occupation or location, who require access to particular information sets. The Designated Community may be composed of multiple user communities and its composition may change over time.

C. The OAIS Functional Entities

The OAIS reference model describes six distinct functional entities defined below, that can be used as a starting point when evaluating digital repository systems.

Pre-Ingest/Ingest: This functional entity represents the boundary between the archival system and incoming information packages. This means that solution coverage is more variable than for the other entities as different solutions effectively draw their own boundaries and might rely upon the task being handled by another system or in another part of the information package's creation.

Archival Storage: This functional entity can be regarded as the foundation of all other repository functionality. Secure long-term storage of digital content and metadata is the prime function of digital repository systems. Ingest, access and preservation planning functionality are all, in a sense, layered on top of archival storage. As such they can be improved and refined over time provided the underlying archival storage is well designed and reliable.

Preservation Planning: An OAIS function that encompasses archival activities required to ensure digital collections remain accessible and comprehensible over time. These activities include developing/creating strategic preservation policies applicable to all digital content, as well as any action plans specific to particular collections or technologies. This is a proactive function which identifies/anticipates changes that may impact the long-term preservation of and access to digital collections. These include internal and external changes to the archival organization, evolving standards and technologies, e.g., storage mediums or file formats, and changes in the needs and expectations of Designated Communities. This is a

difficult function to automate and evaluate as much of it depends on specific institutional requirements.

Access: This functional entity marks another system boundary, in most cases the Designated Community are external consumers, usually researchers. The entity contains the services and functions that make the archival information holdings in the form of digital objects and related services visible to the researcher. This step involves an archivist reviewing the digital objects to make sure that there is nothing that the researcher isn't authorized to view, for example for data protection reasons. This means that a DIP for the archivist to review needs to be created before a redacted DIP for a researcher is created. This practice varies depending on your institution; sometimes the DIP is created when the ingest is made and sometimes it is created upon request.

Data Management: The functional entity of data management is a somewhat biased entity. It contains services and functions for populating, maintaining, and accessing a wide variety of information. This might imply that it is a database solution for handling a number of different statistics like access, billing and security control. It is also the function responsible for managing the repository's descriptive and preservation metadata.

Administration: The administration entity is the catch-all for the organizational, procedural and technical glue that brings the system together, and integrates it within an organization's business as usual activities. It has a broad scope covering non-technical processes and activities, operational management and institutional policymaking. Many of the administration functions are outside of the scope of many, if not all, repository systems. When it comes to infrastructure or process management, existing, dedicated solutions are in place. Leveraging existing domain software to address gaps in the functional coverage of the chosen solution will be the pragmatic, (possibly only) choice.

IV. METHODOLOGY

Organizations embarking on a procurement process can be faced with a choice between a large number of possible approaches and technologies. Attempting a thorough assessment of all available options is often time consuming and impractical. This means assessment can prove an intimidating task and lead to analysis paralysis. Here we present

an overview of a methodology designed to "narrow the field" when assessing potential repository solutions. We divide the analysis into three steps that can be summarized as:

1. Gather Requirements: Define what your organization wants from a repository system.
2. Aggregate and Prioritize Requirements: Create logical groups of requirements based on OAIS functional areas.
3. Research Solutions: Carry out high-level desktop research of potential solutions and/or existing systems.

Following this approach will help institutions to classify their requirements into logical groupings consistent with the OAIS reference model. The methodology could be used for other purposes, for example to perform an assessment or gap analysis of an existing archival system.

A. Gathering Requirements

The first step is to put together a set of requirements, defining what your institution wants from a digital repository. What constitutes a good requirement gathering exercise could be the subject of its own paper. Much depends on the size of your organization and the scale of your digital collection. Generally, they can be divided into two types, functional and non-functional. The distinction isn't always clear but functional requirements describe what the system should do while non-functional requirements describe how it should do them. It can be helpful to look at requirements documents put together by other organizations, particularly organizations of similar size with similar aims. Another quick start might be the OAIS reference model itself, which describes the functional entities that comprise an ideal digital repository in some detail. These should only be used as a starting point though.

At this early stage, it can be helpful to consider system capabilities, rather than focus on overly detailed requirements. For example, the ability to scale a system to meet performance criteria might be more meaningful than trying to stipulate a maximum throughput figure that will be hard to evaluate.

B. Aggregate, Group and Prioritize Requirements

Once a set of requirements has been defined, the next step is to start to gather them into logical groups, defined by the OAIS areas described

previously. The initial evaluation is simply a case of researching publicly available material for compliance with requirements. Again, this is more easily performed if the requirements are more general/coarse-grained. A concrete example might be requirements around integrity checking content held in archival storage. Your organization might have detailed requirements as to digest algorithms and the number of storage nodes supported, e.g., SHA-256 checking across four archival nodes. While it's important to note these requirements, this level of detailed evaluation will come later. For now, aggregating these together as "Audits and integrity checking" under the Archival Storage category is enough.

This simplification process might take some time. We have used six functional areas derived from the OAIS model. The goal should be to have five or six aggregated requirements per functional area. This isn't a prescriptive rule but ten or more requirements is probably a mistake as the detailed information to evaluate them is unlikely to be available for most of the systems. To reiterate, you can reduce the number of requirements to consider by aggregating similar requirements together as a coarse-grained, more general requirement. Considering the priority of requirements is another method, low priority requirements may be left out altogether at this stage. They will be reintroduced later when carrying out a detailed analysis of the solutions shortlisted by this process.

Finally, these grouped requirements should be arranged as the top row of a matrix, in a spreadsheet. Here's an example:

Archival Storage	
Secure preservation storage	Scalable storage and processing
	Audits and integrity checking
	Flexible/resilient IP processing
	Monitoring and reporting

Figure 1 Matrix header row example.

C. Product Research

The next step is to evaluate each of the potential solutions against the criteria listed. The aim is to identify and eliminate solutions that have obvious gaps, not to perform a detailed evaluation. The

product research is carried out simply using information available on the internet such as:

- Product and vendor websites.
- Online manuals and help guides.
- Community forums if available and accessible.

For each of the requirements, only a yes/no answer is required, if the decision is difficult err on the side of generosity and give a yes. A more in-depth evaluation is the place for making trickier decisions.

As the solutions are evaluated against the criteria you simply fill in the appropriate box in the matrix, see example below. The name of the solution and the answers are randomly inserted here to illustrate the methodology.

	Archival Storage					
	Secure preservation storage	Automatic replication	Scalable storage and processing	Audits and integrity checking	Flexible/resilient IP processing	Monitoring and reporting
Solution X	y	y	y	y	y	y
Solution Y	y	n	n	y	y	y
Solution Z	y	y	y	y	y	y

Figure 2 Matrix example.

V. WEAKNESSES

This approach is useful when initially trying to make sense of the options available and eliminating those that are clearly not fit for purpose. It also leaves a reasonably objective record of the solutions considered and the reasons for elimination. It is far from a forensic investigation of detailed system specifications, that is for the full sourcing and procurement process using the solutions that remain.

Because the assessment is performed using information published by vendors it might not be possible to make an informed judgement on all criteria. If detailed information about a particular product is hard to find it may say something about the product itself or the level of support available. The results of this process will only be as good as the work put in. Ensuring that the requirements accurately reflect your institution's priorities and

taking the time to search for detailed information will improve the results.

Another potential issue is that the requirements are "framed" by the OAIS model. This means that it's possible to overlook important criteria that don't align with OAIS. Examples include:

- Relationship with the vendor, ensuring that the vendor is a good cultural fit for your organization.
- Institutional or national policies and regulations. While a cloud-based solution might be attractive there may be good reasons that your organization needs to control the geographic location of data.

VI. NEXT STEPS

This high-level evaluation is only a first step toward sourcing a digital repository system. What comes next depends upon how you use the results of the evaluation. This is still a somewhat subjective exercise, not just a case of counting yes and no scores. "Knock out" criteria can help, these are mandatory features which must be supported by candidate solutions. Beware overusing this blunt instrument, you may arrive at a situation where no available system satisfies all of your mandatory features.

One approach is to consider the next phase and how many options you can realistically evaluate. If the next phase is a full procurement, with vendor interviews and product demonstrations, then it's probably unrealistic to consider more than 5 solutions due to the evaluation effort involved. If ten or eleven candidate solutions remain, you might decide to perform a second sift using the fine-grained requirements that were aggregated earlier. This assumes that the data needed to make more nuanced distinctions are publicly available. It may be necessary to contact vendors for more detail. This could still be more informal than an official procurement, using a questionnaire or a set of scenarios for the suppliers. In the end, it's your decision, and approaches that work for other organizations might not be as well suited to yours.

OAIS-COMPLIANT DIGITAL ARCHIVING OF RESEARCH AND PATRIMONIAL DATA IN DNA

Pierre-Yves Burgi

*IT Division, University of Geneva
Switzerland
pierre-yves.burgi@unige.ch
[0000-0002-4956-9279](tel:0000-0002-4956-9279)*

Jan Krause

*Archives cantonales vaudoises
Switzerland
jan.krause@vd.ch
[0000-0001-9375-2102](tel:0000-0001-9375-2102)*

Linda Meiser

*Switzerland
[linda-meiser@hotmail](mailto:linda-meiser@hotmail.com)*

Dina Andriamahady

*Medical Library, University of Geneva
Switzerland
dina.andriamahady@unige.ch*

Hugues Cazeaux

*IT Division, University of Geneva
Switzerland
hugues.cazeaux@unige.ch
[0000-0002-5618-2670](tel:0000-0002-5618-2670)*

Lamia Friha

*IT Division, University of Geneva
Switzerland
Lamia.friha@unige.ch*

Basma Makhoul Shabou

*School of Business Adm. of Geneva
Switzerland
basma.makhoul-shabou@hesge.ch
[0000-0003-0980-0517](tel:0000-0003-0980-0517)*

Abstract – Nowadays digital data is produced at an exponential rate. The predominant storage and long-term conservation solutions use power-hungry spinning disks. DNA technology represents an ideal candidate for storing data because of the optimum ratio between energy and density of information it can contain, its relative longevity and above all its status at the foundation of life, preserving it from any technological obsolescence. Consequently, in this paper we present the design of a scalable archiving solution based on DNA, interfaced to OAIS-compliant digital repositories, and which allows an effective and efficient implementation on the long term in a perspective of saving research and patrimonial information.

Keywords – OAIS compliant, DNA storage, DLCM technology, OLOS, digital repository

Conference Topics – Innovation; Environment.

I. INTRODUCTION

Sustainable storage and preservation are a major technical and financial challenge for digital data, regardless of the context of its creation and use (scientific, medical, administrative, banking, etc.).

At the current rate of production of digital data, it does not seem economically and ecologically viable to archive all that information over the long term with the available technologies, both in terms of energy and excessive consumption of raw materials. The longevity of DNA and its high storage density (in current practice from 2 PB/gram to 215 PB/gram [1], [2]) make this biomolecule extremely attractive for data storage applications [3], [4]. Current processes are in place to write and store digital information in DNA on the very long term (thousands of years) using relevant conditioning, (such as silica nano-beads or trehalose [5]–[7]), with the possibility to read it back without error thanks to high redundancy and robust error correction algorithms [3]. Random-access of files have also been demonstrated [8], and longevity was evaluated according to enhanced aging experiments [9].

Therefore, to ensure the preservation of human heritage, a DNA-based solution has many advantages: a high density of information despite the redundancy needed to maintain the integrity of the information; no energy needed to preserve the information; no obsolescence (DNA is part of the

building blocks of life). However, a critical next step will be to integrate the experimental work into archival environments that provide self-describing encapsulation of the data and its metadata, long-term integrity and authenticity, self-describing standards, and a study of the degradation of DNA libraries under real preservation conditions.

II. TECHNOLOGY

A. DLCM

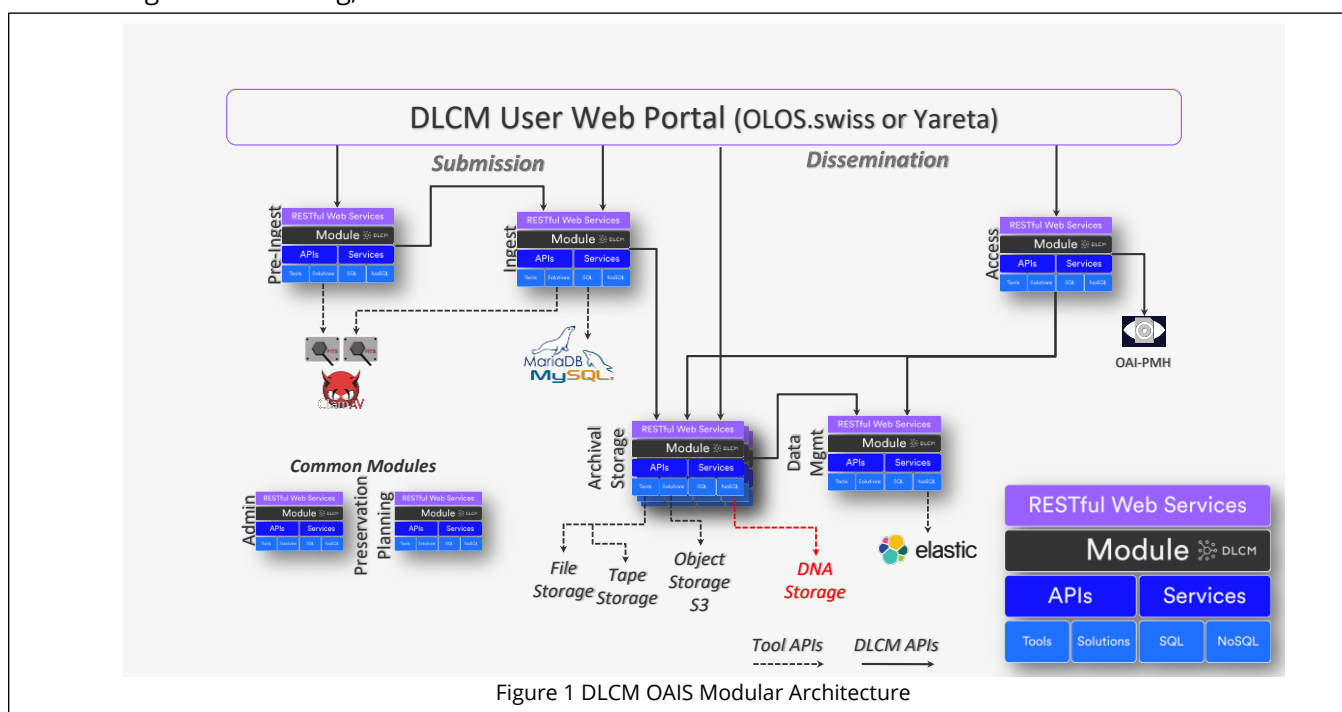
Within the framework of the Swiss national DLCM project (2015-2021) [10], [11], two implementations of long-term preservation systems (yareta.unige.ch and olos.swiss) have been realized in a perspective of safeguarding research and patrimonial information based on Swiss infrastructures. The main competitive advantage of the DLCM technology comes from its modular and distributed architecture, as well as its strict compliance with the OAIS reference model (Figure 1). To the three standard OAIS entities (SIP, AIP, DIP), we have added a pre-ingest module. This module allows for greater flexibility in data management by providing users with the ability to manipulate datasets prior to their final submission. The pre-ingestion comes after the phase of active work on the data, but before the archiving phase, which prevents any further modifications (with the exception of metadata). The architecture is based on PREMIS and DataCite for the metadata, DOI for the persistent identifier, OAI-PMH for indexing and harvesting, and various connectors

to different archive storage systems such as file system, S3, and tapes. For the next stage, we intend to add a new connector to store the information in DNA oligonucleotides. AIPs containing various types of digital data will be DNA encoded, with the DNA segments packaged in substrates that ensure slow degradation [5]–[7], then stored under controlled conditions and periodically checked for readability, which makes this project unique with respect to current technologies.

B. DNA

While DNA synthesis and sequencing is fast and well mastered, it remains to solve the question of finding optimal methods of encoding the information. Indeed, for the DNA segments to be accessed in a random-access mode, they must have an address that allows unique selection. The Polymerase Chain Reaction (PCR) sequencing method use primer sequences, which are short DNA segments at the beginning of a coding sequence, which play the role of this address [3]. However, these addresses must be mutually uncorrelated, so that it is unlikely that one address will be confused with another.

With the current technologies [4], it is possible to synthesize segments of about 1000 base pairs (bps), marked at both ends by specially designed sequences (PCR primers). Adding addresses to shorter segments result in significant storage overhead, while synthesizing blocks longer than



1000 bps is prohibitively expensive and more prone to errors. For that reason, we have chosen 200 bps as a compromise. Each 200-bps data segment is therefore appended at both ends with two unique primers of 20 bps each (one being the complement of the other). These addresses are used for the directed access to specific AIPs. Moreover, each DNA segment, representing a part of an AIP, is indexed (i.e., numbered) to be able to correctly rearrange the content during decoding (Figure 2).

Another aspect of encoding optimization is to provide a robust storage by relying on logical redundancy, both inside and between segments, allowing to decode information packages even in case of segment alteration or loss. Logical redundancy inside segments is referred to as inner code while redundancy between segments is named outer codes. This is necessary to compensate for DNA synthesis, DNA sequencing errors, PCR amplification biases, as well as DNA alteration over time. In practice, to achieve this, Reed-Solomon error correcting codes are applied over two dimensions with optimized parameters (Figure 2) [12].

III. IMPLEMENTATION

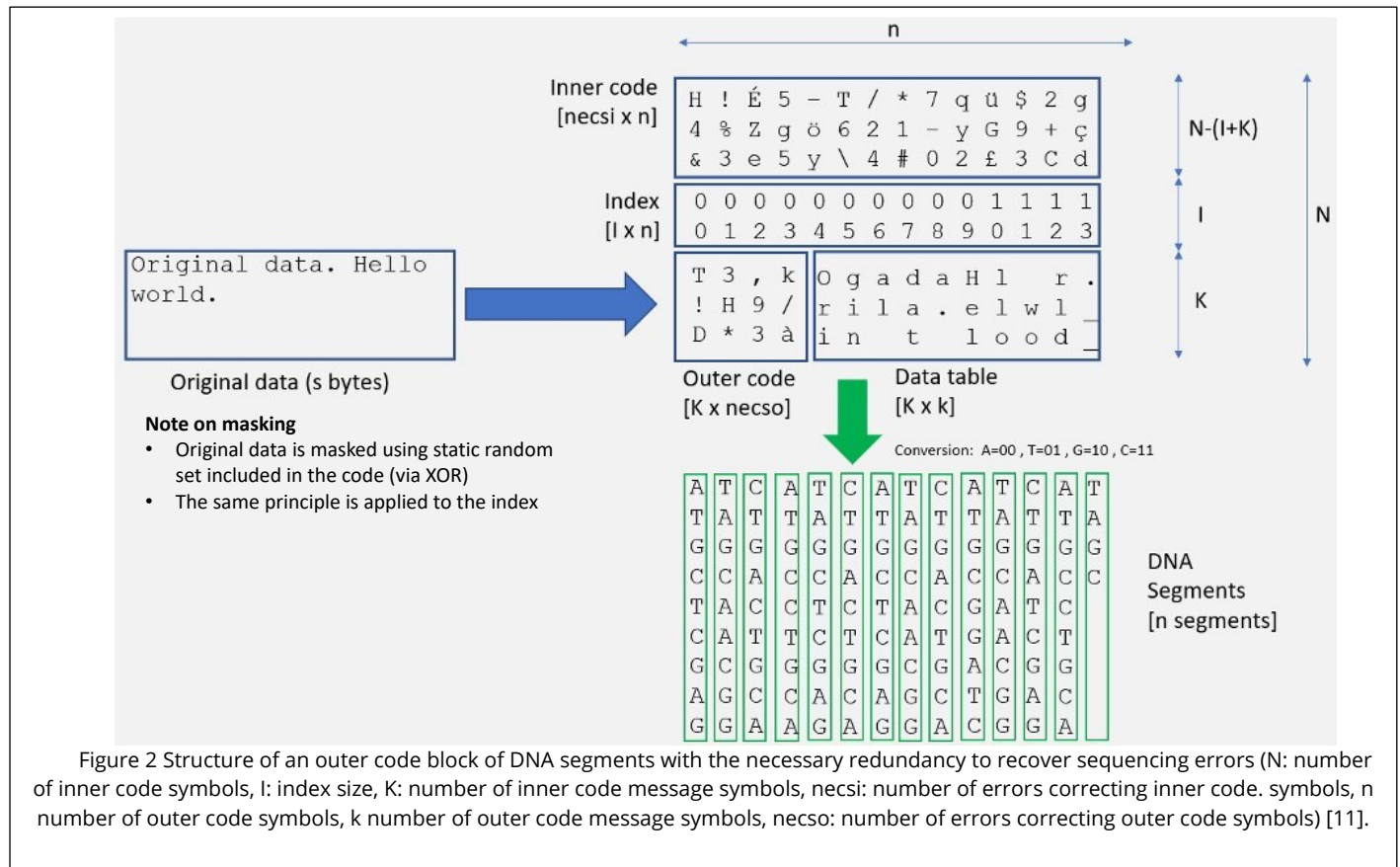
For designing the infrastructure, we extend state-of-the-art methods created and used previously to

store digital information in DNA (e.g., [3], [6], [10]). While previous work has proposed a self-contained DNA storage system that can bring self-explanatory to its stored data without relying on any external tool [13], we have chosen to make the process compatible with both the OAIS standard and the DLCM digital preservation system.

A. Conceptualization of the new data container architecture

Primers are keys to access and copy information stored into DNA, without having to sequence the whole DNA archive. We are developing a method to generate DNA primers to retrieve AIPs in a targeted manner consistent with OAIS identification requirements, i.e., a deterministic way to compute primers based on the digital preservation system AIP identifier. To support large AIPs, we will have to distribute them over many DNA segments, which requires management of DNA segments by blocks. Adding DNA as a novel medium to extend digital preservation systems will require a specific management of the AIP creation logistics with in particular:

- tracking the process from the creation of the DNA AIP until its storage in the archive;



- management of AIP physical storage, that is, location of physical copies: tube, shelf, room, building, etc., in a very similar way than paper preservation;
- management of AIP lifecycle, including events such as access or expiration dates.

B. Writing AIPs in DNA and Reproducibility

For the proof-of-concept, due for mid-2022, we will consider storing several small but valuable AIPs coded with the DLCM format (totaling about 1 MB) on the DNA media using the previously described data container infrastructure. This includes physical and logical organization of the AIPs into DNA data containers to copy and access AIPs, which can be controlled by DNA segments and sets of primers. We will also investigate the requirement for information redundancy to minimize costs while guaranteeing sufficient protection against errors to allow for read-out and accessibility even after very long time periods.

These steps will be followed by DNA synthesis (with about 500'000 copies of each data segment), packaging the resulting DNA segments in substrates that ensure slow degradation [5]–[7], and storage in tubes placed in a secured warehouse at the Cantonal Archives of Vaud, Switzerland. Readability of the AIP will be tested periodically during the project and beyond to assess more precisely the longevity of the media.

Following the OAIS recommendations, a detailed documentation of the whole process is necessary to ensure AIPs readability (lab protocols, calibrating parameters, algorithms, AIP structure, etc.), which will be stored on classical non-acid paper along with the DNA. We intend to verify the DNA readability under real conditions in 1, 2, 5, 10 and 20 years after encapsulation. This will consist in DNA amplification and decoding of files, which will be matched with checksums to assess the integrity of the extracted information. Error rates will also be evaluated at this stage to confirm that the redundancy mechanisms are correctly parametrized and the whole lifecycle (from AIP encoding in DNA to information retrieval) is operational.

IV. CONCLUSIONS

The DNA archiving approach offers a very promising green option for the management of digital data at reasonable costs, with less risk regarding the sustainable access to digital data.

Different methodologies of DNA storage have already been used to store music and historical documents, to demonstrate the feasibility of this technology, see for instance the Montreux Jazz Digital Project [14] and the deposit of digital archives encoded on DNA at the French National Archives [15]. However, the use of DNA to archive research and patrimonial information based on an OAIS architecture has not yet been done to our knowledge.

Different technologies allow DNA to be stored for decades or hundreds of years if kept at a temperature between 10 and 15°C [6], [7], [16]. With passive building cooling strategies, which are becoming increasingly common in the archival sector [17]–[19], such a temperature range can be achieved without consuming additional energy.

Given our proof-of-concept is due for mid-2022, we hopefully will be able to present the first results of our implementation during the iPres conference in September 2022. This innovative approach offers a new paradigm in data archiving for data scientists and archive professionals. Specific skills and expertise would be worth to be developed in this field.

REFERENCES

- [1] N. Goldman *et al.*, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, Feb. 2013, doi: 10.1038/nature11875.
- [2] J. Koch, S. Gantenbein, K. Masania, W. J. Stark, Y. Erlich, and R. N. Grass, "A DNA-of-things storage architecture to create materials with embedded memory," *Nat Biotechnol*, vol. 38, no. 1, pp. 39–43, Jan. 2020, doi: 10.1038/s41587-019-0356-z.
- [3] L. C. Meiser *et al.*, "Reading and writing digital data in DNA," *Nat Protoc*, vol. 15, no. 1, pp. 86–101, Jan. 2020, doi: 10.1038/s41596-019-0244-5.
- [4] L. Ceze, J. Nivala, and K. Strauss, "Molecular digital data storage using DNA," *Nat Rev Genet*, vol. 20, no. 8, pp. 456–466, Aug. 2019, doi: 10.1038/s41576-019-0125-3.
- [5] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes," *Angew. Chem. Int. Ed.*, vol. 54, no. 8, pp. 2552–2555, Feb. 2015, doi: 10.1002/anie.201411378.
- [6] D. Paunescu, M. Puddu, J. O. B. Soellner, P. R. Stoessel, and R. N. Grass, "Reversible DNA encapsulation in silica to produce ROS-resistant and heat-resistant synthetic DNA 'fossils,'" *Nat Protoc*, vol. 8, no. 12, pp. 2440–2448, Dec. 2013, doi: 10.1038/nprot.2013.154.
- [7] L. Organick *et al.*, "A empirical comparison of preservation methods for synthetic DNA data storage," *Synthetic Biology*, preprint, Sep. 2020. doi: 10.1101/2020.09.19.304014.

- [8] J. L. Banal *et al.*, "Random access DNA memory using Boolean search in an archival file storage system," *Nat. Mater.*, vol. 20, no. 9, pp. 1272–1280, Sep. 2021, doi: 10.1038/s41563-021-01021-3.
- [9] K. Matange, J. M. Tuck, and A. J. Keung, "DNA stability: a central design consideration for DNA data storage systems," *Nat Commun*, vol. 12, no. 1, p. 1358, Dec. 2021, doi: 10.1038/s41467-021-21587-5.
- [10] P.-Y. Burgi, E. Blumer, and B. Makhlof-Shabou, "Research data management in Switzerland: National efforts to guarantee the sustainability of research outputs," *IFLA Journal*, vol. 43, no. 1, pp. 5–21, Mar. 2017, doi: 10.1177/0340035216678238.
- [11] P. Y. Burgi and B. Makhlof Shabou, "Le projet Data Life-Cycle Management (DLCM) en Suisse : une gestion des données de la recherche pensée pour ses utilisateurs," *I2D - Information, données & documents*, vol. n° 2, no. 2, pp. 87–95, Nov. 2021, doi: 10.3917/i2d.212.0087.
- [12] Jan, Krause-Bilvin, *archive2dna*. 2022. Accessed: Feb. 22, 2022. [Online]. Available: <https://github.com/jbkrause/archive2dna>
- [13] M. Li *et al.*, "A self-contained and self-explanatory DNA storage system," *Sci Rep*, vol. 11, no. 1, p. 18063, Dec. 2021, doi: 10.1038/s41598-021-97570-3.
- [14] K. Osborne, "All that jazz: Researchers preserve iconic musical performances in DNA." Allen School News, Sep. 29, 2017. [Online]. Available: <https://news.cs.washington.edu/2017/09/29/all-that-jazz-researchers-preserve-iconic-musical-performances-in-dna/>
- [15] P. Petit, "Stocker pour 50 000 ans : des textes historiques sur ADN entrent aux Archives nationales." France Culture, Nov. 24, 2021. [Online]. Available: <https://www.franceculture.fr/sciences/stocker-pour-50-000-ans-des-textes-historiques-sur-ADN-entrent-aux-archives-nationales>
- [16] K. Washetine *et al.*, "DNAShell Protects DNA Stored at Room Temperature for Downstream Next-Generation Sequencing Studies," *Biopreservation and Biobanking*, vol. 17, no. 4, pp. 352–354, Aug. 2019, doi: 10.1089/bio.2018.0129.
- [17] T. Padfield, M. Ryhl-Svendsen, P. K. Larsen, and L. Aasbjerg Jensen, "A Review of the Physics and the Building Science which Underpins Methods of Low Energy Storage of Museum and Archive Collections," *Studies in Conservation*, vol. 63, no. sup1, pp. 209–215, Aug. 2018, doi: 10.1080/00393630.2018.1504456.
- [18] K. Kompatscher, R. P. Kramer, B. Ankersmit, and H. L. Schellen, "Indoor Airflow Distribution in Repository Design: Experimental and Numerical Microclimate Analysis of an Archive," *Buildings*, vol. 11, no. 4, p. 152, Apr. 2021, doi: 10.3390/buildings11040152.
- [19] B. Bøhm and M. Ryhl-Svendsen, "Analysis of the thermal conditions in an unheated museum store in a temperate climate. On the thermal interaction of earth and store," *Energy and Buildings*, vol. 43, no. 12, pp. 3337–3342, Dec. 2011, doi: 10.1016/j.enbuild.2011.08.028.

DNA DATA STORAGE FOR LONG TERM DIGITAL PRESERVATION

Euan Cochrane

*Yale University Library
United States
euan.cochrane@yale.edu
[0000-0001-9772-9743](tel:0000-0001-9772-9743)*

Daniel Chadash

*Twist Bioscience
United States
dchadash@twistbioscience.com
[0000-0002-9712-6034](tel:0000-0002-9712-6034)*

Abstract – Long term digital preservation can benefit from long term storage solutions. DNA data storage is a new technology that offers unique benefits to the digital preservation community. With the cost of DNA data storage rapidly decreasing Yale University Library has partnered with Twist Bioscience to investigate the benefits and feasibility of deploying DNA data storage for long term preservation. In this paper we discuss DNA data storage, outline the pilot project we have undertaken and discuss the technology and its potential future applications.

Keywords – DNA, Storage

Conference Topics – Innovation, environment.

I. INTRODUCTION

In our increasingly volatile geopolitical world those of us working to preserve the cultural, scientific, and spiritual knowledge and records of our global societies in digital form have increasing justification to look for methods of storing these digital artifacts in ways that would outlast long periods of environmental and geopolitical volatility. DeoxyriboNucleic Acid (DNA) data storage technology provides an increasingly practical, sustainable, durable, and cost-effective method for mitigating many of the potential risks to digital storage infrastructure that seem increasingly likely to become realized as destructive issues in upcoming years.

In addition to DNA data storage being a reliable, durable and long lasting medium, media diversity is very important in any long-storage digital preservation strategy that can significantly de-risk the chances of data loss.

DNA data storage in its current form is a type of 'Write Once, Read Forever' digital storage media (WORF) [1]. WORF media is so-called because it can

only be written once (cannot be edited) and once written will survive and be accessible 'forever' or at least for a hugely greater time than standard digital storage media. Other digital storage media described in the same way include:

1. M-DISC optical media - with a claimed lifespan of '1000 years' [2]
2. PIQL Film technology - Film-based storage using binary data stored as Q R codes on the film, with an expected lifetime of over 500 years [3].
3. Digital Optical Technology System (DOTS) - "DOTS physically encodes data on an archival tape coated in a phase-change alloy" and has a claimed lifetime of over 100 years [4].

Using DNA data storage for long term bit-preservation is particularly attractive in the context of these other available options for a number of reasons. Specifically DNA data storage has unique qualities that make it attractive compared to these alternatives, including:

1. While future post-disaster humans may never re-develop (for example) a blu-ray reader, no matter how significant the disaster that could befall humanity, surviving humans ought to eventually want and need to read their DNA once more. In doing so this will provide the ability to unlock a potential deluge of valuable knowledge from the past.
2. DNA data storage is extremely information dense. This makes it very suitable for use in time capsules and data caches that could be deliberately hidden to prevent their destruction in times of volatility. Furthermore the relatively tiny [Figure 1] physical size of current DNA data storage

containers make them less likely to be affected by any sort of physical event simply due to their minimal attack surface.



Figure 1 a DNA data storage capsule
Image courtesy of Imogene SA

The small physical size also means that the cost of dispersing copies of data stored as DNA is relatively low. This may potentially enable more varied and effective storage-risk mitigation strategies to be implemented at the same cost as it would be to implement a higher risk strategy using larger, more expensive media.

3. DNA data storage is increasingly cost-effective. Twist predicts that the cost will reduce to less than \$1000 per terabyte within the next few years and the cost could be further reduced in the future. Equally relevant, once a set of data has been synthesized, the cost of additional copies of the dataset is very insignificant and mostly dependent mostly upon the cost of the (inexpensive - under tens of dollars each) physical container used to encapsulate the DNA material and the low cost of reagents to copy the DNA. This has wide implications for storage-risk mitigation strategies as it allows for significant physical redundancy to be implemented in storage strategies that use DNA data storage. For example, many additional copies of the same data may be able to be placed in locations with wide variations in their risk profiles at minimal additional cost.
4. DNA data storage has significant redundancy built into it which provides the opportunity for very effective error correction functionality to be included in the implementation of the data-to-DNA

synthesis and recovery/reading processes. This means that it is far less likely that a single 'copy' of data stored as DNA will not be able to be read at any point in the future¹.

5. Unlike some alternative options² DNA has already been proven to last for a very long time in the real world [5]. In addition, existing examples of readable DNA samples have been recovered from physical contexts that are extremely far from the ideals that can be achieved using the encapsulation technologies we have available today. Such technologies coupled with a deliberate plan for locating the physical media in many low-risk locations make data stored as DNA extremely likely to be accessible in the future.
6. Environmental Benefits
 - a. DNA data storage only uses energy when being read or written rest. At rest it requires no energy consumption It can be stored at room temperature and has no no active cooling requirements.
 - b. Physically little material is necessary to store large volumes of data. A room of DNA would likely easily store all the data that exists in the world in 2022.
 - c. The technology consists of very easily recyclable components. Just metal and organic material are required and the capsules can also be washed and reused.
 - d. The sourcing of materials is easier than alternatives. For example there is no need for rare metals that might necessarily be sourced from conflict zones.
7. Offline storage benefits - DNA data storage can be kept offline (disconnected from the public internet or local intranet). This enables a higher level of security as for most risk profiles it prevents various threats such as ransomware attacks.

II.VALUE IN LONG TERM DIGITAL PRESERVATION

¹ 'Copy' is not entirely accurate here as each DNA data storage capsule includes a huge number of copies of the data being stored.

² Though not all - e.g. the PIQL technology uses time-proven film media for its physical storage mechanism.

WORM media has a long history in digital preservation³. The community has never fully embraced it for a number of previously justifiable reasons. Amongst those reasons are:

- a. The need to be able to make changes to preserved data or its metadata over time.
- b. The need to check datasets for integrity over time, and the perceived need to continually change the digital objects over time so they work in current technologies (i.e. to undertake migration of content between files of different formats).
- c. Digital preservation also continues to be a relatively nascent field that has not been around long enough to verify vendor claims of longevity. This has meant a natural lack of trust exists in unproven solutions
- d. From an historic perspective the nascent digital preservation community has been operating in a period of relative global stability, especially in the western world in which most of the practitioners operate. WORM media is most easily justifiable as a hedge against loss due to instability. Without the risk of such threats it has been harder to justify the cost of mitigating against them by using WORM solutions.

For these reasons, to date, the digital preservation community has not had any compelling justification for implementing WORM media. However this may be changing.

Emulation solutions have become more widely used [7] and provide a novel solution for ensuring future generations can decode digital files into meaningful information. As opposed to migration, as a solution for ensuring long term access to content in preserved digital files, emulation can be implemented in a way that requires minimal or no change to stored digital files over time. Emulators, and the full computing environments needed to access stored digital files, can be stored using the same mechanisms as the digital files themselves (i.e. in this case using DNA data storage). Over time if an emulator becomes incompatible with current technology a new one can be written to replace the old emulator, and stored alongside the old ones, or a new emulator can be created into which the old one can be nested. In either case, the primary digital

objects being preserved do not need to be altered at all, making write-once media feasible.

Long Term Preservation Challenges of DNA Data Storage.

As discussed, emulation offers a practical option for ensuring future users can access digital objects stored using DNA data storage without requiring the digital objects be regularly replaced. However neither emulation nor migration can mitigate against large gaps in time between storing objects and future users trying to access them, during which time the computing technology (and related knowledge) required may have been lost. For such a scenario we will need additional infrastructure in place to bootstrap future generations to the point where they can make sense of binary data (for example there is an attempt to create a “Manual for [rebuilding] civilization”[8]).

III. TURNING DNA DATA STORAGE INTO A WRITE-MANY SERVICE

While this article has focussed on the use of DNA data storage as a WORM technology Twist has plans in place to offer an implementation of the technology that would make it functionally similar to re-writable media. In the planned configuration DNA data storage will serve a similar role to archivally-configured cloud storage like AWS Deep Glacier [9] and can offer a new cold layer in the archival tiers.

Current large scale tape based digital storage solutions generally involve the use of a central management server with local higher speed cache storage, one or more tape drives, a large storage rack to store tape cartridges, and a ‘tape robot’ that can fetch cartridges and insert them in the drive to be read and their data cached to the local server for end-user access. A similar approach is used for large scale blu-ray storage solutions [10]. The difference between those two however is that blu-ray disks are generally implemented as write-once media meaning edits are not possible and deletions require destroying entire discs. The blu-ray based approach is currently the most similar to that which is planned for the future utilizing DNA data storage. Using the new approach it will be possible to implement automated machines to synthesize data as DNA, and to read the data back from the synthesized DNA before re-synthesising the data with changes and storing the DNA back in it’s storage containers.

³ See the summary section in “Site visit report #1” describing WORM use in 1988 in [6]

Implementing a process like this will effectively turn DNA data storage into a practical automated medium/long term storage mechanism.

Unfortunately the current cost of synthesizing and reading DNA data are both too high for practical use in this way [11]. In addition the time taken for both activities is prohibitively high such that it would not be practical for usage in active-archive scenarios where access time and frequency are important. However it might be within the acceptable range for becoming a competitor for offline tape storage or whatever technology is backing the deep-archive solutions offered by cloud vendors such as Amazon's 'Glacier Deep Archive', which currently only guarantees retrieval within 12 hours.

An even simpler potential use for DNA data storage for shorter term (less than 'forever') storage requirements will arise if the cost for synthesis/storage reduces even more dramatically. If the cost of DNA data storage reaches a level significantly less than that of other options it will become viable to use it as an additional risk-mitigating back-up copy. Used in this way, edits to stored data would simply involve deletions and inexpensive re-encodings of changed data, and partial-deletions would be similar.

IV. THE DNA DATA STORAGE PILOT PROJECT AT YALE UNIVERSITY LIBRARY

The Yale University Digital Preservation unit provides digital preservation services across the libraries, archives and museums on campus. The services are used to preserve a huge variety of collections ranging from the Fortunoff Holocaust Testimonies, to architecture records, to video games, and more [cite]. Most of Yale's preserved collections are unique and irreplaceable and as such are important to the global historic record for informing future generations. The Digital Preservation Services (DPS) team is organized as an internal service provider and works with collection owners to provide services so that they can preserve their content. As part of these services DPS provides multiple options for storing digital content in ways that have different risk profiles associated with them (different sets of specific risks that may lead to data being lost, with different likelihoods associated with them). DPS does not decide the storage options that are used as part of the bit preservation strategies employed to preserve Yale's collections. Instead DPS discusses the options with collection owners and they decide what option best fits their risk appetites and budgets.

As part of providing these services the DPS team has recently been investigating options for offering WORM media offerings to its users. The DPS team expects that at least initially only the most high-value collections will find these options attractive due to the cost, however this may be naive, especially if costs continue to decrease as predicted by some service providers.

In 2021 the DPS team began a pilot project to test the DNA data storage technology offered by Twist Bioscience. The team's goals in the pilot were:

1. To learn about the technology, e.g. how it works, how feasible it is to implement.
2. To test the technology over time
3. To socialize the idea of DNA data storage and seek feedback from our stakeholders about their interest in implementing it
4. To use this interesting new technology to raise awareness about digital preservation in general.

The pilot involved storing approximately 15 megabytes of data at a cost of nearly \$1000 per Megabyte. The DPS team worked with librarians at the Harvey Cushing/John Hay Whitney Medical Library to select sample data that was open, relevant, and interesting. The hope was to include data that could be publicly shared in publications used for awareness raising about the project and which would attract interest.

The data was packaged into a single .zip file as the current implementation of the DNA data storage technology does not include a filesystem, preventing multiple files being stored in the same DNA sequence. Following this the data was passed on to Twist Bioscience (Twist) to be converted from binary to a DNA sequence. After this the DNA sequence was synthesized, replicated, and stored in 40 capsules. The DPS team requested 40 capsules for two purposes. Firstly the team intends on undertaking test-reads of the data in the capsules at regular intervals, and so needed multiple copies to support this. Secondly some of the capsules will be given away as keepsakes as part of awareness raising activities.

The DNA synthesis process was relatively short and the DPS team received the capsules with the data within a few weeks of sending off the file for synthesis. Following this the team reached out to researchers in the Yale Center for Genome Analysis (YCGA) who offered to conduct an initial reading of the DNA in one of the capsules. This process was

undertaken using standard DNA sequencing technology that the YCGA researchers use on a regular basis. After sequencing the DNA the resulting DNA sequence had to be converted back to binary data. This process was undertaken using a virtual machine provided by Twist, containing the decoding algorithm and software implementation. Converting the sequence to binary resulted in the original file that had the original checksum associated with it. The DNA Data Storage Alliance is planning to develop an industry standard encoding and decoding algorithm and implementation. These are intended to be open source and freely available.

The DPS team intends to replicate this reading process in (at a minimum) 2, 5, and 10 years time and report the results publicly in order to provide a set of benchmark data for others in the digital preservation community to learn from. We (Twist and the DPS team) are also planning to undertake a second pilot in 2022 to take advantage of the new iteration of the storage technology that has an estimated cost of US\$1000/GB, which is an improvement of three orders of magnitude of the cost.

V. THE DNA DATA STORAGE TECHNOLOGY

Storing data on DNA is not a new concept. It was first demonstrated a few decades ago.

The dominant workflow today for storing and reading data on DNA is built from 6 steps:

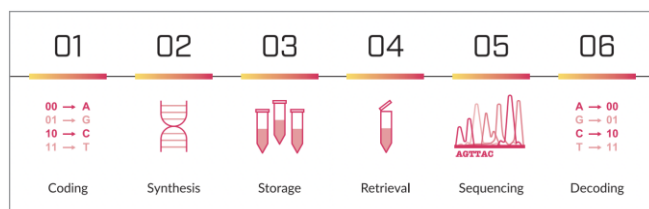


Figure 2 The DNA Data Storage Workflow

1. **Coding** - Any digital file at its base is built from 1s and 0s, using an encoder we translate the 1s and 0s into AGTC (the building block of DNA), and divide the long string of letters into short sequences so it will fit current synthesis technologies. In addition we add metadata and error correction codes to help us with reading the data back in case of sequencing errors.
2. **Synthesis** - The short sequences are sent to synthesis, where we translate the text strings of ATGC into real physical DNA using a DNA "printer".

3. **Storage** - The DNA comes out of synthesis in liquid form, it is dehydrated for long-term storage purposes and is encapsulated and sealed in capsules for storage.
4. **Retrieval** - Once access to the data is needed, the capsule is opened and the DNA is rehydrated and prepared for sequencing/reading.
5. **Sequencing** - The liquid DNA is read using a device called sequencer. The sequencer transforms the physical sample into a digital form of the short strings of DNA letters.
6. **Decoding** - Using the codec that was used in step 1, the decoder is able to reconstruct the digital binary file from the short DNA sequences using the metadata in each short string and is able to correct errors that originated from the sequencing process.

This workflow is mainly deployed in demonstrations and proof of concepts of the technology through the past decade mainly because of the cost of DNA synthesis, as there was no new technology to enable a drastic cost reduction.

Twist Bioscience was founded in 2012 with a breakthrough technology for synthesizing DNA. Using the advancements in chip and silicone manufacturing Twist was able to synthesize DNA on a silicon chip that allowed miniaturization of the components that enabled the lowest price point to date for custom DNA. A new generation of that original chip was used for the first part of the pilot at Yale University Library which enabled the first commercial offering of DNA Data Storage at \$1000/MB.

In order to reach commercial viability, there still needs to be an improvement of a few orders of magnitude of the cost.

Scaling the Technology

The technology that was used in the first part of the pilot is the same one that is being used to synthesize DNA for a broad range of applications in the healthcare and biotechnology industries. Very high quality DNA, without many errors that is suited for applications that can't tolerate a moderate error rate.

The advantage of using DNA for data storage is that we can incorporate many well-proven algorithms from information/communication theory that allows us to recover the data even if there are errors in the synthesis/sequencing processes, in the same way

that our cellphones are able to compensate on noise/error in the cell signal.

This property allows us to develop new technologies that are specific for DNA data storage and will allow a massive cost reduction. Most of the cost reduction will come from miniaturizing the features and building more reagent-efficient chips and technologies.

The second chip (POC chip) that will be used for the second part of the pilot is a DNA data storage dedicated chip that was developed by Twist specifically for that purpose and shows a staggering improvement of three orders of magnitude in the cost (from \$1000/MB to ~\$1000/GB).

As described in the figure below, Twist is not stopping at that price point and plans to continue with the cost reduction until it will be able to compete with prices of current archival storage mediums (tape and HDD).

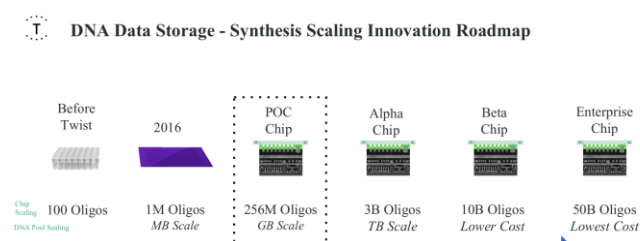


Figure 3 The DNA data storage synthesis scaling innovation roadmap.

Without synthesis nothing further would be possible with DNA as a data storage technology. However it's worth noting that there are more challenges down the road, and we focused here mainly on the synthesis as it's the main enabler for the successful large scale use of DNA as a long term storage technology.

In addition to lowering the cost of synthesis there needs to be a drastic reduction in the cost of sequencing (few orders of magnitude) as without it customers won't adopt the technology outside as a last-resort backup option.

Automation and scalability are also very important factors. Any technology that would be deployed in the data center needs to be highly automated and provide the same ease-of-use as current technologies and important features that other technologies provide today for archival applications.

VI. CONCLUSION

DNA data storage offers an increasing practical option for the long term storage of digital data. The

successful pilot project trialing DNA data storage at Yale University offers a promising window into this new technology. With costs consistently decreasing DNA and new implementations in development that will turn DNA data storage technology into an effectively write-many media, we expect to see an increasing uptake in its use for long term preservation storage.

REFERENCES

- [1] *Write Once, Read Forever (WORF)—Proof-Of-Concept Demonstrated For Archival Data Storage Using Interference Spectra*. Solomon, R. et al. Archiving 2015 Final Program and Proceedings. 2015. <https://www.ingentaconnect.com/contentone/ist/ac/2015/0002015/00000001/art00023>
- [2] *Accelerated Life Cycle Comparison of Millenniata Archival DVD*. I. Svrcek. Naval Air Warfare Center Weapons Division. 2009. https://web.archive.org/web/20210211221229/https://www.esystor.com/images/China_Lake_Full_Report.pdf
- [3] *PIQL Technology website*. PIQL. Retrieved 6th March 2022. <https://www.piql.com/about/technology/>
- [4] *DOTS—Long-Term, Human-Readable Archival Data Storage*. Warner, A. The Long Now Foundation. 2015. <https://web.archive.org/web/20220123123847/https://longnow.org/ideas/02015/12/27/dots-long-term-human-readable-archival-data-storage/>
- [5] *Million-year-old DNA sheds light on the genomic history of mammoths*. van der Valk, T., Pečnerová, P., Díez-del-Molino, D. et al. *Nature* 591, 265–269 (2021). <https://doi.org/10.1038/s41586-021-03224-9>
- [6] *Digital-Imaging and Optical Digital Data Disk Storage Systems*, The Technology Research Staff. The National Archives at College Park. 1994. Retrieved 6th March 2022. <https://www.archives.gov/preservation/technical/imaging-storage-appendix.html>
- [7] *Emulation as a Service Infrastructure (EaaS)*. EaaS team. 2018. Retrieved 6th March 2022. <https://www.softwarepreservationnetwork.org/emulation-as-a-service-infrastructure/>
- [8] *How Can We Create a Manual For Civilization?*. A. Kabil. The Long Now Foundation. 2017. <https://longnow.org/ideas/02017/06/07/how-can-we-create-a-manual-for-civilization/>
- [9] *New Amazon S3 Storage Class – Glacier Deep Archive*. J. Barr. Amazon Web Services. 2019. Retrieved 6th March 2022. <https://aws.amazon.com/blogs/aws/new-amazon-s3-storage-class-glacier-deep-archive/>
- [10] *Inside Facebook's Blu-Ray Cold Storage Data Center*. R. Miller. Data Center Frontier, 2015. Retrieved 6th March 2022. <https://datacenterfrontier.com/inside-facebooks-blu-ray-cold-storage-data-center/>
- [11] *Preserving our digital legacy: an introduction to DNA data storage*. The DNA data storage alliance. 2021. Retrieved 6th March 2022. <https://dnastoragealliance.org/dev/wp-content/uploads/2021/06/DNA-Data-Storage-Alliance-An-Introduction-to-DNA-Data-Storage.pdf>

THE DESIGN AND IMPLEMENTATION OF A NECESSARY AND SUFFICIENT SYSTEM FOR THE LONG-TERM ARCHIVAL RETENTION OF DIGITAL DOCUMENTS

Viv Cothey

*Gloucestershire County Council
UK*

Claire Collins

*Gloucestershire County Council
UK
claire.collins@gloucestershire.gov.uk*

Abstract – we describe how the design of a digital preservation system suitable for the long-term archival retention of digital documents mimics conventional archival practice with regard to provenance, authenticity and workflow. The design further ensures the evidential nature of digital “documents” in the Archive. Exit plans recognize both the routine expiry of time limited supplier agreements and the effects of a disorderly supplier exit. The design has been successfully implemented and is now being migrated from test to production as a “business as usual” component of the archivist role.

Keywords – Archive, Provenance, Authenticity, Fixity

Conference Topics – Resilience; Community

I. INTRODUCTION

This paper emphasizes that the conventional underpinnings of archival procedures and practices as seen in (English) local authority record offices should apply also when accepting custodial responsibility for digital documents.¹ We describe how the experience of ensuring the continuing evidential quality of physical documents has informed the design and implementation of a digital preservation system by concentrating on provenance and authenticity.

The perceived role and purpose of the Archive has evolved as evidenced by The National Archives (TNA) reporting now to the Department for Digital,

Culture, Media and Sport. Although being undoubtedly a major part of the heritage sector the Archive is much more than a library of old stuff. Schellenberg carefully distinguishes between the organization, operation and management of a library and that of an Archive [1]. These distinctions are readily apparent when comparing the technology of library catalogues and their standardized bibliographic records and classification schemes with the hierarchical arrangement of archival descriptions [2] and institutionally devised individual arrangements. A significant area of contrast is that, mostly, a library is a collection of published documents whereas an Archive comprises documents that have not been published. The Archive receives documents from its parent body which evidence the bureaucratic systems and decision making which are the activity of that parent. A further feature is that access to many archival documents is restricted, that is, the documents are “closed”.

Whereas the lack of demand for an item may be a cause of concern to a library, for an Archive it is to be expected. Most archival documents will never be requested. However what is important is that every document has the potential to be accessed, possibly by users yet unborn. The existence or lack thereof of a document in the Archive can have life changing consequences for the individual.

¹ “document” is here used in its broadest possible sense to refer to an identifiable unit, that is, regardless of form or extent

Jenkinson [3] asserts that the archivist has a duty to both preserve its intellectual properties, for example how the document relates to other documents, and to protect the physical document. Taken together these two duties can be summarized as preserving the provenance and the authenticity of the document. We know where and why the document was created and that the document is authentic, that is, it remains to be what it purports to be.

Archival procedures and practice address these duties by maintaining archival “provenance” and the archive “strongroom”. Archival provenance is established and maintained by the descriptive entries in the hierarchical catalogue. Access to documents is strictly controlled; they are stored under lock and key in environmentally controlled vaults [4]. Also, document access is mediated including acclimatization and other conservation procedures as necessary.

A special property of the Archive is the evidential nature of its documents. Investigations into the behavior of institutions or the need to revise previous decisions often rely upon the demonstrable authenticity of the document. The information retained and organized in Archives protects people and has legal force. It is not an exaggeration to say that users trust the integrity of information managed by archivists and rely upon it “to hold government and organizations to account” [5]. In a similar vein, Procter [6] says,

*“[Archivists] are often unaware of... the way in which the characteristics of archives – an ability to provide information and evidence and **sustain rights** – have provided, and continue to provide, the rationale for their maintenance over time.”* [emphasis added]

(p xv)

However simply producing an authentic document may not be enough to access its meaning. Many older documents employ either or both an archaic style or language, and an archaic script. Paleography is the study of the script but even a modern transliteration does not remove the need to understand an archaic usage. Words and phrases change their meaning over time and there are traps for the unwary [7].

The design for a system to support the archival retention of digital documents by Gloucestershire

County Council (GCC) is based on the need to preserve provenance and authenticity.

II. PREVIOUS WORK

Retaining digital information over the long-term Cothey [8] proposes a system architecture to preserve both provenance and authenticity information. The paper introduces the notion of “authentic preservation” which entails the known survival of a digital document. Authentic preservation requires that a) information, including provenance, must survive, b) surviving information must be authentic, and c) authenticity can be demonstrated.

The ‘Archives First’ consortium’s report Further investigations into digital preservation for local authorities [9] documented relevant digital preservation issues and options. In particular the report proposes a long-term authentic preservation architecture that is based on a sequence of interlinked short-term authentic preservation systems. The report therefore draws attention to the need to manage exit plans to successfully transfer curated documents to successor systems.

III. DESIGN

The scope of the design is limited. It is assumed that co-lateral digital information hygiene, such as information security, disaster recovery (DR) and business continuity plans are in place and are regularly tested. The main co-lateral threats manifest suddenly and unexpectedly. In contrast the main threats to a long-term retention system are gradual and expected. An important exception here is supplier failure giving rise to a disorderly exit.

The system is considered to comprise two components. The first is “operational”, that is, it forms part of the day to day operational IT of the Archive. Information is dynamic and frequently modified. Like many local authority record offices the operational IT is managed through an outsourced facilities management contract which has a time limited duration.

The second component is the storage of the digital documents, stored information is static and infrequently accessed. Like the operational IT this store is provided and managed via a time limited outsourced facilities contract.

In addition the system workflow is integrated with archivists' conventional workflows in order to support a business as usual approach.

A. Preserving provenance

The preservation of provenance relies on the continued existence and accessibility of the Archive's hierarchical catalogue. This is threatened by the gradual obsolescence of supporting technologies and by the need to migrate or roll-over the proveniential information at the expiration of a management contract. Neither of these threats are mitigated by DR but must be managed through an exit plan. Supplier failure represents a disorderly exit and plans here cannot assume any supplier support. As with DR, exit plans must be regularly tested.

All archival descriptive metadata for digital documents is maintained by this catalogue.

The design response is for the hierarchical catalogue to be ISAD(G) compliant and to ensure that frequent system agnostic information exports are generated. Exported information can be "round tripped" or re-imported to simulate a recreation of the hierarchical catalogue. Importantly this is designed to be achievable without any support from the catalogue provider.

B. Preserving authenticity

Digital documents are uniquely fragile. Access to their information is based on reading a stored bitstream which can become corrupted. A demonstration of a lack of corruption is a consistent cryptographic hash or message digest of the bitstream. This is known as a fixity value for the document in question. Different cryptographic algorithms provide supplementary digests.

A sufficient long-term fixity management system is the principal design challenge facing the archival system. The question of fixity arises three times, firstly when the Archive deposits the document into the store, secondly when the Archive requests the document from the store and thirdly when monitoring the integrity of a particular store. A particular instance is verifying every stored document when managing an exit.

As identified above, the Archive protects the authenticity of its physical documents by keeping them under lock and key. For digital documents an equivalent is to maintain a copy of relevant digests

independently of the storage supplier. These are retained within the operational IT and are thereby covered by appropriate business continuity arrangements. The three fixity questions are addressed by;

1) on initial deposit the store returns two or more digests which the Archive compares with digests computed independently. These are retained in an operational fixity database.

2) on request the Archive computes two or more digests for the returned document and compares these with the digests retained in the fixity database.

3) periodically the storage system provides digests for a selection of stored documents which the Archive compares with the fixity database.

IV. IMPLEMENTATION

During early 2021 GCC issued a request for proposals in respect of a storage fixity manager (SFM) having the design features described above [10]. This complemented a similar procurement to implement an ISAD(G) compliant catalogue also with features as described above. The catalogue is now in production and exit plans are being tested. Here we present the implementation of the SFM.

The SFM has been implemented as a Web based service that mediates interactions with multiple cloud storage providers in order to eliminate single provider vulnerabilities and to support sustainable storage provider exit planning. Cloud based commodity storage is used in order to benefit from both economies of provision and geo-diversification. In addition to "upload" and "download" the service provides progressive reporting of current fixity values.

This service is complemented by an independent (open source) desktop digital curation application [11] used by the archivist to both interact with the SFM and to access the operational fixity database. Archivist workflows replicate practice in respect of non-digital documents when accessioning documents and storing them in strongrooms – our business as usual approach. In particular the desktop application also supports the creation of OAIS archival information packages and dissemination packages.

V. CONCLUSION

The successful implementation of the ISAD(G) catalogue, storage fixity manager and desktop digital curation application provides the necessary attributes of an authentic preservation system that includes planning and testing the management of disorderly exits. This authentic preservation system is also sufficient. It mimics existing procedures and practice, in particular mediating access to documents and for working with closed documents. It is anticipated that any future challenges when rendering authentic bitstreams will be addressed by paleographers skilled in archaic digital formats as well as archaic scripts.

DISCLAIMER

The views and opinions expressed in this paper do not necessarily represent those of the institutions to which the authors are affiliated.

REFERENCES

- [5] Schellenberg T. R. Modern archives: principles and techniques. [Midway reprint] Chicago: University of Chicago Press. 1956.
- [6] ISAD(G). (1999). ISAD (G): general international standard archival description: adopted by the Committee on Descriptive Standards, Stockholm, Sweden, 19-22 September 1999.
- [7] Jenkinson H. A manual of archive administration: new and revised edition. London: Percy Lund, Humphries & Co. 1937.
- [8] The National Archives. A guide to environmental management of archival material. <https://cdn.nationalarchives.gov.uk/documents/information-management/environmental-management.pdf>. 2017.
- [9] The National Archives. Consultation on a new strategic vision for the archives sector. 2016.
- [10] Procter M. Introduction to the English edition. In Delsalle P. A history of archival practice [Translated and revised by Margaret Procter] London: Routledge. 2018.
- [11] Shoemaker R. Why Naomi Wolf misinterpreted evidence from the Old Bailey Online. <http://www.historymatters.group.shef.ac.uk/naomi-wolf-misinterpreted-evidence-bailey-online/>.
- [12] Cothey V. Retaining digital information over the long-term. <https://www.gloucestershire.gov.uk/media/2087704/retainin-g-digital-information-over-the-long-term.pdf>.
- [13] Cothey V. Archives First: digital preservation: further investigations into digital preservation for local authorities. <https://www.gloucestershire.gov.uk/media/2094490/digital-preservation-for-local-authorities.pdf>.
- [14] Forbes H. and Collins C. Request for proposals. <https://www.gloucestershire.gov.uk/media/2113461/gloucestershire-archives-request-for-proposals.pdf>.
- [15] Cothey V. SCAT is Curation And Trust. Available from claire.collins@gloucestershire.gov.uk. 2022.

EMA: BRAZILIAN CULTURAL HERITAGE IMAGE DATASET

Towards AI-based metadata annotation of digital collections

Vagner Inácio de Oliveira

University of Campinas
Brazil
vagner.inol@gmail.com
[0000-0002-0947-2990](tel:0000-0002-0947-2990)

Dalton Martins

University of Brasília
Brazil
daltonmartins@unb.br
[0000-0002-6244-6791](tel:0000-0002-6244-6791)

Paula D. Paro Costa

University of Campinas
Brazil
paulad@unicamp.br
[0000-0002-1534-5744](tel:0000-0002-1534-5744)

Abstract – Metadata annotation in digital collections is typically conducted by several specialized professionals, configuring a complex, labor-intensive, and time-consuming activity, leading to human failure, high costs, and problems in retrieving information accordingly. Recent advances in artificial intelligence, particularly Deep Learning techniques, have shown their potential in performing visual recognition and interpretation of objects on images. In this context, the present work introduces EMA, a Brazilian cultural heritage image dataset with over 11,000 labeled images of objects from seventeen Brazilian museums. EMA dataset is a contribution towards the development of automated metadata annotation tools. The paper also presents baseline ResNet50 results for the dataset, resulting in an over 86% recognition rate.

Keywords – Digital Cultural Heritage, Thesaurus, Automatic Annotation, Deep Learning, Computer Vision
Conference Topics – Innovation; Resilience.

I. INTRODUCTION

Digital collections are a powerful way to open museums' cultural heritage to exploration by the public. They are particularly relevant in a country like Brazil, where museums that preserve the country's history are thousands of kilometers apart, making them inaccessible to most people and difficult to study by historians and researchers in general.

Digital collections also play an important role as a strategy for preserving cultural heritage. In Brazil, in less than a decade, three museums caught on fire: the Museum of the Portuguese Language in São Paulo in 2015, the Historical National Museum in Rio de Janeiro in 2018, and more recently, the Natural History Museum in Minas Gerais in 2020. Unfortunately, a vast amount of objects were not digitized. This type of disaster unveils the lack of resources, of all kinds, faced by many

museum administrations around the world, particularly the smaller ones.

Despite all the difficulties faced by Brazilian museums, the country has a relevant amount of digitized collections. The Brazilian Institute of Museums (the Brazilian body that manages public museums) gives access through the internet to more than 15,000 items, from seventeen museums. Following the same philosophy of other collections worldwide, the Brazilian digitized collection enables access to annotated metadata with historical context for their items. The key information technology behind it is Tainacan, an open-source repository platform for creating digital archives in WordPress that also enables programmable access to the database of items [1].

Complete and reliable metadata annotation is fundamental to aggregate meaning to images in a museum's digital collection. The picture of a fork, for example, becomes an irrelevant image of an object if it is not indicated that it was used by some historical character during a dinner where great decisions were made or that its material represents a whole historical period.

Such metadata annotation is typically conducted by several specialized professionals and is a complex, labor-intensive activity and time-consuming process, frequently leading to high costs, human failure, and misunderstanding. As a result, numerous digitized collections in Brazil and worldwide suffer from a lack of metadata information, making the cultural assets unattractive and their full potential untapped.

To tackle the problem, this work proposes the use of machine learning algorithms, specifically, computer vision models, as aiding tools for specialized professionals to conduct more efficient, reliable, and

potentially less expensive metadata annotation processes.

State-of-the-art artificial intelligence (AI) algorithms have proven their success in tasks such as object recognition and automatic image captioning [2, 3]. However, they are highly dependent on the diversity and the volume of the set of images, or datasets, used in their training. For this reason, the recognition of historical objects on photos and the automatic annotation of relevant, historical, and contextual metadata, remains a challenge.

In this context, the present work describes the construction of an image dataset as a necessary step to the development of AI-based metadata annotation tools for cultural heritage assets.

As the main contribution of this work, we present EMA¹, a public dataset with approximately 12,000 images of more than 2,900 Brazilian historical objects, associated with 31 different labels. The EMA dataset can be adopted in different contexts to improve the training or to evaluate the performance of automated image captioning algorithms.

We also present baseline results using EMA dataset to train a ResNet50 artificial neural network [4]. We obtained 86.7% of accuracy in category recognition. The obtained results indicate that the dataset is consistent and can be used to implement and evaluate automated metadata annotation models.¹

The paper is organized as follows. In Section II we introduce basic concepts related to machine learning algorithms applied to digital cultural heritage, also mentioning state-of-the-art related work. Section III describes the EMA dataset construction methodology, and our approach to obtain baseline results for the new dataset. Finally, in Section IV, we discuss our results and findings.

II. RELATED WORK

A Digital Cultural Heritage (DCH) asset is a digital representation of an object with the same characteristics as the original physical object that conducts the past and present knowledge to the future [5]. In recent years, digital cultural heritage collections have been the core for museums due to the global pandemic restricting researchers from accessing the physical asset and their ease of use. Governments have been encouraged to increase research in museums to improve their IT systems, search engines, and the curation of assets [6].

In this context, many research efforts are being made worldwide to ensure that the cultural heritage assets have a reliable metadata annotation. Some institutions have made their collections available, offering authentic and ground truth data for studies to improve DCH storage, classification, and annotation [1,

7, 8]. The greater access to high volumes of data enabled the application of Deep Learning (DL), a set of machine learning techniques based on artificial neural networks, to the problem of annotation and classification in the cultural heritage field.

In 2012, DL techniques showed their potential in multi-class classification problems when Krizhevsky and colleagues proposed AlexNet [2], the winning network of ImageNet Large Scale Visual Recognition Challenge (ILSVRC 2012), successfully demonstrating the potential of Convolutional Neural Networks (CNN). Since then, different DL networks have been proposed, successively beating the image recognition rates of their predecessors like VGGNet [9] in 2014, ResNet [4] in 2015, and SENet [10] in 2017, up to the point that DL models make fewer errors than humans in ImageNet recognition task.

A. Deep Learning in Cultural Heritage

DL models are also being applied to the field of cultural heritage. In [11] and [12], the authors used deep learning to classify fine arts and obtained successful results for artist classification by adopting techniques that manipulate image structure at varying scales and resolutions. In [13], the authors used CNN to classify architectural heritage images. The network architectures adopted were AlexNet and Inception V3 and two residual networks, ResNet and Inception-ResNet-v2. ResNet achieved the best accuracy on a 64 × 64 pixel image size. In [14], the authors used Mexican architectural heritage images produced from video content and categorized styles of buildings as prehispanic, colonial, or modern style.

In [5], the authors propose classification and completion frameworks for paintings using ResNet50, showing the potential of this DL architecture in challenging classification problems.

III. METHODOLOGY

A. Dataset construction

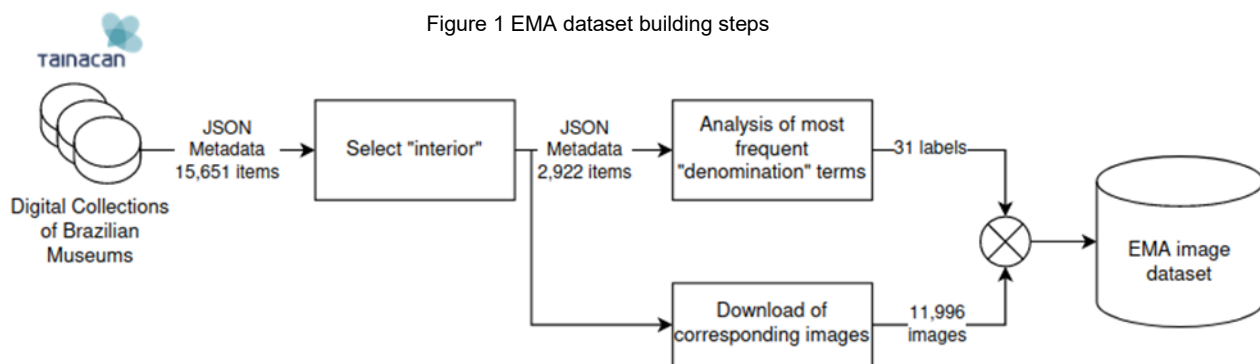
As an intermediate and necessary step to developing a cultural heritage annotation model for Brazilian context, the present work focused on the construction of a labeled image dataset.

The first step in our methodology involved the study of the digital collection managed by the Brazilian Institute of Museums (Instituto Brasileiro de Museus, IBRAM) integrated by Tainacan [1], an open-source repository platform for creating digital archives in WordPress.

We have downloaded the metadata for all the objects in the collection, which contains 15,651 objects from seventeen museums ("JSON Metadata" in Figure 1).

¹ EMA is the name of a giant flightless bird native to eastern South America that lives in Cerrado, one of the five Brazilian biomes, where the University of Brasília is located. EMA is also a Brazilian

Portuguese acronym for "Extração de Metadados Automática", or Automated Metadata Extraction.



Each object in the IBRAM's collection is categorized according to a thesaurus. A thesaurus is defined as a set of concepts, called terms or descriptors, determined according to their function or structure, ordered clearly and unambiguously, based on establishing relationships between them [16].

As a first approach to the problem, we focused on the most frequent thesaurus term in the collection, "interior", corresponding to 18.6% of the total items. In the context of cultural heritage, the term refers to daily life objects used in the interior of the houses, such as a charcoal-fired iron used to iron clothes when electricity was not available.

We also performed an interview with an IBRAM museologist who confirmed that many museums in Brazil are dedicated to showing how people lived in the past, showing, for example, how white and black people lived during slavery times. She also emphasized the relevance of developing automatic or semi-automatic tools to help museologists generate metadata for digitized items.

After filtering the original collection to keep only 2,922 "interior" objects, we analyzed which metadata field could be used to label their corresponding images. We identified that the metadata fields "title," "denomination," "material type," and "technique" are the ones that provide a general description of an item. However, we found that the fields "material type" and "technique" were not always filled and that the "title" field sometimes replaced an accurate description with an alias that does not describe the object accordingly. For this reason, we adopted the field "denomination" as the target field to extract the labels of our image dataset.

Once again, we faced a vast amount of terms used to describe the "interior" objects of the collection, and we decided to analyze the most frequent words used to describe the objects. As a result of this analysis, we decided to keep only the 31 most frequent words as image labels.

A. Image data

The match of text labels and their related images can be found in the original JSON and retrieved from the "denomination" field. To be precise and straightforward, the labels were sorted with the folder's name in the image database as it can be easy to use at the model as well. Some examples of "interior" objects

retrieved from the collection, and their labels, can be seen in Figure 2.



Figure 2 Examples of "interior" objects belonging to the EMA dataset. Their labels are *candlestick* (castiçal), *pan* (panela) and *kerosene lamp* (lâmpião), respectively.

An interesting aspect of the images downloaded from the digital collection is that we found many repeated images for the same object (probably a human error when uploading images to the Tainacan platform). We also found images within a broad spectrum of resolutions. The most frequent resolution in the dataset is 1024 x 684 pixels, but we also found images with low and unusual resolutions, such as 205 x 137 (possibly indicating that crops were made in the pictures), and high resolutions such as 2560 x 2435 pixels. While this scenario can be considered an extra challenge to train deep learning models, we kept all the non-repeated images in the dataset.

IV. RESULTS AND EVALUATION

As the main result of the present work the EMA image dataset contains 11,996 images, corresponding to 2,922 “interior” objects cataloged in seventeen Brazilian museums, that are labeled according to 31 classes.

Actual	Predicted	occurrences
spoon	cutlery	58
table knife	cutlery	44
cutlery	spoon	34
cutlery	table knife	32
sideboard	curtain	31
luminaire	sconce	26
sconce	luminaire	17
cutlery	fork	16
curtain	sideboard	15
fork	cutlery	11
bed	jug	7
fork	spoon	5
dish	spoon	4
cup of tea	spoon	3
mirror	chest	3
saucer	cup of tea	3
sideboard	table	3
chest	mirror	2
cup	spoon	2
cup of tea	saucer	2
luminaire	table	2
table	luminaire	2
table	sideboard	2
chest of drawers	chest	1
chest of drawers	fork	1
glass	spoon	1
lamp	jug	1
luminaire	table knife	1
table knife	spoon	1

Table 1 Most confused classifications

As a proof-of-concept of the use of EMA Dataset to train a DL model for cultural heritage recognition we built an image classifier that relies on the pre-trained network ResNet50. Adopting the transfer learning method, we trained the final layer using the original images with no data augmentation or any transformations. We used 80% of the images to train the model and the remaining images were used for validation and tests. The model was applied with fastai, an open-source deep learning library built on top of PyTorch, one of the leading modern and flexible deep learning frameworks. The training, validation and testing steps were performed in Google Colab [17].

The training and validation accuracy at the end of 6 epochs was 86.7%. The most confusing classifications are summarized in Table 1 and they show limitations of our methodology. For example, our methodology resulted in four labels to identify cutlery: fork, table-knife, spoon, and also cutlery. Those four labels resulted in many misclassifications since the cutlery label englobes fork, table-knife and spoon.

We also note, for example, the confusion between the classes luminaire and sconce. A luminaire can have parts of a sconce, so it is not straightforward to solve this kind of classification.

V. CONCLUSION

Metadata annotation in digital collections is a challenging task. Typical problems include lack of information and misclassifications mainly due to significant differences between modern objects and their equivalents in the past. These issues can cause data retrieval problems or associate an item to the wrong context, making it difficult to access the knowledge the object can offer.

In this paper, we presented our first steps towards developing AI-based metadata annotation tools to help museologists improve the overall quality of digital collection annotation. In particular, we presented EMA, a labeled image dataset with over 11,000 images of historical objects found in seventeen Brazilian museums. The code implemented to run all the processing and classification steps described in the present paper and the instructions to request the dataset are available in the project’s repository [18].

We also presented baseline results for this dataset through a ResNet50 DL model training. Our model could obtain 86.7% of accuracy in object recognition, showing the consistency of the dataset and the potential of this approach.

Future work includes exploring the performance of other DL architectures and increasing the dataset with other cultural heritage collections towards a generalization of the model. We also plan to develop an application that suggests labels during annotation processes.

ACKNOWLEDGMENT

We thank Amanda Oliveira, from IBRAM for the valuable information she provided regarding her experience in Brazilian museums. We also thank the Tainacan community for their support.

REFERENCES

- [1] “Tainacan – A Flexible and Powerful Repository Platform for WordPress.” Accessed March 6, 2022. <https://tainacan.org/en/>.
- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012, pp. 1097–1105.
- [3] K. Xu, J. Ba, R. Kiros, et al., “Show, attend and tell: Neural image caption generation with visual attention,” *arXiv:1502.03044 [cs]*, Apr. 2016. *arXiv: 1502.03044 [cs]*.
- [4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015. *arXiv:1512.03385.[Online].Available: <http://arxiv.org/abs/1512.03385>*.
- [5] A. Belhi, A. Bouras, A. K. Al-Ali, and S. Foufou, “A machine learning framework for enhancing digital experiences in cultural heritage,” *Journal of Enterprise Information Management*, 2020.

- [6] E. Lorang, L.-K. Soh, Y. Liu, and C. Pack, "Digital libraries, intelligent data analytics, and augmented description: A demonstration project," 2020.
- [7] A. Isaac and B. Haslhofer, "Europeana linked open data. europeana.eu," *Semantic Web*, vol. 4, no. 3, pp. 291–297, 2013.
- [8] T. Mensink and J. van Gemert, "The rijksmuseum challenge: Museum-centered visual recognition," in *ACM International Conference on Multimedia Retrieval (ICMR)*, 2014.
- [9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [11] N. Van Noord, E. Hendriks, and E. Postma, "Toward discovery of the artist's style: Learning to recognize artists by their artworks," *IEEE Signal Processing Magazine*, vol. 32, no. 4, pp. 46–54, 2015.
- [12] N. Van Noord and E. Postma, "Learning scale-variant and scale-invariant features for deep image classification," *Pattern Recognition*, vol. 61, pp. 583–592, 2017.
- [13] J. Llamas, P. M. Lerones, R. Medina, E. Zalama, and J. Gómez-García-Bermejo, "Classification of architectural heritage images using deep learning techniques," *Applied Sciences*, vol. 7, no. 10, p. 992, 2017.
- [14] A. M. Obeso, M. S. G. Vázquez, A. A. R. Acosta, and J. Benois-Pineau, "Connoisseur: Classification of styles of mexican architectural heritage with deep learning and visual attention prediction," in *Proceedings of the 15th international workshop on content-based multimedia indexing*, 2017, pp. 1–7.
- [15] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," 2016. *arXiv: 1603. 05027 [cs.CV]*.
- [16] H. D. Ferrez, "Tesauro de objetos do patrimônio cultural nos museus brasileiros," Rio de Janeiro: Fazer Arte. Gerência de Museus da Secretaria Municipal de Cultura, 2016.
- [17] J. Howard and S. Gugger, "Fastai: A layered api for deep learning," *Information*, vol. 11, no. 2, p. 108, 2020.
- [18] V. De Oliveira, P. D. P. Costa and D. L. Martins, "EMA's Project Repository" <https://github.com/AI-Unicamp/ema>, Last Access: June, 2022.

OPTIMIZING MEMORY FOR LEGACY DOS SYSTEMS

Denise de Vries

Swinburne University of Technology
Australia
dbdevries@swin.edu.au

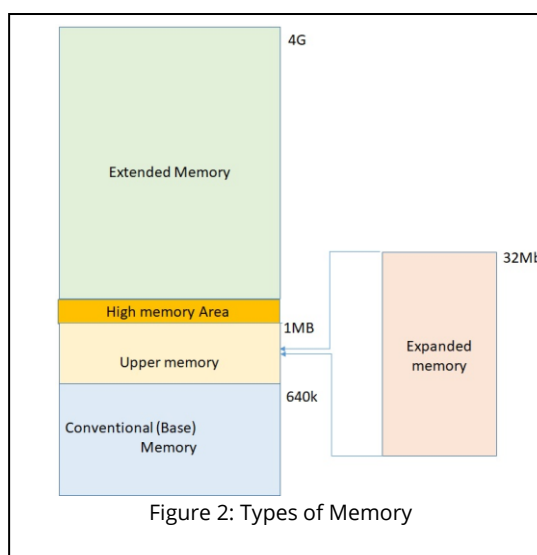
Flinders University
Australia
denise.devries@flinders.edu.au

[0000-0001-9061-6471](tel:0000-0001-9061-6471)

Abstract – Running software from the early 1980s and 1990s, often problems are encountered due to the limitation of conventional (base) memory. Even though the system may have a sizable amount of memory, only conventional memory was used to load and run programs. Customizing the system to load other necessary executable files such as drivers into memory was required. In this paper, an overview of the memory architecture of IBM-compatible personal computers is given and approaches to memory management configuration are presented.

Keywords – DOS, memory management, IBM-compatible

Conference Topics – Resilience



I. INTRODUCTION

The Play it Again 2 project “Preserving Australian video game history of the 1990s” (LP180100104) is focusing on 51 games from that period across eleven different computer platforms. The DOS and MS Windows games number 28 games from 1992 to 2000, i.e. MS DOS versions 6 and later and MS Windows versions 95 to 2000. When executing these video games difficulties arise with respect to memory limits due to the 640K conventional memory barrier.

In 1988, Microsoft launched a specification to manage previously unmanaged regions of memory in an IBM-compatible personal computer [1]. While this memory management solved the problem of applications overwriting memory addresses used by other applications, it introduced the need for the computer owner to configure a computer’s use of available memory. There were various disk operating systems (DOS) released, such as PC DOS (IBM), MS DOS (Microsoft) and DR DOS (Digital Research), all of which offered memory management. All variants are referred to as DOS in this paper. This memory management architecture was also implemented in MS Windows versions to Windows 98.

Therefore, configuring original hardware and emulators to run software from this era, it is necessary to have an understanding of the requirements of the day. In this article, I present a background to the system architecture of the IBM PC-compatible of the early 1990s and examples of configurations to optimize system memory.

II. TYPES OF MEMORY

Figure 1 shows types of memory in the IBM PC-compatible systems; each of these is described as follows.

A. Conventional (Base) Memory

The original PC/XT-type computer was designed to use 1M of memory workspace (RAM). This 1M of RAM was divided into several sections. DOS could read and write to the entire megabyte, but could manage the loading of programs only in the portion of RAM space called **conventional memory**, which at the time the first PC was introduced was 512K. The other 512K was reserved for use by the system itself. IBM decided that only 384K was needed for these reserved uses, and then began marketing PCs with 640K of user memory.

Thus, 640K became the standard for memory that could be used by DOS for running programs and resulted in what was known as the 640K memory barrier. The remaining memory after 640K was reserved for use by the graphics boards, other adapters, and the motherboard ROM BIOS. [2]

B. Upper Memory Area

The term Upper Memory Area (UMA) describes the reserved 384K at the top of the first megabyte of system memory. While this area was termed *reserved memory*, it was possible to use unused regions of this memory to load device drivers and memory-resident programs to free up the conventional memory they would otherwise require.[2, 3]

C. High Memory Area

The High Memory Area (HMA) is the first 64K of extended memory. The HMA's size was fixed, no matter how much extended memory was available.[3] HMA was used to load device drivers and memory-resident programs, to free up conventional memory. [2]

D. Extended Memory

Extended memory is all memory past the first megabyte, which could only be accessed while the processor was in protected mode. The extended memory specification (XMS) was developed to specify how programs would use extended memory. Extended memory required an extended-memory manager, such as HIMEM.SYS.[2, 3]

E. Expanded Memory

Some DOS programs used a type of memory called Expanded Memory Specification or EMS memory. Expanded memory was installed on an expanded memory board and came with an expanded memory manager. As EMS was designed for 8-bit systems, The memory manager EMM386 was used instead to convert extended to expanded memory for backwards compatibility. Only programs written specifically to make calls to expanded memory require it. Therefore, some DOS programs use expanded memory while others do not.[2, 4]

III. OUT OF MEMORY? BUT THERE IS A LOT!

An often encountered message is "not enough conventional memory". As explained in the previous section programs must be loaded into the conventional memory of the system (640K). So no matter what the total memory is of the system, if

there is not enough space in conventional memory, the program cannot run.

Most DOS and many early Windows systems load numerous device drivers and TSR (terminate-and-stay resident) programs during the boot cycle. These programs are, by default, loaded into conventional memory, taking up valuable space. Memory management techniques are needed to load these device drivers and TSRs into the upper memory, allowing more conventional memory to be made available to programs.

Customizing the CONFIG.SYS and AUTOEXEC.BAT files to manage the placement of device drivers and TSRs into upper memory blocks (UMBs) in the upper memory area (UMA) on booting, in maximizes the conventional memory available for applications.

The CONFIG.SYS file is a text file containing commands that configure the computer's hardware components (memory, keyboard, mouse, etc). When DOS starts, it carries out the commands in the CONFIG.SYS file first.

The AUTOEXEC.BAT file is a batch program that DOS runs immediately after carrying out the commands in the CONFIG.SYS file. The AUTOEXEC.BAT file contains the commands to be executed when the system is started. DOS carries out the commands in both the CONFIG.SYS and AUTOEXEC.BAT files **each** time the computer is started.[4, 5]

A. CONFIG.SYS

Each hardware component of your computer is called a device. The keyboard, mouse, display, printer, disk drives, and memory boards are all devices. Each device has its own characteristics that can be customized. DOS has built-in device drivers for the keyboard, display, hard drives and diskette drives, and communication ports. Other devices, such as memory boards, a mouse, or CD-ROM have device drivers that are not built into DOS. Such a device driver is called an **installable** device driver: these are installed by adding a command to the CONFIG.SYS file. [5, 6]

While most CONFIG.SYS commands can appear in your CONFIG.SYS file in any order. The order of the

device and devicehigh commands is important because some device drivers enable devices that are needed by other drivers. The HIMEM.SYS extended-memory driver must be loaded before any drivers that use extended memory.

The order in which device drivers should appear in the CONFIG.SYS file is as follows:

1. DEVICE=C:\DOS\HIMEM.SYS
2. DEVICE=C:\DOS\EMM386.EXE NOEMS
3. DOS=HIGH,UMB
4. Any other device drivers.[5, 6]

The DEVICE commands load the HIMEM.SYS and EMM386.EXE device drivers. The HIMEM.SYS driver manages extended memory. The EMM386.EXE driver, provides access to the upper memory area and simulates expanded memory. The DOS=HIGH,UMB command runs DOS in the high memory area and specifies that programs should have access to the upper memory area.[5, 6]

If programs require expanded memory (EMS), start EMM386 with the NOEMS switch. This can give you an additional 64K of UMBs. The NOEMS switch instructs EMM386 not to create an EMS page frame in the upper memory area. If EMM386 is started with the NOEMS switch, programs will be unable to use expanded memory.

B.

In Code 1 is an example of CONFIG.SYS commands to load the device drivers for memory management, followed by loading DOS onto upper memory. On line 4 is the command to load the driver

```
DEVICE=C:\DOS\HIMEM.SYS
DEVICE=C:\DOS\EMM386.EXE NOEMS
DOS=HIGH,UMB
DEVICEHIGH /L:2=C:\CDROM\CDROM.SYS /D:MSCD000
```

Code 1: Example CONFIG.SYS

for a CD-ROM drive. This is specific for the installed drive. The driver might have another name and be located in another directory.

The parameter /L:2 sets the UMB block where the driver should be loaded. The /D: MSCD000 is the device name, not the driver. It is important to note that this name must match what is in AUTOEXEC.BAT. If these names do not match, the CD-ROM drive will not load.

Other commonly used commands in CONFIG.SYS are:

Memory Type	Total	Used +	Free
Conventional	640K	136K	504K
Upper	71K	47K	24K
Reserved	0K	0K	0K
Extended (XMS)	31,545K	2,752K	18,978K
Total memory	32,256K	2,752K	29,504K
Total under 1 MB	711K	183K	528K

Total Expanded (EMS)	24,960 (25,559,040 bytes)
Free Expanded (EMS)*	24,576 (25,165,824 bytes)

*EMM386 is using XMS memory to simulate EMS memory as needed
Free EMS memory may change as free XMS memory changes
Largest executable program size 503K (515,328 bytes)
Largest free upper memory block 16K (16,624 bytes)
MS-DOS is resident in the high memory area.

Figure 3 An Example MEM Report

- BUFFERS to specify how much memory is reserved for transferring information to and from disks.
e.g. BUFFERS=20
- COUNTRY to set the language conventions for the system.
e.g. COUNTRY=031,850,C:\DOS\COUNTRY.SYS for The Netherlands.
- FILES to specify how many files can be open at a time.
e.g. FILES=30
- LASTDRIVE to set the number of valid drive letters.
e.g. LASTDRIVE=J

AUTOEXEC.BAT

```
@ECHO OFF
LH:/L:2 C:\DOS\MSCDEX /D:MSCD000
LH /L:2 C:\MOUSE\MOUSE
SET SOUND=C:\SB16
SET BLASTER=A220 I5 D1 H5 P330 T6
SET MIDI=SYNTH:1 MAP:G
PROMPT $p$g
PATH C:\DOS;C:\SB16
SET TEMP=C:\TEMP
```

Code 2: Example AUTOEXEC.BAT

AUTOEXEC.BAT executes each command in the exact order in which they have been placed. These commands specify where device programs are loaded, and DOS environmental settings are set.

In Code 2 the example AUTOEXEC.BAT file shows loading the CD DOS extension to access a CD drive matching the device specified in CONFIG.SYS. LH loads the driver to high memory.

The mouse driver may have a different name and be in a different directory from the example.

The directory is set the location of the sound card files. The sound card is installed at Address 220 with IRQ (Interrupt Request) 5, Low DMA (Direct Memory Access) on DMA-channel 1, High DMA on DMA-channel 5, MIDI (Musical Instrument Digital Interface) address 330 and that the soundcard is Type 6 (Sound Blaster 16 compatible).

SET MIDI-sets how MIDI files are played. MAP:G ensures that both basic MIDI and extended MIDI play.

The DOS prompt is set to the current drive and path followed by ">".

The search path for executable files is set.

Directories names for temporary files generated by programs are set. Both TEMP and TMP were used as variable names for DOS programs.

C. Analyzing Memory

In DOS version 4, the command MEM was provided to analyze memory. With DOS version 6 the MEM command came with additional switches to produce more detailed reports on memory usage.[3, 5]

By running MEM a report similar to Figure 2 is produced showing the different types of memory and what has been used and still available. Most importantly can be seen what the largest executable program size is.[3, 5]

To further analyze memory usage, the /C switch lists all the drivers and programs that have been loaded with their installed sizes.

Altering the CONFIG.SYS and AUTOEXEC.BAT files , rebooting the computer and then running the MEM command enables more control over memory usage by fine tuning the computer's configuration.

IV. SUMMARY

In the late 1980s, a memory management architecture was implemented in IBM-compatible personal computers. The memory comprised four defined types: conventional, upper memory, high memory blocks, extended memory plus in some systems expanded memory.

The 640K limit of conventional memory became the standard memory size for running programs in DOS which became known as the "640K memory barrier".

In order to use other areas of memory available, it is necessary to customize the CONFIG.SYS and AUTOEXEC.BAT files to load the operating system, device drivers and TSRs into these regions.

Examples of possible customizations of commands have been presented to explain the syntax and recommended order in which these commands should be executed on booting the system.

Analyzing the result of configuration amendments can be made using the MEM command which displays a summary of the memory configuration. It shows how much of each kind of memory there is, how much is currently in use, and how much is currently free.

Whether original hardware or emulated systems are used to run legacy software from the late 1980s to the 1990s, knowledge of the memory architecture and how to optimize it are required for a successful experience.

REFERENCES

- [1] Gibson, S., *Placing the IBM/Microsoft XMS Spec into Perspective*. Info World, 1988. **10**(33).
- [2] Mueller, S., *Upgrading and repairing PCs*. 4 ed. 1994, Indianapolis USA: Que.
- [3] Prosise, J., *PC magazine guide to DOS 6 memory management with utilities*. 1993, Emeryville, USA: Ziff-Davis Press.
- [4] Microsoft Press, *Microsoft MS-DOS programmer's reference* 1993, Redmond, WA USA: Microsoft Press.
- [5] Microsoft Corporation, *Concise User's Guide Microsoft MS-DOS 6*. 1993, USA: Microsoft Press.
- [6] IBM, *PC DOS 7*. 1995: International Business Machines Corporation.

METADATA ENCODING AND TRANSMISSION STANDARD (METS) VERSION 2

Karin Bredenberg

*Kommunalförbundet
Sydarkivera, Sweden*
karin.bredenberg@sydarkivera.se
[0000-0003-1627-2361](tel:0000-0003-1627-2361)

Aaron Elkiss

*HathiTrust
USA*
aelkiss@hathitrust.org
[0000-0002-2904-9559](tel:0000-0002-2904-9559)

Inge Hofsink

*KB National Library of the
Netherlands*
inge.hofsink@kb.nl
[0000-0002-8366-8983](tel:0000-0002-8366-8983)

Juha Lehtonen

*CSC - IT Center for Science
Finland*
juha.lehtonen@csc.fi
[0000-0002-9916-5731](tel:0000-0002-9916-5731)

Andreas Nef

*docuteam AG
Switzerland*
a.nef@docuteam.ch
[0000-0003-2324-6444](tel:0000-0003-2324-6444)

Tobias Steinke

*German National Library
Germany*
t.steinke@dnb.de
[0000-0002-3999-1687](tel:0000-0002-3999-1687)

Robin Wendler

Harvard University, USA
robin_wendler@harvard.edu
[0000-0003-2158-4319](tel:0000-0003-2158-4319)

Abstract – The METS Editorial Board is working on version 2 of the Metadata Encoding & Transmission Standard (METS), work which aims to make METS easier to use and implement. Version 2 simplifies the schema, makes it more consistent, and removes reliance on the outdated XLink standard. It aims to retain a clear path for migration from METS 1 for most use cases. In this paper the METS Editorial Board presents the changes in a short form and invites comments and thoughts on the evolution of METS.

Keywords – METS, evolution, transfer formats
Conference Topics – Exchange; Resilience.

I. INTRODUCTION

METS, the Metadata Encoding & Transmission Standard, has been used for describing digital objects since 2001. The METS XML schema version 1.x (METS 1) is used both as an interchange and storage format by numerous systems in the digital preservation space [1,2]. A METS document can describe the manifest of files that make up a digital object, their structural relationship to each other, and include a variety of metadata about the digital object and its component files.

Since the release of METS 1.x, other standards have emerged; most notably, the Portland Common Data Model (PCDM) [3] and the International Image Interoperability Framework (IIIF) [4]. These standards complement rather than replace METS. PCDM is focused on describing digital objects via RDF/linked data, while METS is an XML representation. IIIF is focused primarily on delivering and describing digitized images; METS is often used for this purpose as well, but is considerably more general. Other standards such as BagIt [5] and the Oxford Common Filesystem Layout (OCFL) [6] standardize manifests and directory layouts for digital objects; METS complements these standards by providing a way to describe structure and to link content with metadata.

A. Motivation

METS 1 has been largely stable for many years. No new elements have been added to the schema since 2010; changes since then have primarily been to allow new values for specific attributes and to allow arbitrary attributes to appear on a variety of elements (via `xsd:anyAttribute`). Around 2011, the METS Editorial Board started exploring potential future directions for METS, areas where METS has

been successful, and areas where METS has not been as successful [7]. This work did not result in a new version of METS at that time. However, in recent years, the METS Editorial Board has been made aware of a variety of issues and incompatibilities related to the XLink schema used in METS 1 [8]. After discussion, it became clear that the best solution to the XLink issues was to move forward with the design of a new major revision of METS that did not need to maintain strict backwards compatibility. This also enabled consideration of a more general overhaul of the METS schema, building on the earlier exploration in [7].

The basic idea of this new version of METS (METS 2) is to make METS simpler and more flexible by removing rarely-used features and by improving consistency between its various parts. From the beginning of the design process, it was a goal to maintain the general concepts of METS, to continue to support the major use cases of METS, and to make it easy to adapt and migrate a large majority of existing uses of METS 1 to METS 2.

At the same time, the METS Editorial Board recognized that not all systems will migrate from METS 1 to METS 2. The METS 1 schema will continue to be available and will continue to be supported for the foreseeable future. In particular, implementations which rely on elements such as `<structLink>` and `<behaviorSec>` in METS 1 will continue to be supported with METS 1; if there are any bugs found, a new version of METS 1 could be released, but most effort from the Board will be on METS 2 going forwards.

Usage of every element and attribute was checked against registered METS profiles [2]. Known problems and inconsistencies of METS 1 were discussed, and possible solutions were considered in terms of their fit with the overall concepts of METS. The result is a kind of “METS Light”, improving consistency and ease of implementation without giving up flexibility or versatility.

II. CHANGES IN METS 2

The changes in the METS 2 schema all serve to simplify usage by making the schema more consistent and by removing some rarely-used features. As METS 2 is not backwards-compatible with METS 1, there is a new namespace URI for the schema. METS 2 reorganizes the major sections of the METS file: it removes the `<structLink>` and

`<behaviorSec>` sections entirely, simplifies the `<dmdSec>` and `<amdSec>` metadata sections into a single `<mdSec>` section, and adopts a parallel organization for the remaining major sections. METS 2 also removes reliance on the XLink specification [9] and removes most lists of allowed attribute values from the schema in favor of suggested external controlled vocabularies. The details of each change and motivation behind each specific change are discussed below.

METS 2 is still in an early stage of development, but is now ready for discussion and feedback. The draft schema, generated documentation, and instructions for feedback are all available in GitHub at <https://github.com/mets/METS-schema>.

A. Removing XLink

When METS 1 was first drafted in 2001, XLink was in the process of being adopted as a W3C recommendation, and seemed promising for future adoption. In the intervening years, XLink has had little uptake. Although XLink was revised in 2010 [9], there is no browser support for XLink beyond basic XLinks in SVG, and SVG 2 deprecates XLink entirely [10]. The continued inclusion of XLink in METS can also cause validation problems when using METS alongside other XML schemas that also reference XLink but include reference to a slightly different XLink schema [8]. Schemas which used XLink in the past have moved away from it: notably, schemas often used with METS such as PREMIS 3 and EAD 3 drop XLink entirely in favor of schema-local attributes [11,12].

METS 2 follows this trend by removing extended XLinks entirely, dropping references to rarely-used XLink attributes, and using a locally-defined LOCREF attribute instead of the `xlink:href` attribute in `<FLocat>` and `<mdRef>` elements. The draft METS 2 schema also allows LOCREF to be any string, not just a URI as with `xlink:href` in METS 1. In practice the `xlink:href` attribute was used even when the location was not actually a URI – for example, locally-defined identifiers, or relative paths defined without reference to a base URI. Changing the attribute name and type removes this potential semantic confusion.

B. Unrestricted Attribute Values

The draft METS 2 schema removes the restriction on allowed values for several attributes such as MDTYPE, LOCTYPE, CHECKSUMTYPE, and others. Enumerated lists of values in the schema limit

extensibility and flexibility for users, delay the availability of values for new standards, and add to the proliferation of schema versions over time. With METS 1, implementers could use (for example) `MDTYPE="OTHER"` and provide an `OTHERMDTYPE` value, but this literally serves to “other” and devalue data types not explicitly approved by the METS Editorial Board. Alternatively, implementers could request new approved values, but the overhead of evaluating and approving requests and releasing a new version of the schema means there is a significant delay between the request and the availability of the new term for use. With METS 2, recommended standards and values for these attributes will instead be documented externally to the schema itself, as it was done for PREMIS 3 [13]. This will both reduce changes required to the schema and reduce the barrier to extending lists of possible attribute values.

C. *Removing Rarely-Used Sections*

Neither the `<structLink>` nor `<behaviorSec>` sections are included in METS 2. Both these sections are rarely used in METS 1, as determined through a review of registered profiles as well as general web and GitHub searches.

The `<structLink>` element was added in METS 1.1 for recording hyperlinks between media represented by `<structMap>` nodes. These hyperlinks were represented by extended `XLink` objects that could be used to record links between `<structMap>` nodes separately from the `<structMap>` nodes themselves. The primary documented use case for `<structLink>` was to indicate links between web pages described in a METS object [14]. However, in the intervening years the Web ARChive (WARC) file format has emerged as a standard way of capturing web archives, minimizing the need for METS to handle this use case. Likewise, `XLink` (especially extended links) did not come into widespread usage. Thus, METS 2 removes support for `<structLink>`.

The `<behaviorSec>` element was added to the ‘epsilon’ revision late in the design process of METS 1 to support referencing executable code from METS objects. This was primarily to support a use case for earlier versions of the Fedora digital repository system [15]. Fedora has since moved away from the use of METS and XML in general, and this section has not been widely used or supported by other METS implementations.

D. *Simplified Metadata Section*

In METS 1, metadata is recorded in purpose-specific sections and elements. Descriptive metadata is recorded in section `<dmdSec>`; all other metadata is recorded in an administrative metadata section `<amdSec>`. This parent section `<amdSec>` is separated into four subsections for technical metadata (`<techMD>`), intellectual property rights (`<rightsMD>`), analog/digital source metadata (`<sourceMD>`), and digital provenance metadata (`<digiprovMD>`). Multiple instances of the element `<amdSec>` can occur within a METS document and multiple instances of its subsections `<techMD>`, `<rightsMD>`, `<sourceMD>` and `<digiprovMD>` can occur in one `<amdSec>` element.

METS 2 makes all metadata sections more generic by using general elements `<mdSec>`, `<mdGrp>` and `<md>` following the hierarchy of the file section. All these elements can be referenced using the general `MDID` attribute instead of the more specific `DMDID` and `ADMID` attributes from METS 1. METS 2 does not prescribe a specific vocabulary or syntax for encoding metadata. These changes simplify the schema and in turn processing software while enhancing flexibility in the structuring of the metadata.

In METS 2, the metadata section `<mdSec>` contains all metadata pertaining to the digital object, its components and any original source material from which the digital object is derived. The optional `<mdGrp>` element allows grouping related kinds of metadata. This could be all metadata of a particular type, all metadata coming from a particular source, all metadata pertaining to a certain file or set of files, or any other relevant grouping; the `<mdGrp>` can then be referenced from an `MDID` attribute elsewhere. The `<md>` element records any kind of metadata about the METS object or a component thereof. As with metadata elements in METS 1, the `<md>` element can include the metadata inline with `<mdWrap>`, reference it in an external location via `<mdRef>`, or both. The `<mdSec>` element can contain any number of `<mdGrp>` elements which in turn contain any number of `<md>` elements, or it can include `<md>` elements directly if grouping is not needed. As in METS 1, included or referenced metadata can be in any format, XML or otherwise. METS 2 replaces the varied element names with a `USE` attribute comparable to that on `<fileGrp>`. Values could include `DESCRIPTIVE`, `TECHNICAL`, `RIGHTS`, `SOURCE`, `PROVENANCE` to correspond to the

various metadata sections available in METS 1, or could use any other value according to local needs.

Example with METS 1:

```
<mets>
...
<dmdSec>...</dmdSec>
<amdSec>
  <techMD>...</techMD>
  <rightsMD>...</rightsMD>
  <sourceMD>...</sourceMD>
  <digiprovMD>...</digiprovMD>
</amdSec>
...
</mets>
```

Example with METS 2, preserving METS 1 semantics

```
<mets>
...
<mdSec>
  <mdGrp USE='DESCRIPTIVE'>
    <md USE='DESCRIPTIVE'>...</md>
  </mdGrp>
  <mdGrp USE='ADMINISTRATIVE'>
    <md USE='TECHNICAL'>...</md>
    <md USE='RIGHTS'>...</md>
    <md USE='SOURCE'>...</md>
    <md USE='PROVENANCE'>...</md>
  </mdGrp>
</mdSec>
...
</mets>
```

Example with METS 2 without <mdGrp>

```
<mets>
...
<mdSec>
  <md USE='DESCRIPTIVE'>...</md>
  <md USE='TECHNICAL'>...</md>
  <md USE='RIGHTS'>...</md>
  <md USE='SOURCE'>...</md>
  <md USE='PROVENANCE'>...</md>
</mdSec>
...
</mets>
```

E. Removing Nested File Groups

METS 2 retains the file section <fileSec> that lists the files that comprise the digital object described in the METS document.

METS 1 supported arrangement of files with nested file group (<fileGrp>) elements, and allowed mixing both <fileGrp> and <file> elements at the same level. In METS 2, the <fileGrp> element is made

optional, and <fileSec> may contain either <fileGrp> elements or <file> elements directly. This simplifies the schema and processing software; it also makes <fileSec> / <fileGrp> / <file> consistent with <mdSec> / <mdGrp> / <md>.

In METS 1, nested file groups were sometimes used to describe structural information about the object. METS 2 clarifies that <fileSec> is for listing a manifest of files in the object and that the <structMap> element is the way to represent structure. As in METS 1, <file> elements themselves may still be nested, which is often useful for representing the physical structure of archive formats such as .zip, etc.

In this METS 1 example, the dprov-001 metadata section applies to all nested <fileGrp> elements:

```
<mets>
...
<fileSec>
  <fileGrp ADMID="dprov-001"
    USE="Images">
    <fileGrp USE="Original">
      <file ID="file-001"
        ADMID="tech-001">
        <FLocat LOCTYPE="URL"
          xlink:href="https://example.org/img001.tif"
        />
      </file>
    </fileGrp>
    <fileGrp USE="Thumbnails">
      ...
    </fileGrp>
  </fileGrp>
  <fileGrp ADMID="dprov-002"
    USE="Documents">
    <file ID="file-doc-001"
      ADMID="tech-doc-001">
      <FLocat LOCTYPE="URL"
        xlink:href="https://example.org/doc001.pdf"
      />
    </file>
  </fileGrp>
</fileSec>
...
</mets>
```

In METS 2, there may be no more than one level of <fileGrp> elements, so the reference to dprov-001 is repeated across multiple <fileGrp> elements:

```
<mets>
...
<fileSec>
  <fileGrp MDID="dprov-001">
```



```

    USE="Original Images">
    <file ID="file-001" MDID="tech-001">
      <FLocat LOCTYPE="URL"
        LOCREF="https://example.org/img001.tif" />
    </file>
  </fileGrp>
  <fileGrp MDID="dprov-001"
    USE="Thumbnails">
    ...
  </fileGrp>
  <fileGrp MDID="dprov-002"
    USE="Documents">
    <file ID="file-doc-001"
      MDID="tech-doc-001">
        <FLocat LOCTYPE="URL"
          LOCREF="https://example.org/doc001.pdf" />
      </file>
    </fileGrp>
  </fileSec>
  ...
</mets>

```

As with <mdSec> and <md>, if there is no need for multiple groups, <file> elements may be added directly under the <fileSec> element:

```

<mets>
...
<fileSec>
  <file ID="file-001"
    MDID="tech-001 dprov-001">
    <FLocat LOCTYPE="URL"
      LOCREF="https://example.org/img001.tif" />
  </file>
  ...
  <file ID="file-doc-001"
    MDID="tech-doc-001 dprov-002">
    <FLocat LOCTYPE="URL"
      LOCREF="https://example.org/doc001.pdf" />
  </file>
</fileSec>
...
</mets>

```

Instead of repeating the references to dprov-001 and dprov-002 in <file> elements, these could be included in <div> elements in a structural map.

F. Structural section

METS 1 and 2 both support including multiple structural maps. In METS 1 the structural maps were included directly to the main level as <structMap> elements. In METS 2, these <structMap> elements are included in a new structural section <structSec>.

Although <structMap> was required in METS 1, to match other sections and to support using METS as a simple manifest of files, <structSec> is optional in METS 2.

III. FUTURE WORK

In addition to accepting and discussing comments and feedback, the METS Editorial Board will undertake additional work before publishing METS 2 as a released standard:

- Release a white paper describing these changes in greater detail and providing additional examples
- Create an XSLT transformation and/or other tools to aid in migration from METS 1 to METS 2
- Update the METS primer and tutorial for METS 2
- Publish vocabularies for attributes whose allowed values are no longer encoded in the schema
- Review and update the METS profile schema to support METS 2

Our hope is that METS 2 simplifies and further encourages the adoption of METS for the purposes of describing, preserving, and providing access to digital objects.

ACKNOWLEDGMENT

Thanks to Bertrand Caron for proofreading and to the entire METS Editorial Board for design and discussion of the METS 2 draft schema.

REFERENCES

- [1] METS, COPTR, 2021 <https://bit.ly/3CA6OMV>
- [2] METS Profiles, METS Editorial Board, 2017. <https://bit.ly/3Mx6KBM>
- [3] Portland Common Data Model, DuraSpace, 2016. <https://pcdm.org/models>
- [4] International Image Interoperability Framework. <https://iiif.io>
- [5] J. Kunze, J. Littman, E. Madden, J. Scancella, C. Adams, "The BagIt File Packaging Format (V1.0)", RFC 8493, 2018. <https://www.rfc-editor.org/info/rfc8493>
- [6] A. Hankinson, N. Jefferies, R. Metz, J. Morley, S. Warner, A. Woods, "Oxford Common File Layout Specification 1.0", 2020. <https://ocfl.io/1.0/spec/>
- [7] Reimagining METS: An Exploration for Discussion, METS Editorial Board, 2011. <https://bit.ly/3Coy8xy>
- [8] T. Habing, et al, "Primer Xlink Issue", 2019. <https://github.com/mets/METS-board/issues/19>
- [9] S. DeRose, E. Maler, D. Orchard, N. Walsh, "XML Linking Language (XLink) Version 1.1", W3C, 2010. <https://www.w3.org/TR/xlink11/>

- [10] A. Bellamy-Royds, et al, "Scalable Vector Graphics (SVG) 2", W3C, 2018 <https://www.w3.org/TR/SVG2/>
- [11] R. Denenberg, "PREMIS Preservation Metadata XML Schema Version 3.0", 2016. <https://bit.ly/3MvnNo8>
- [12] Encoded Archival Description Tag Library - ver. EAD3, Soc. Amer. Arch, 2019. <https://www.loc.gov/ead/EAD3taglib/EAD3.html>
- [13] Preservation Schemes (all), Library of Congress. <https://id.loc.gov/vocabulary/preservation.html>
- [14] J. McDonough, "Question regarding StructMap and StructLink." METS Listserv, 2004. <https://bit.ly/3KpZVjG>
- [15] J. McDonough, "METS Meeting Summary." METS Listserv, 2001. <https://bit.ly/3LXS2C3>

ACCESS QUALITY METRICS FOR NET ART

Xiao Ma

Independent Researcher

USA

xm75@cornell.edu

[0000-0001-6134-3531](tel:0000-0001-6134-3531)

Dragan Espenschied

Rhizome

USA/Germany

dragan.espenschied@rhizome.org

[0000-0003-1968-6172](tel:0000-0003-1968-6172)

Lyndsey Jane Moulds

Rhizome

USA

lyndsey.moulds@rhizome.org

[0000-0002-4858-0417](tel:0000-0002-4858-0417)

Rhizome's ArtBase is a public archive holding copies of more than 800 works of net art. Most pieces allow for different points of access: they might be available live from a Rhizome web server or alternatively from a web archive, with both versions potentially incomplete and in different states of restoration. Visitors might view these works via a period-adequate browser in an emulator, or whatever setup they are running on their devices. Discussed below is a system using technical metadata and curatorial information to calculate an access quality score that can help visitors choose which artworks, versions, and modes of access will best meet their needs.

**Keywords – Digital Art, Net Art, Access, Emulation
Conference Topics – Innovation; Resilience**

I. PRESENTING NET ART IN AN ARCHIVAL CONTEXT

Rhizome's ArtBase, an online archive started in 1999, holds pieces of net art that have entered the collection through different mechanisms, including open accession (1999-2008), curation (2011-2020), and open calls (starting 2021) [1]. Methods used to package and stabilize the works varied depending on what tools and concepts were available at the time (artist-submitted file copies, web archiving, disk imaging, etc.) as well as how the artworks were made and conceptualized as objects [2].

The quality of access to archived born-digital art can be thought of as a result of two factors: first, the availability of stable and complete resources, and second, the capabilities of the software environment used to perform the works. Net art introduces further complexity: many works are not self-contained, but instead present a “blurry” object boundary [3] that may not be easily understood or accurately demarcated in the moment of archival. As an example, an artist might submit an incomplete set of files to the archive, and omissions might not become apparent until external resources fall offline much later.

Additionally, net art is usually produced for and accessed via whatever devices and software internet users have available and is in most cases not tied to a canonical software environment. Over time, this mix of operating systems, browsers, and other applications to access online materials change in their forms and capabilities. These changes range from the drastic, like deprecation of certain file formats and programming languages, to less noticeable changes such as the deprecation of features allowing browsers to open popup windows, play MIDI music, or draw certain UI widgets [4].

Preservation and restoration actions can in many cases retroactively supply missing resources and, via emulation, prepare software environments that provide the best possible circumstances for the digital artifacts to be performed. The result of each preservation action is a new “variant” of the artwork [5]. Each variant is composed of a set of stabilized artifacts and a software and network environment.

For each of these variants, Rhizome aims to provide an access quality score that is an expression of these possible states of a variant. This is done to direct newcomers to highlights of the collection and manage users' expectations of artworks that expose deficiencies. The score is especially useful while an artwork is transitioning from being best accessible on the live web to being best represented in a controlled, encapsulated environment constructed for preservation purposes. Users will have to make the tradeoff between accessing a variant of the artwork that is integrated into the present landscape of the internet but may be less functional, versus a variant that is more separate from the live internet but offers a reliable, reproducible performance. The access quality score can guide them to the variant that fits their intention for access.

Described below is the data model and process required to compute a single access quality value per variant that can be displayed as a simple 3 level “stop light” indicator on access links: green variants should be expected to be as complete as possible and have all current preservation goals met, yellow variants have known problems, red variants must be expected to be incomplete or at least partly non-functional [6].

II. DATA MODEL AND DATA SOURCES

ArtBase is built as a Linked Open Data repository with Wikibase. Artworks are modeled in the following manner:

- Variants are represented as a combination of *artifacts* and *machines*.
- Artifacts can be collections of files, disk and media images, containers, web archives, etc. [7], with their components described based on the PRONOM file format registry.
- Machines are configured virtual machines, emulators, and containers managed via EaaS, or an approximation of the software environments widely used (see below). They are described by the *software* that is installed on the disk image they boot from.
- Each software is a self-contained, installed package with its *capabilities* described by the data formats it can handle, again using the PRONOM file format registry.
- Finally, the capabilities of a machine represent the sum of the capabilities of the software installed on it, which then can be matched against the components of the artifacts.

This technical data can be automatically generated (artifact composition can be determined by a tool like Siegfried) and observed and recorded in experiments (supported data types can be elicited by trying to run software with specimens of that data type).

One special type of machine is the “default access machine,” representing the capabilities of an assumed contemporary software environment that approximates the lowest common denominator of different devices, operating systems, browsers, etc. that are available to regular web users. New machines are described in sync with the general landscape of contemporary software changing, and assigned to variants that are accessed “directly,”

Table 1 Data Model

subject	predicate	object	note
machine	has part	software	Software installed on machine
software	has part	software	Optional nesting for bundles
software	handles	data format	Capabilities of a software package
artifact	made of	data format	Artifact has at least one occurrence of a data format
variant	has artifact	artifact	Artifact used in variant
variant	has machine	machine	Machine used this variant
variant	handles	data format	Optional curatorial information overriding machine values
variant	made of	data format	Optional curatorial information overriding machine values
	relevance	relevance value	Qualifier holding a multiplier value indicating a data format’s relevance for the intended purpose of the variant.
variant	access quality	access quality value	Computed value expressing the variant’s access quality.

rather than via an emulator. In addition to representing an approximation of currently available capabilities, modeling a projected default access machine can be used to project the effects of upcoming software changes on a collection, for instance when a browser vendor announces that support for a particular video codec or plugin will be discontinued.

Information recording the capabilities of software based on PRONOM has already been proven meaningful to create matches of existing configured machines with artifacts in a library context [8][9]. When applied in the context of art and access quality, the considerations need to be slightly different, in sometimes counterintuitive ways:

- 1) There is no correlation between the number of occurrences of a certain data format (as in “how many files of this type are part of an artifact?”) with the relevance of that data format for an artwork’s performance. For instance, a Microsoft Word file being part of an artifact might be an artist’s

description of their work submitted as a package to Rhizome and not be referenced or linked to in the actual artwork at all. As a result, the machine used for access does not need to provide software to render this file if the goal is to present the artwork. In another access context, like the analysis or exhibition of artists' descriptions of their work, the capability to render this type of file would be essential. Each access scenario needs to be modeled as its own variant, combining the same artifact with different machines. A machine needs to provide the capabilities to render a data type if it occurs more than zero times and is relevant. The machine does not need to provide capabilities if the data type occurs zero times or is deemed irrelevant. The actual number of occurrences greater than one is not producing better scoring results. Even if only one jpeg file out of a set of ten is deemed relevant, the machine used will have to support that format.

2) Unidentifiable or misidentified data formats are common in digital art, in which oftentimes artists employ tools that have little relevance in the library field and hence are not represented in the PRONOM registry, or at least not at the required level of detail that would enable correct automatic detection. Additionally, the adherence to standards like "well-formed XML" that would make format detection more reliable has little relevance in the production of digital art. 34% of the works in ArtBase contain at least one occurrence of an unidentified data format. "Clean" solutions—implementing a new format detection rule, or manually assigning a synthetic format ID to every occurrence—is considered too laborious and demanding too much expert knowledge to implement in day-to-day collection management. Instead, both cases are handled via a value manually added to the variant that denotes the relevance of a particular data format for the access quality calculation. Since the relevance of a file format is tied to the intention of making the variant available, it does not make sense to record it with the artifact.

III. FROM DATA TO READINESS SCORE

Defining Readiness Score: Baseline

We have established above that a variant is not just the static files (artifacts) associated with it but is a combination of artifacts performed in a particular environment (machine).

$$\text{Variant} = \text{Artifact} \times \text{Machine}$$

Therefore, we define a "readiness" score for a variant as a feature of the variant that indicates how likely the variant's performance can be perceived as complete by the user.

The most basic definition of a readiness score can be:

$$\text{readiness_score}_0^{\text{artifacts,machine}} = \frac{\text{num_supported_data_formats}}{\text{num_data_formats}}$$

Here is a toy example:

Table II Example Variants

Variant ID	Artifact	Artifact composition	Machine ID	Supported (inferred)
1	1	doc	97	True
1	1	jpeg	97	True
1	1	mp3	97	True
2	1	doc	98	True
2	1	jpeg	98	False
2	1	mp3	98	False

Let's say we have a variant with ID 1 composed of an artifact that contains three types of files, doc, jpeg, and mp3. We have two machines that might support the artifacts, therefore we have two variant IDs. The readiness score of variant 1 is 1 because all file types are supported. The readiness score of variant 2 is 0.33 because only one file type is supported.

1. Variant-Specific Relevance Score

The baseline score assumes that each data type is equally important to the artwork. As established above, depending on the purpose of the variant being made accessible, some data types might be crucial for the intended performance while others do not require support.

To make the readiness score more accurate, we can augment the baseline with human curatorial information. A human curator can examine each variant and assign a relevance score to file types on a Likert scale of 1-5, with 1 being not important at all (the viewer's experience will just be fine if this file type is not supported), and 5 being very important (the experience is meaningless without this file type being supported).

Essentially, for each variant, we can use the importance score as a weight to modify baseline readiness score.

In this example, the human curator will examine the variant 1, and assign scores of importance based on the artifact and machine combination.

Table III Variants with Curatorial Relevance Rating

Variant ID	Artifact ID	Artifact composition	Machine ID	Supported (inferred)		Relevance score (curatorial label)
1	1	doc	97	True		1
1	1	jpeg	97	True		1
1	1	mp3	97	True		5
2	1	doc	98	True		5
2	1	jpeg	98	False		1
2	1	mp3	98	False		1

1. Ignore the unknown files in the readiness score computation: this method is easy but may not be accurate.
2. Default “null” to false or true for consistency: easy to implement but may not be accurate.
3. Have human curators examine the unknown file, correct the file type if known, and assign a supported true/false label, as well as an importance score to override the unknown. This

Table IV Variants with Curatorial Relevance and Support Rating

Variant ID	Artifact ID	Artifact composition	Machine ID	Supported (inferred)	Supported (curatorial label)	Relevance score (curatorial label)
1	1	doc	97	True		1
1	1	jpeg	97	True		1
1	1	mp3	97	True		5
1	1	unknown	97	Null	False	5
2	1	doc	98	True		5
2	1	jpeg	98	False		1
2	1	mp3	98	False		1
2	1	unknown	98	Null	True	3

The curatorially supported readiness score can be calculated as follows:

$$readiness_score_{1artifacts,machine} = \frac{\sum int(supported_by_machine) * relevance_score}{\sum relevance_score}$$

In the toy example above, variant 1 would have a readiness score 1 of $(1+1+5) / (1+1+5) = 1$ (very good support) again; while variant 2 would have a readiness score 1 of $(1*5) / (5+1+1) = 0.7$ (good support), reflecting that the human curator has indicated that on the machine ID 98 environment, jpeg files are not relevant enough to require support. The augmented readiness score therefore incorporates curatorial knowledge and can be a more accurate estimate than the baseline readiness score.

2. Handling Unknown Data Types

As established above, a consistent way to handle unknown file types in the computing of the readiness score is essential for the digital art use-case.

In the case that file types are unknown, the automated pipeline to infer whether a file is supported would return null value. In these cases, we have a few options:

method is more label intensive, yet considered to be attainable in day-to-day collection care, and should provide the most accurate data and estimate of readiness score. (See Table IV.)

3. Computing and Presentation

Once a variant's readiness score is computed it can be stored as a property of that variant. On the user interface, the value can be used to draw the access quality stoplight indicator and for providing ranking and filtering functions.

Each time an element involved in the computation of this value changes—such as a new data format being detected in an artifact due to a PRONOM update, a software installed on a machine is found out to have different capabilities than originally thought, etc.—the value must be re-computed.

IV. FUTURE WORK

There is, of course, room to improve in terms of accuracy for the readiness score as defined above.

The most important omission in the current calculation is concerning the grade of completeness of available artifacts, which we plan to include in an upcoming version.

In addition, we plan to leverage timestamps as a source of improving data quality. Say if an artwork was created in the year 2000, and the machine contained software released earlier or much later,

we can infer that the machine might not support a file type even if the corresponding identifiers would match.

Finally, we can explore machine learning techniques to learn to parse the composition of artworks based on curatorial information. For example, we can try to predict the importance score of a file type—based on features such as file size, last modified time, and past curatorial importance labels. These computing techniques can further improve the efficiency of digital preservation staff and more quickly provide users with an access quality score.

REFERENCES

- [1] Rossenova, L. (2020) "ArtBase Archive—Context and History: Discovery Phase and User Research 2017–2019". Available from: https://lozanaross.github.io/phd-portfolio/docs/1_Report_ARTBASE-HISTORY_2020.pdf
- [2] Espenschied, D. 2021. "Digital Objecthood," in Selçuk Artut, Osman Serhat Karaman, Cemal Yılmaz, eds. Technological Arts Preservation, Sabancı University, Istanbul, 2021. ISBN 978-625-7329-16-3 <https://www.sakipsabancimuzesi.org/en/page/technological-arts-preservation>
- [3] Espenschied, D., Rechert, K. "Fencing Apparently Infinite Objects," in: Proceedings of iPRES 2018. DOI 10.17605/OSF.IO/6F2NM. <https://phaidra.univie.ac.at/view/o:923620>
- [4] Espenschied, D., Kreymer, I. "Oldweb.today: Browsing the Past Web with Browsers from the Past" in Gomes, D., Demidova, E., Winters, J., Risse, Th. (Eds.), The Past Web, Springer, 2021. ISBN 978-3-030-63290-8
- [5] Rossenova, L., de Wild, K., and Espenschied, D. "Provenance for Internet Art: Using the W3C PROV data model," in: Proceedings of the 16th International Conference on Digital Preservation iPRES 2019, Amsterdam, The Netherlands, pp.297-305. <https://osf.io/qc9u5/>
- [6] Rossenova, L (2020) "ArtBase Redesign—Prototype Development: Design Exploration and Evaluation 2018–2019". Available from: https://lozanaross.github.io/phd-portfolio/docs/4_Report_DESIGN_EXPLORATION_2020.pdf
- [7] Espenschied, D. "Artifacts and Infrastructure." Rhizome, 2021. <https://almanac.rhizome.org/pages/artifacts-infrastructure>
- [8] Giessl, J., Gieschke, R., Rechert, K. and Cochrane, E. "Automating the Selection of Emulated Rendering Environments for Born-Digital Data-Sets," in: Proceedings of the 25th International Conference on Theory and Practice of Digital Libraries, TPD 2021, Virtual Event, September 13–17, 2021. Springer. https://link.springer.com/chapter/10.1007/978-3-030-86324-1_12
- [9] Thornton, K. "Wikidata for Digital Preservationists. DPC Technology Watch Guidance Note". 2021. Digital Preservation Coalition. DOI 10.7207/twgn21-19

DIGITAL PRESERVATION PIPELINE FOR DATA STORAGE MEDIA AT THE CINÉMATHEQUE SUISSE

Imaging and extracting data and metadata from Special Collections media

Robin FRANÇOIS

*Cinémathèque suisse
Switzerland*

*robin.francois@cinematheque.ch
[0000-0001-5334-4438](tel:0000-0001-5334-4438)*

Rebecca ROCHAT

*Cinémathèque suisse
Switzerland*

*rebecca.rochat@cinematheque.ch
[0000-0003-3908-3854](tel:0000-0003-3908-3854)*

Abstract – Bit rot and technical obsolescence are threatening the ability to read data storage media received by GLAM institutions. This paper presents the work in progress to build a pipeline of autonomous steps to correctly preserve information on data storage media. Inspired by video game and computer science preservation communities, our pipeline relies on promising open-source software, such as Aaru or HxCFloppyEmulator, and hardware such as Pauline. Challenges and limitations of our approach are discussed.

Keywords – data storage media, digital archival imaging, migration, obsolescence, media dump
Conference Topics – Community; Innovation

I. INTRODUCTION AND RATIONALE

The mission of the Cinémathèque suisse (Swiss National Film Archives) is to collect and preserve film archives, with an emphasis on Swiss cinematic heritage. Such collected archives are of a wide variety and include data storage media, which need to be processed and accessioned as any other acquisition. Archive collections have progressively collected these types of media, but cannot apply the same treatment as an analog collection. Alternative methods have to be explored.

Such data storage media, exclusively computer readable, cannot be handled by documentalists and archivists as straightforwardly as human-readable media. It requires specific equipment – which might not be easily acquired or used due to technological

obsolescence – combined with technical skills that could pose a barrier for most staff members.

To streamline data storage media preservation, we have been building a pipeline. Analog media and digital audiovisual media (e.g., MiniDV, DVCAM, etc.) are excluded from this pipeline as they are handled by a separate team at the Cinémathèque suisse. This paper will focus solely on computer data storage media, such as floppy disks, optical disks or hard drives.

II. COLLECTION CHALLENGES

Data decay, also known as bit rot, can cause degradation to media that could complicate data reading [1]. Even if media can be correctly read, obsolete file systems and file formats can hinder any descriptive work. In addition, data storage media are data containers that can contain a very large amount of documents. The information stored on the ephemeral media should be copied and analyzed to prepare it both for accessioning and long-term archiving.

Data storage media are not human-readable and thus require a specific process for the necessary accessioning and descriptive work. Their longevity, even under good climatic conditions, has proven difficult to evaluate without extensive knowledge of the materials used [2] and the production process or the very precise examination of the media. This lack of detailed information has led most institutions to

consider data storage media as fragile and in need of intervention before their end of life.

However, to unearth the information contained within, a reading device and specific software are required. Both elements are subject to obsolescence and access to the data is often prevented, not due to the conditions of the media but by technical and technological limitations.

To prevent a blind spot in the collections and a digital dark age, time is a key factor and imaging should be performed as soon as possible, before hardware and software become scarce and decay damages the media.

While a survey is still ongoing, more than 5,000 data storage media have already been identified at the Cinémathèque suisse, with mostly optical media (CD and DVD), 3.5" floppy disks and ZIP disks. It is expected that this figure will at least double once the survey is completed.

III. IMAGING COMPUTER DATA STORAGE MEDIA

In order to properly preserve information on data storage media, the different data levels should be considered.

On digital media, the information is encoded as a sequence of binary values and stored using a physical property of the media (e.g., pits and lands for optical media). The devices required to read digital media will measure the variations in the physical property (the *signal*) and, through various operations performed by the device controller, generate a *bit stream*. The bit stream consists of *raw data* and control data (e.g., checksums, error corrections bytes, headers, etc.). Raw data can be interpreted as *user files and metadata* (e.g., modification date) through file systems and partitions [3].

As shown in Figure 1, the lower the level, the greater the amount of data, but the data is not always relevant. Accurately archiving media requires capturing all the necessary and meaningful data.

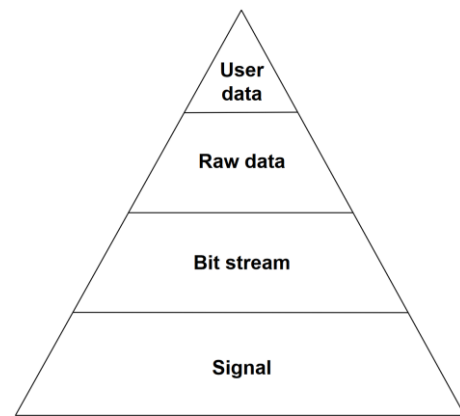


Figure 1 Levels of data

The level of necessary detail remains an open question, and it might be highly dependent on the institution and its missions. Initiatives such as DANNNG [4] are exploring and questioning the current situation.

Video game and computer science preservation communities have been approaching media imaging with great care and have developed methodologies and tools. These communities have been focusing on creating the most accurate images – sometimes to a very low level such as the signal level – and making the most of the processing in software *a posteriori*. This attention to detail has emerged due to two main factors: the importance of a near exact copy of the physical item and the copy protections that usually rely on peculiarities of the media.

IV. CHOICE OF HARDWARE AND SOFTWARE TOOLS

As Aristotle reportedly wrote, “He who can do more, can do less”. Our pipeline was greatly inspired by the video game and computer science preservation communities, notably the Game Preservation Society [5] and the dumping.guide project [6].

As part of a GLAM institution, we also had to comply with our IT department to ensure that our software and hardware choices would integrate with a modern IT infrastructure. The priorities were acquiring brand-new equipment where possible and installing a supported operating system.

In order to fit some internal drives, 5.25" slots were needed. New tower workstations from major brands (e.g., HP, DELL, Lenovo, etc.) were evaluated. With limited choice, an HP Z2 G4 tower was selected as it covered our hardware requirements

(connectors, storage, performances) and fitted with the IT department's strategy.

As for the operating system of our imaging PC, our choice was a Linux distribution for its wide interoperability and large toolbox. As recommended by our IT department, we are using CentOS Linux.

For the drives, priority was given to new and internal drives with current connectivity (i.e. USB), rather than used drives and obsolete connectivity. The following list of drives covers most media found in the Cinémathèque suisse collections:

- Internal SATA ASUS BW-16D1HT optical drive, picked for its broad compatibility with optical media and the selected software suite;
- External USB Iomega ZIP 750 (used), capable of reading ZIP drives of all capacities;
- SONY MPF 920 3.5" floppy disk drive (new, manufactured in April 2009).

For controlling the 3.5" floppy disk drive, we have decided not to use KryoFlux [7] for various reasons: it is not a community-led, open-source project; the hardware relies on USB connectivity that is prone to errors; and the development of the project seems to have slowed in recent years. Instead, we have selected the open-source hardware and software project Pauline [8], developed by a consortium of non-profit organizations focusing on video game preservation (MO5.com, La Ludothèque française and the Game Preservation Society).

The Pauline daughter board, plugged into a Terasic DE10-Nano FPGA [9] running a Linux distribution, becomes a standalone hardware solution for reading the signal from floppy disk drives with Ethernet connectivity. The solution uses a small web server to give instructions and the resulting files can be retrieved via network share.

Signal files generated by Pauline are in the hxcstream format and can be opened with HxCFloppyEmulator [10]. Based on the recorded signal, the HxC software can reconstruct the bit stream, calculate checksums, identify reading errors or damaged sectors, and export the raw data as a raw image to be further analyzed with additional tools.



Figure 2 HxC floppy disk track analyzer. Red parts indicate sectors with errors.

For media other than floppy disks and to manipulate raw images, we rely on the open-source Aaru Data Preservation Suite [11]. Aaru's philosophy is "to allow any user to create the best image (dump) that their hardware allows of the media they have" [12]. In addition, Aaru can be used to compare or convert existing images and list or extract files from an image with a broad range of supported image formats and file systems.

AaruFormat, Aaru's own image format, allows lossless compression and deduplication but also contains image metadata, contents metadata and dump hardware information.

Since Aaru is central to our preservation pipeline, the Cinémathèque suisse has joined the Technical Committee of the Aaru Data Preservation Suite project to provide input and feedback on roadmap items and to ensure that the open-source project thrives.

V. PRESERVATION PIPELINE

The main objectives of this pipeline are the following:

- Offload the archivists, documentalists, and the IT department of the digital archival imaging tasks
- Avoid the loss of fragile data on media due to physical and technological obsolescence
- Generate images of the media to allow the best possible software processing
- Provide the files and metadata to the archivists and documentalists to enable them to carry out their work
- Preserve and archive files in the best possible conditions to enrich digital collections and avoid the digital dark age

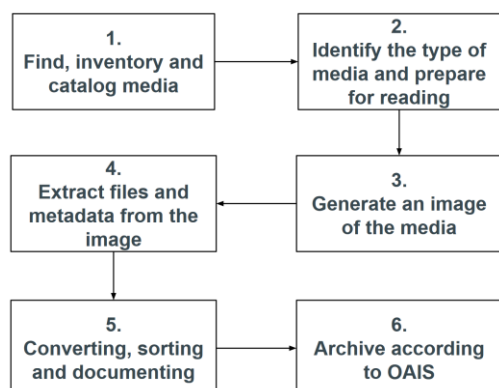


Figure 3 High-level view of the preservation pipeline for data storage media at the Cinémathèque suisse

Our preservation pipeline is summarized in Figure 3 in six steps.

The first step, “Find, inventory and catalog”, is very specific to our institution and its collections. Currently, data storage media have not always been identified and described, and are not often separated from the rest of the “analog” collections. Ongoing work consists of improving the handling of new acquisitions and applying the new procedures retroactively to the entirety of the Special Collections. Regrouping data storage media would allow batch processing, as well as improved conservation measures and packing. Attributing a unique identifier to each media is essential for further processing and associating the data extracted from the media to the right context.

The second step, “Identify the type of media and prepare for reading”, is necessary to address two issues. Media imaging requests might come from other teams that do not have full knowledge of our imaging capabilities, and some requests might require drives or tools that we do not have. In this case, an assessment will be made as to whether acquiring the necessary equipment and testing the methodology is relevant or if the imaging process is to be outsourced. The second issue is that precise media identification could need expertise, tools, and time that archivists and documentalists might not have. Cleaning and taking a picture of the media is also part of this step.

The third step, “Generate an image of the media”, consists of imaging the media at the lowest level necessary for adequate preservation, but also at the lowest level possible with the available drives and tools. In the current state of media drives, signal level imaging is only attainable for floppy disks. For other

data storage media, bit-stream level imaging is the usual limit of the drives, and raw level imaging is often satisfactory for most situations. Capturing metadata about the media and tools (paradata) is also performed at this step.

The philosophy of this step is very similar to digitization: create a digital surrogate of the media with the best accuracy and quality. Any intellectual processing can be performed in separate steps.

Once the best image possible has been made from the media, the fourth step is “Extract files and metadata from the image”. Images are rarely directly useful for archivists and documentalists to perform their work. Files and folders need to be extracted from the partitions and file systems stored on the image.

Interestingly, this step can be performed in a decoupled way from the imaging step and can be automated or performed by a separate staff member. We are using Aaru for this step, except for file systems not yet supported (see Chap. VI.).

Once files, folders, and metadata have been extracted from the image, the process of accessioning and digital archiving (following the OAIS model) begins: files need to be converted (normalized) to allow archivists and documentalists to sort, accession, and describe the documents that will be added to the digital collections and the OAIS archive.

The last step of the pipeline is the archiving of the selected documents according to the OAIS model. The description of this step is beyond the scope of this article.

VI. CHALLENGES AND LIMITATIONS

The Cinémathèque suisse is still investigating the best approach for appraising “digital bulk” extracted from some media, such as the personal 2TB hard drive belonging to a film director. Generally, extracting files from media can be a technical challenge, but adapting existing appraisal processes to digital assets is always a challenge.

Mass digitization is still being evaluated, notably due to the large amount of optical media in our Special Collections. The Acronova Nimble USB Plus automated loading system [13] is being considered, but compatibility with our current tools and pipeline needs to be tested.

Furthermore, the Aaru software has several limitations:

- Aaru is mainly a command line tool, which limits its use by less technical users.
- Imaging of large disks in AaruFormat is suboptimal due to bugs or missing functionalities. Imaging in QCOW2 is a good workaround while AaruFormatv2 is developed.
- The Cinémathèque suisse receives numerous media created in an Apple environment, with specific file systems. Aaru is currently not able to extract files from HFS and HFS+ file systems, which forces us to turn to a more manual approach using standard Linux tools.
- Metadata currently generated by Aaru needs to be transformed to fit into METS and PREMIS metadata used for digital archiving.

These limitations have been communicated to the Aaru project and should be addressed in future releases.

ACKNOWLEDGMENT

We thank our colleagues from the Cinémathèque suisse and the dedicated communities and institutions developing and supporting the Aaru, HxC and Pauline open-source projects.

REFERENCES

- [1] M. Baker et al., "A fresh look at the reliability of long-term digital storage", *SIGOPS Oper. Syst. Rev.* 40, 2006, pp. 221–234.
- [2] J. Iraci, "Longevity of Recordable CDs, DVDs and Blu-rays", Canadian Conservation Institute (CCI) Notes 19/1, 2019
- [3] B. Hayes, "Bit rot", *American Scientist*, 1998, vol. 86, no 5, pp. 410-415.
- [4] Digital Archival traNsfer, iNgest, and packagiNg Group, <https://dannng.github.io>
- [5] Game Preservation Society - <https://www.gamepres.org>
- [6] Dumping guide project by Hit-Save! - <https://dumping.guide>
- [7] Kryoflux - <https://www.kryoflux.com>
- [8] Pauline project webpage <https://www.laludotheque.fr/projets-en-cours/preservation-des-disquettes-pauline>
- [9] Terasic DE10-Nano FPGA product webpage <https://www.terasic.com.tw/cgi-bin/page/archive.pl?No=1046>
- [10] HxCFloppyEmulator, <https://sourceforge.net/p/hxcfloppyemu/>
- [11] Aaru Data Preservation Suite, <https://www.aaru.app>
- [12] N. Portillo, "Why is Aaru different?", Blog post, 24 July 2020 <https://blog.claunia.com/why-is-aaru-different/>
- [13] Acronova Nimble USB Plus product webpage, <https://disc.acronova.com/product/auto-blu-ray-duplicator-publisher-ripper-nimble-usb-nb21/9/review.html>

DATA CURATION AND AGROECOLOGY

Examining data requirements for short supply chains

Sarah Higgins

Aberystwyth University
United Kingdom
sjh@aber.ac.uk
[0000-003-1433-6923](tel:0000-003-1433-6923)

Christopher I. Higgins

ecodyfi
United Kingdom
chris@ecodyfi.cymru

Abstract – Digital preservation discourse tends to focus the organisational and technical processes required to make data, held in a recognised custodial environment, accessible and usable over the long-term. It rarely focuses on data needs requirements across the full lifecycle as defined by the *DCC Curation Lifecycle Model* (ref). This paper introduces the problem space for a project in mid-Wales which is taking a holistic approach to data curation and preservation. The *Tyfu Dyfi food, nature and well-being* project is supporting and developing agroecological practice in the UNESCO designated Dyfi Biosphere towards a more resilient and participatory local food system. Gaps in the creation and distribution of data necessary for successful collaborative food production and marketing are currently being identified. Next steps are the analysis of information flows across the partners to identify requirements for their long-term capture and access.

Keywords – digital curation, agroecology, food security, supply chains

Conference Topics – Environment; Resilience

I. INTRODUCTION

Data are created, stored, managed and accessed across all sectors of society making digital preservation and curation a cross disciplinary activity that is a meta-discipline of information science [1]. The widely cited foundational model adopted for the discipline, the *OAIS Reference Model* describes a custodial paradigm for data that is to be preserved [2]; while the *DCC Curation Lifecycle Model* looks outside the 'archive' to consider the full data lifecycle including the original creator and their creation of the data, and the use and reuse that might be made of it [3]. Best practice advice suggests that data curators should work with data creators in designing

a data architecture that is fit for use, reuse and preservation. Studies have used the *DCC's Model* to analyse post-creation requirements for improving data curation methods [4], [5], and research into co-creation of arts and humanities data curation have been undertaken [6] but only a few have worked within a commercial sector to understand their needs and ensure curatable data at the pre-creation stage. Notable in this space are Martin *et al.* [7] who are developing interoperable data architectures the curation of fisheries data towards their eco-system-based management.

This paper identifies agroecological supply chains as an information space that requires robust data modelling, and a data architecture that embraces the full lifecycle from creation to post-creation curation, access and use, to ensure the success of the sector in the marketplace. It describes how the *DCC Curation Lifecycle Model* will be used to structure research into data requirements for the sector, in tandem with anthropological research methods, to profile information needs as part of *Tyfu Dyfi*. This pilot project aims to contribute towards showing how food can be sustainably produced using agroecological practices, contributing to food security and resilience in a local community, while improving the area's biodiversity and reducing greenhouse gas emissions.

II. FOOD SYSTEMS AND AGROECOLOGY

Access to adequate and sufficient food is a human right [8] and in the UK the COVID-19 pandemic, Brexit and more recently the war in Ukraine have highlighted the need for sustainable food systems to safeguard this right for its citizens

[9], [10]. The complex global networks that ensure food availability can be readily disrupted and are generally outwith the control of those who need to eat [11]. Global food systems adopt factory style production methods which create large farm units, use chemical and intensive livestock production systems to maximize output, and prioritize quantity over quality or environmental protection [11].

Local food systems can create greater food security while also bringing better quality produce to the table. Community resilience is improved through local participation in food production and short supply chains, enabling appropriate responses to supply and demand and effective responses to national social priorities such food poverty and improved health and wellbeing [12], [13]. Local food systems can also play a part in mitigating the effects of climate change and biodiversity loss through their agroecological focus [12], [13].

Agroecology 'is a holistic and integrated approach [to agriculture] that simultaneously applies ecological and social concepts and principles to the design and management of sustainable agriculture and food systems' [14]. This transdisciplinary approach to food production 'includes the ecological, socio-cultural, technological, economic and political dimensions of food systems, from production to consumption' [8]. The values accrued by local food systems and agroecological methods are at the heart of the European Commission's *Farm to Fork Strategy* [10] and the reform of their Common Agricultural Policy (CAP) [15]. The UK's Department for Environment, Food and Rural Affairs' *Path to Sustainable Farming* acknowledges 'the connection between environmentally sustainable farming and an effective food supply chain' [16]; while The Welsh Government's report *Codesign for a Sustainable Farming Scheme* recognizes that farms need to be environmentally sustainable, the value of family run farms in protecting local cultures and the opportunities afforded by short supply chains [17].

III. FOOD SYSTEMS INFORMATION

In the era of 'big data', like other sectors of the economy (e.g., finance, manufacturing, and

healthcare), 'big' farming is now information driven. Precision technology creates data which is used to increase yields in both crop and animal production. Spatial data tools including remote sensing and geographical information systems (GIS) technologies are used to monitor environmental conditions such as weather, soil health and crop growth. Through licensing agreements for the technologies used, this data, although created by the farmer, is increasingly in the ownership and control of big technology companies [13]. Similarly, market access to food for the consumer is also driven by data and more crucially access to that data.

Agroecology also needs information systems to create successful food systems but those created for 'big' farming do not scale down, as the user needs are so different. The full information ecology for supporting agroecology has received little research attention. Sligo *et al.* [18] identified that small to medium scale farmers tend to rely on interpersonal relationships for their information, while Clavert-Mir *et al.* [19] notes that traditional knowledge, shared through informal networks, invaluable for agroecology, are fragile. Their *e-CONNECT Project* aims to capture these with 452 landraces – traditional or locally adapted species – documented. Several projects have focused on digitisation as a means to preserving published advice and research [20]–[22]. Meanwhile, a number of advice portals targeted at the sector aggregate information sources as advisory services. The Food and Agricultural Organization's (FAO) *Agroecology Knowledge Hub*¹ offers an international service that distributes information on research, legal frameworks, policies and programmes. In the UK, The Soil Association² offers a certification service and both advice and a brokering system between producers and markets for organic produce; while the Land Workers' Alliance³ offers support, guidance and training for 'agroecological and sustainable land-based sectors.' Garden Organic⁴ undertakes policy campaigning and research, offers an advisory service, and operates a heritage seed library.

The Open Food Network's⁵ offer is more immediately applied to the problem, providing an online marketplace to support the development of a

¹ Food and Agricultural Organization's (FAO) Agroecology Knowledge Hub: <https://www.fao.org/agroecology/policies-legislations/en/>

² Soil Association: <https://www.soilassociation.org/>

³ Land Workers' Alliance: <https://landworkersalliance.org.uk/>

⁴ Garden Organic: <https://www.gardenorganic.org.uk/our-work>

⁵ Open Food Network: <https://openfoodnetwork.org.uk/>

sustainable food network along with its advisory information. Crucially for data systems to support agroecology their national Lottery funded *Food Data Interoperability Project*⁶ is collaborating with the Data Food Consortium to develop open standards for technical and semantic interoperability that can 'be applied to tools for short food supply chain systems'.

IV. TYFU DYFI

Tyfu Dyfi - Food, Nature and Wellbeing is a pilot project supporting and developing agroecology in the UNESCO Dyfi Biosphere Reserve. Funded by the Welsh Government's Enabling Natural Resources and Well-being (ENRaW) scheme it is bringing together a range of partners and producers using agroecological methods to demonstrate 'how communities can be involved in their local food systems and enumerating the multiple benefits that accrue'.⁷ It aims to 'provide a national exemplar demonstrating how multiple organisations can cooperate on local food systems.' It is supporting and training people to grow and cook food, investigating the potential for a community led agriculture initiative, undertaking field scale trials of crops with local growers, and developing a local market for products. The project is led by ecodyfi, an NGO that co-ordinates activity across the Biosphere Reserve and partners with Aberystwyth University, Garden Organic and local organisations concerned with food growing, sustainability and food justice - Penparcau Community Forum, Aber Food Surplus, Mach Maethlon and the Centre for Alternative Technology.

The project aligns with the UNESCO Dyfi Biosphere Reserve's vision to 'be recognised and respected internationally, nationally and locally for the diversity of its natural beauty, heritage and wildlife, and for its people's efforts to make a positive contribution to a more sustainable world'.⁸ Designated in 2009, and home to a bi-lingual community of around 26,000 people, the continuing development of a locally-based economy which is 'more self-reliant and less carbon intensive, based largely on local culture, resources, products and environmental assets' is among its 5 principles.⁸

V. THE INFORMATION GATEWAY

Tyfu Dyfi builds on the results of the project *Mixed Farming: Histories and Futures*.⁹ This project, funded by the Welsh Government's LEADER Scheme and the Laura Ashley Foundation, was an ecodyfi led collaboration with Aberystwyth University, National Library of Wales and Environment Systems Ltd. It developed a local advice portal around suitable land and crops for those considering agroecology in the Dyfi Biosphere by identifying former agricultural practices to inform possible future practice. It used archival material and oral history, combined with layers of contemporary data from the National Forest Inventory,¹⁰ OpenStreetMap¹¹ and multi-temporal satellite imagery to build GIS based *Information Gateway* identifying former arable land, and opportunity maps of land suitable for reinstatement of arable practices. It focused particularly on land-use information captured in the apportionment schedules which accompanied the tithe maps created in the 1830s-1840s and digitized and transcribed through crowdsourcing by the National Library of Wales, to identify land that was designated arable before the opening of the railway line from 1862-1864 disrupted the local food markets. This was supported by data from the 1930s *Land Utilisation Survey of Britain* by Dudley Stamp of the London School of Economics and digitized by the *Great Britain Historical GIS Project* at the University of Portsmouth. Oral histories with older farmers in the area collected narrative memories of crop production before EU subsidies changed agricultural priorities and emphases towards livestock farming. Crops formerly grown, in fields, now given over to sheep and cattle rearing, included different cereals, swedes, potatoes, carrots and carrots.

Mixed farming also identified current key enterprises involved in producing and distributing fresh local produce, mapping their location against former arable land, enabling a start point for building networks in *Tyfu Dyfi*. This mapping activity revealed a range of small-scale food producers working both solo and in collaboration to market their produce in a small range of specialist shops. The organic

⁶ Food Data Interoperability Project: <https://about.openfoodnetwork.org.uk/introducing-the-food-data-collaboration/>

⁷ Tyfu Dyfi: <https://www.dyfibiosphere.wales/tyfudyfi>

⁸ UNESCO Biosphere Reserve: <https://www.dyfibiosphere.wales/dyfi-biosphere-wales>

⁹ Mixed Farming: Histories and Futures: <https://www.dyfibiosphere.wales/mixed-farming-histories-and-futures>

¹⁰ National Forest Inventory: <https://www.forestresearch.gov.uk/tools-and-resources/national-forest-inventory/>

¹¹ Open Street Map: <https://www.openstreetmap.org/>

vegetable bag scheme, run by *Tyfu Dyfi* partners Mach Maethlon, brought many producers together in this shared guaranteed market. This scheme orchestrates crop production across participating growers ensuring a variety of products for their customers.

VI. INFORMATION NEEDS REQUIREMENTS

1. First steps

The Open Food Network's marketing platform was piloted by Mach Maethlon as the Bwyd Dyfi Hwb (Dyfi Food Hub) for 3 months during the 2020 vegetable growing season. This brought together a total of 8 food producers and 8 buyers (shops and hospitality) in one information space. Producers and distributors were found to be in a vicious cycle where producers wouldn't sell on the online marketplace if distributors didn't buy, and distributors wouldn't buy unless there was a range of produce available. Additionally, an increase in production was identified as a criterion for success, which requires greater access to arable land than currently possible for the producers, or more producers entering the market.

This negative feedback cycle can be broken through timely data interventions and a mechanism for its creation, sharing, curation and analysis. Data gaps identified by the pilot for the development of a more resilient local food system can be seen in Table 1.

Table 1
Data gaps for agroecology in the Dyfi Biosphere

Actor	Data Gap	Data purpose
Producers	Land availability	Enables plans for new entrants or existing producers to expand
	Distribution systems	Ensures easy delivery to distributors
	Crops preferred by the market	Enables informed decisions on what to plant
	Market prices	
	Range and quantity of crops being produced by other producers	Enables co-ordination so that a range of produce is available
Distributors	Range and quantity of crops being produced	Enables informed decisions on what to buy
	Supply dates	

2. Next steps

The next step towards developing a local data ecology to support agroecology in the Dyfi Biosphere

is a more detailed analysis of the data requirements of both producers and distributors participating in *Tyfu Dyfi*. This will use an anthropological research strategy to gain a more granular understanding of the data gaps experienced and the datasets required. An anthropological research strategy is appropriate for illuminating the problem, as it uses mixed methods approaches to understand how people behave and their motivations [23]. This will facilitate appropriate requirements gathering towards the development of a data architecture which is fit for purpose.

Qualitative data will be collected through interviews and observation; and analysed to identify both further data gaps and the specifics of the datasets and data classes required to ameliorate each identified gap. This will be complemented with quantitative research around inputs and outcomes from the actors' activities to inform requirements for data flow between the different datasets needed to address the gaps. Focus groups with both producers and distributors will help to identify interoperability issues between the datasets required by the two groups.

The *DCC Curation Lifecycle Model* will then be used to benchmark each of the identified and specified datasets to ensure that all the actions it specifies can be adequately addressed. This will enable the further specification of a data architecture which is usable and accessible, and that methods can be put in place to ensure that the data created can be curated and preserved over the long-term to support ongoing needs.

VII. CONCLUSIONS

Data creation and its ongoing curation and preservation underpins much of the commercial sector. The data needs of global food systems have received attention due to their possible commercial exploitation. The agroecological sector's data needs have received less attention, but timely data interventions could ensure greater resilience of local food systems and help improve human health, and the degradation of the environment. *Tyfu Dyfi* has started to address this data gap at a local level and is moving forward with research to understand data requirements in more detail.

ACKNOWLEDGMENT

The Tyfu Dyfi Project and the Mixed Farming Project (and this research) were funded through the Welsh Government Rural Communities - Rural Development Programme 2014-2020, which is funded by the European Agricultural Fund for Rural Development and the Welsh Government.

REFERENCES

- [1] S. Higgins, "Digital curation: the development of a discipline within information science," *Journal of Documentation*, vol. 74, no. 6, pp. 1318–1338, 2018, doi: 10.1108/JD-02-2018-0024.
- [2] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS)," 2012. [Online]. Available: <https://public.ccsds.org/pubs/650x0m2.pdf>
- [3] S. Higgins, "The DCC Curation Lifecycle Model," *The International Journal of Digital Curation*, vol. 3, no. 1, pp. 134–140, 2008, [Online]. Available: <http://www.ijdc.net/index.php/ijdc/article/viewFile/69/48>
- [4] E. Fairley and S. Higgins, "Curated Databases in the Life Sciences: The Edinburgh Mouse Atlas Project," Digital Curation Centre, 2009. [Online]. Available: http://www.dcc.ac.uk/sites/default/files/documents/scar_p/SCARP_EMAP_Final13Jul09A.pdf
- [5] C. Rusbridge, L. Lyon, C. Neilson, and A. Whyte, "DCC SCARP disciplinary approaches to sharing, curation, reuse and preservation," 2010. [Online]. Available: http://www.dcc.ac.uk/sites/default/files/documents/scar_p/SCARP-FinalReport-Final-SENT.pdf
- [6] I. Faniel, A. Austin, S. Whitcher Kansa, E. Kansa, J. Jacobs, and P. France, "Identifying opportunities for collective curation during archaeological excavations," *International Journal of Digital Curation*, vol. 16, no. 1, p. 17, Apr. 2021, doi: 10.2218/ijdc.v16i1.742.
- [7] A. Martin, C. Chazeau, N. Gasco, and G. Duhamel, "IJDC | General Article Data Curation, Fisheries and Ecosystem-based Management: The Case Study of the Pecheker Database 2 | Data Curation, Fisheries and Ecosystem-based Management," *International Journal of Digital Curation*, vol. 16, no. 1, pp. 31–32, 2021, doi: 10.2218/ijdc.v16i1.674.
- [8] United Nations Human Rights, "The right to adequate food," 2010. Accessed: Mar. 06, 2022. [Online]. Available: <https://www.ohchr.org/Documents/Publications/FactSheet34en.pdf>
- [9] S. Coe *et al.*, "The effect of the war in Ukraine on UK farming and food production," *House of Commons Library*, July 2022. [Online]. Available: <https://researchbriefings.files.parliament.uk/documents/CDP-2022-0147/CDP-2022-0147.pdf>
- [10] R. Ranta and H. Mulrooney, "Pandemics, Food (in)security, and leaving the EU: What does the Covid-19 pandemic tell us about food insecurity and Brexit?" *Social Sciences & Humanities Open*, vol. 3, no. 1, pp. 2–5, 2021, doi: 10.1016/j.ssaho.2021.100125
- [11] C. Leroux, "Agriculture & digital: are we really going in the right direction?," *Aspexit Precision Agriculture*, Oct. 26, 2021. <https://www.aspexit.com/agriculture-digital-are-we-really-going-in-the-right-direction/> (accessed Mar. 08, 2022).
- [12] European Commission, "Farm to fork strategy: for a fair, healthy and environmentally-friendly food system," 2020. Accessed: Mar. 08, 2022. [Online]. Available: https://ec.europa.eu/food/document/download/472acca8-7f7b-4171-98b0-ed76720d68d3_en
- [13] A. Cooper *et al.*, "AgroEcoTech: how can technology accelerate a transition to agroecology?," no. July, 2021, [Online]. Available: <https://www.soilassociation.org/media/22821/agroecotech-soil-association-report.pdf>
- [14] Food and Agricultural Organization of the United Nations, "Agroecology Knowledge Hub." <https://www.fao.org/agroecology/overview/en/> (accessed Mar. 08, 2022).
- [15] European Committee of the Regions, "Agroecology: the answer to Europe's agricultural, social and environmental challenges," *The EU's Assembly of Regional and Local Representatives*, 2021. <https://cor.europa.eu/en/news/Pages/answer-to-agricultural-social-environmental-challenges.aspx> (accessed Mar. 08, 2022).
- [16] E. & R. A. Department for Food, "The path to sustainable farming: An agricultural transition plan 2021 to 2024," 2020. Accessed: Mar. 08, 2022. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/954283/agricultural-transition-plan.pdf
- [17] Welsh Government, "Co-design for a sustainable farming scheme for Wales," 2022. Accessed: Mar. 08, 2022. [Online]. Available: https://gov.wales/sites/default/files/publications/2021-09/sustainable-farming-scheme-co-design-future-farming_0.pdf
- [18] F. X. Sligo, C. Massey, and K. Lewis, "Informational benefits via knowledge networks among farmers," *Journal of Workplace Learning*, vol. 17, no. 7, pp. 452–466, 2005, doi: 10.1108/13665620510620034.
- [19] L. Calvet-Mir *et al.*, "The contribution of traditional agroecological knowledge as a digital commons to agroecological transitions: The case of the CONECT-e platform," *Sustainability (Switzerland)*, vol. 10, no. 9, 2018, doi: 10.3390/su10093214.
- [20] J. Oleen and L. Olsen, "Creating new partnerships: An examination of two collaborative, grant-funded digitization projects," *Journal of Agricultural & Food Information*, vol. 12, no. 3–4, pp. 370–376, Jul. 2011, doi: 10.1080/10496505.2011.619380.
- [21] D. C. Becker and K. M. Monks, "Slicing and dicing the digital preservation and dissemination of land grant research: a survey of and model for digital collections of agricultural experiment station and extension publications," *Journal of Agricultural & Food Information*, vol. 14, no. 3, pp. 225–241, Jul. 2013, doi: 10.1080/10496505.2013.793163.
- [22] A. L. Meger and D. Draper, "Digital preservation and access of agricultural materials," *Journal of Agricultural & Food Information*, vol. 13, no. 1, pp. 45–63, Jan. 2012, doi: 10.1080/10496505.2012.637437.
- [23] J. Fontein, "Doing research: fieldwork practicalities," in *Doing anthropological research: A practical guide*, 2nd ed.,

N. Konopinski, Ed. Taylor & Francis Group, 2013, pp. 70–90.

DESIGN PATTERNS IN DIGITAL PRESERVATION

Understanding Information Flows

Andrew N. Jackson

The British Library

UK

andrew.jackson@bl.uk

[0000-0001-8168-0797](tel:0000-0001-8168-0797)

Abstract – This paper proposes a framework to help understand the different ways digital preservation goals can be achieved, and the contextual factors these choices depend on. This is done through a worked example: three different design patterns representing the three possible modes of archival information flow, each illustrated with realistic examples and practices. By helping a wider audience understand these different approaches, we can ensure implementation choices are informed by practice, rather than by default, or without even realizing a choice has been made.

Keywords – OAIS, design patterns, risk management.

Conference Topics – Community; Innovation

I. INTRODUCTION

To a newcomer to digital preservation, it must seem like every question they have has one of two answers: “OAIS”, or “it depends...”. Standards like the Open Archival Information Systems reference model can provide useful high-level frameworks, but cannot provide concrete guidance when it comes to how to actually implement a digital preservation program. At the other extreme, over the years the digital preservation community has done a good job of sharing information on local implementations and workflows (through conferences like *iPres*, and via things like the *Community Owned Workflows* wiki [1]). These individual reports are a rich source of information, but synthesising this information is a very difficult challenge for a newcomer to take on, as they are least well equipped to know how relevant or current any particular piece of work might be.

The *NDSA Levels of Digital Preservation* [2] have helped bridge this gap from the top down, by providing a concrete roadmap of goals. But any such

universal path necessarily must remain abstracted away from the details that depend on context. And working from the bottom up, summaries of recent work, like the *Digital Preservation Coalition’s Technology Watch Publications* [3] provide very helpful in-depth summaries of known areas of interest. But this is still a lot to work through, and it remains difficult to quickly find and filter through the wide range of information on different approaches, and understand how the details depend on your context.

Is there some complementary format or framework that would allow us to navigate this gap, providing some kind of map that helps us find our place, and understand how our work should adapt to our context? In short, if the answer is “it depends...”, then the next question is “what does it depend on?”.

Here, I propose that a practice-driven approach is needed in order to answer this question. Specifically, one where concrete experience of implementing digital preservation processes are drawn together based on common patterns, thereby allowing the common contextual factors to be revealed. What follows are the results of analysing one aspect of digital preservation – the overall flow of information through the archive – within this “grassroots” framework. By publishing this here, I hope to get constructive feedback on whether this approach is useful, especially from newcomers to digital preservation. This will help shape future work on completing the analyses of a wider range of aspects of digital preservation.

II. PATTERNS OF INFORMATION FLOW

The Open Archival Information System (OAIS) reference model [4] is generally very high-level. There are a *lot* of different ways of running an archive that would still fit.

But one thing it is very clear about is the overall flow of information through the archive.

1. Flow 1: The Line

OAIS §1.4 - CONFORMANCE: "A conforming OAIS Archive implementation shall support the model of information described in 2.2"

OAIS §2.2.3 - INFORMATION PACKAGE VARIANTS: "Within the OAIS one or more SIPs are transformed into one or more Archival Information Packages (AIPs) for preservation. [...] In response to a request, the OAIS provides all or a part of an AIP to a Consumer in the form of a Dissemination Information Package (DIP)."

Under OAIS, Submission Information Packages (SIPs) come in from Producers, are managed as Archival Information Packages (AIPs), and these AIPs are then used to generate the Dissemination Information Packages (DIPs) that serve the needs of the archive's user community a.k.a. Consumers. This fundamental design pattern for digital archives – this way of describing the overall flow of archival information – can be visualised via this simple diagram:



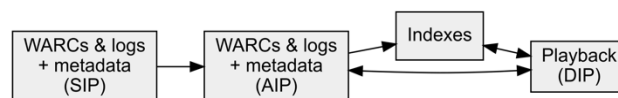
This linear flow is one of the strengths of the OAIS model, because by generating the DIP from the AIP, we ensure that all the information we might need to (re)create a new DIP is in the archive.

For an implementation to follow this pattern, it must ensure that *only* information that is in the AIP was used in the creation of the DIP, and there is no dependency on the SIP. OAIS does this by specifying that the DIP is generated from the AIP held on Archival Storage, after the SIP has been discarded. This could be immediately after ingest, or some time later, or on demand – the important thing is that it's post-ingest. To understand what this means in practice, it is useful to look at some examples of digital archives that work in this way.

1. The UK Web Archive

Like most web archives, the UK Web Archive generates access copies on demand. Before that can happen, the archived web resources are captured in

WARC files, which are then placed on archival storage before the necessary indexes are generated from them in order to enable access. When individual items are requested, the relevant records are looked up in the indexes before being copied from the AIP, and the access version (DIP) is generated from the archival version during the playback process.



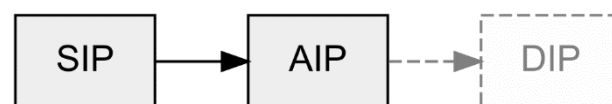
Because the transformation is done on demand, this makes it easier to modify how playback is performed in response to changes in web technology. In essence, we preserve the software that gives us the ability to generate access copies, rather than preserving the access copies directly.

The Internet Archive

While their web archive generates derivatives on demand (as above), the Internet Archive's documentation makes it clear that their archival system generates various derivatives for items after they have been ingested/uploaded, depending on the content. See for example this documentation on type of derivatives they generate [5], and this summary table showing all the format conversions they do [6]. The derivatives are stored rather than re-generated on demand, but doing this post-ingest means the derivation process can't block the process of ingesting items and ensuring they are safely replicated.

2. Flow 2: The Stop

The linear workflow described above may seem obvious, perhaps even inevitable, but it's not. Indeed, it's not unusual to find archives where there is *no outflow at all*. This corresponds to so called "dark archives" - ones that cannot be accessed except by the people who manage the archive. Usually, this is because the content is available elsewhere, and the archive is acting as a 'backup copy' in case the original goes away.



In these cases, the focus tends to be on ingesting the content, and not on how that content might be used in the future. There may be DIPs, but they are not well tested because no-one really uses them much. Or there may be no DIPs at all, or just some theoretical DIP based on an imagined future. The

danger here is that it's possible to miss some information you need from the SIP, or from the wider context, because you'd only realise you needed it when your user community started trying to access the material. Ideally, to mitigate against this risk, it is necessary to either encourage real usage, or at least simulate it, so that access problems can be identified while there is still some hope of resolving them. Otherwise, your AIPs are like a backup that's never been tested.

1. The UK Web Archive

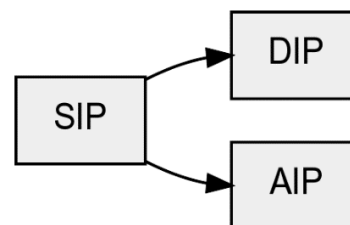
While the UK Web Archive has thousands of open-access websites, we also archive many millions of sites that can only be accessed under Non-Print Legal Deposit terms. This places heavy restrictions on access to those sites, and this combined with our limited capacity for manual quality assurance means it is unlikely that many of these archived pages will be viewed quickly enough to pick up any issues while we can still resolve them. One way we are trying to mitigate this risk is by starting to automate some of the manual QA processes. For example, for many years we have been collecting screenshots of how the websites looked when the web crawler originally visited them. The next step will be to add a post-ingest process that takes a screenshot of the *archived* web page in the same way as the crawler, allowing the images to be directly compared. This should highlight any issues and focus our efforts on where things need to be improved. We have also been working on publishing non-consumptive datasets based on our holdings. These are no substitute for the original web pages, but the surrogates are rich enough that meaningful research can be done with them, and this can help identify gaps in our collection.

2. Electronic Journals

Due to their importance and economic value, ejournals are often 'backed up' in dark archives like CLOCKSS or Portico. Usually, access to ejournals is via the publisher sites, and the archived copies only become active if the publisher shuts down. Here, AIPs usually contain all the SIP information, but ensuring the completeness of these packages is highly dependent on how fully the structure and content can be validated. This is particularly challenging when it comes to the handling of supplementary material, as this can come in a wide range of formats, or may only be supplied as a reference to resources held elsewhere.

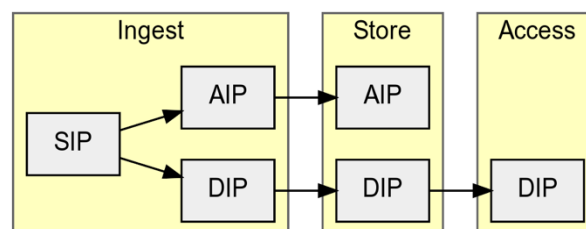
3. Flow 3: The Fork

The final class of information flow breaks the linear model entirely: while there is always a line from ingest to access, there is sometimes a fork in the road.



In this kind of split workflow, both the AIP and the DIP are generated directly from the SIP. In contrast to the linear flow, this design pattern *does not automatically ensure* that the AIP contains all the information we need to generate the DIP. This is not to say that this risk cannot be addressed, rather the point is that the implementation has to do *additional work* to make sure that this is the case.

One of the reasons why this can happen is that, during the design and implementation of archival workflows, it is common to try and perform a lot of processing up-front. Schematically, the result can look something like this:



Here I am using *Ingest*, *Store* and *Access* to identify processing contexts, covering all the functions associated with that phase of the workflow. During the *Ingest* phase, the incoming packages have been used to generate the DIP before the AIP has been transferred to the archive store. This means it is *possible* for the AIP and DIP to diverge.

In many cases, especially those where the critical payload of each package is a single file, this is a modest burden, because when the number of files involved is small, it's easier to keep track of individual files, their checksums, and their relationships. As things get more complex, it becomes easier for errors to creep in, especially if some network or system outage occurs while an item is being processed. These potential risks are certainly manageable - the point is simply that these risks arise because of the decision to perform more processing up front. This does not mean "The Fork"

is a *bad* pattern, indeed in those cases where we can confidently verify the packages are complete, this means we can pick up problems before items make it to the archival store. The downside is that if there are problems during ingest, this can lead to a backlog of material that spends far too long stuck on non-archival storage.

1. *An Outsourced Digitization*

This example involves a digitization project where much of the work had been outsourced to a third-party. The decision had been made to get the external partner to generate “access copies” as well as the long-term “preservation copies”. The preservation version consisted of high-resolution, full page TIFF files and associated metadata tying each set of TIFFs together, mapping back to the original publication. However, the access versions were not simply JPEGs of the preservation versions, because the items were broken up at the level of individual articles. This worked well for access purposes at the time.

The problem was the article segmentation process had not been properly documented. The coordinates on the TIFF version that corresponded to the positions of the JPEG versions had been lost. This meant that when the access system became obsolete, it was not possible to replace it like-for-like, i.e. while still preserving the same article-level experience.

2. *Preservation System Design*

When looking at tools and services that implemented digital preservation processes, it can be very difficult to tell exactly what’s going on *under the hood* and therefore difficult to determine what kind of information flow is being implemented. However, both [Archivematica](#) and [Rosetta](#) have a significant amount of documentation that is openly accessible online, so this allows us to gain some insight into how things are done.

Reading [the Archivematica documentation](#) [7] it is clear that Archivematica *usually* performs all processing up front, prior to ingest: the DIP is clearly generated directly from the SIP, but the system takes steps to ensure nothing is lost by making the original SIP form the basis of the AIP. Archivematica does make it possible to derive access copies in a post-ingest process as part of the ‘re-ingest’ workflow, but this is not the normal mode of operation. Rosetta also supports post-ingest generation of access copies via a [Create Derivative Copy Representation](#)

workflow [8], but like Archivematica, this appears to be seen as a secondary mode of operation, with ingest-time processing being the primary focus.

Note that this is not a criticism of these tools, which are merely implementing the workflow that their users have requested. Note also that it is not possible to draw comparisons with other preservation systems because this analysis is necessarily skewed towards those tools that provide detailed documentation online.

3. *Summary*

These three distinct information flow patterns show that there are fully-functional and widespread archival information flows that are not strictly OAIS conformant. This is not a problem with the archives or the software, but just alternative modes of operation that are not fully represented by OAIS as it stands. Each information flow pattern has its place, but it’s important to be aware of the balance of benefits and risks of each.

III. CONCLUSIONS

In this paper we have used the idea of design patterns as a way of capturing different choices that can be made when implementing a digital preservation process, focussing on overall information flow. The same tactic could also be applied to other aspects of digital preservation, such as:

- Communities: can we identify different classes of communities and environments, helping us understand how to engage with them?
- Ownership: is the archive part of the organisation that own the records, or do we hold records on behalf of others? How does this change what we need to do?
- Assessing Preservation Actions: what are the different meanings of Significant Properties, and what other methods can we use to assess our interventions?
- Archival Packaging Patterns: what are the different approaches to defining information packages.
- System Architectures: what are the different ways we can implement the OAIS functional requirements, i.e., which system or systems covering which functions?

This paper seeks feedback on this approach, and on whether future work along these lines would be of use to the wider community.

REFERENCES

- [1] Community Owned Workflows wiki. https://coptr.digipres.org/index.php/Workflow:Community_Owned_Workflows
- [2] Phillips, Megan, et al. "The NDSA levels of digital preservation: Explanation and uses." *Archiving conference*. Vol. 2013. No. 1. Society for Imaging Science and Technology, 2013.
- [3] Technology Watch Publications, Digital Preservation Coalition. <https://www.dpconline.org/digipres/discover-good-practice/tech-watch-reports>
- [4] Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-B-1; Consultative Committee for Space Data Systems: Washington, DC, 2002.
- [5] Files, Formats and Derivatives – Basic Guide, Internet Archive. <https://help.archive.org/help/files-formats-and-derivatives-a-basic-guide/>
- [6] Table of derivatives for different types of content. Internet Archive. <https://archive.org/help/derivatives.php>
- [7] Archivematica 1.13 Documentation – Ingest – Normalise. <https://www.archivematica.org/en/docs/archivematica-1.13/user-manual/ingest/ingest/#normalize>
- [8] How does Rosetta manage Derivative Copy representations? Ex Libris Knowledge Base. https://knowledge.exlibrisgroup.com/Rosetta/Knowledge_Articles/How_does_Rosetta_manage_Derivative_Copy_representations%3F

PASSIVE DIGITAL PRESERVATION ON PAPER IN PRACTICE

Vincent Joguin

Eupalia

France

vincent.joguin@eupalia.com

[0000-0003-0627-8778](tel:0000-0003-0627-8778)

Jean-Noël Dumont

Andra

France

Jean-Noel.Dumont@andra.fr

Abstract – Radioactive waste management encompasses timescales ranging from centuries to hundreds of thousands of years. Therefore, it is concerned with the long-term preservation of complex analog and digital information that are key to safeguarding the environment and human health against the potential threat of radioactive waste. The multi-century timescale and criticality of the radioactive waste repositories call for information preservation strategies based on durable, robust and secure data carriers and technologies.

Andra, the French national radioactive waste management agency, uses permanent paper to archive analog material. Over 2020 and 2021, it tested the Micr'Olonys solution from Eupalia, a novel approach to transcribe digital data on paper for passive long-term retention and accessibility.

This paper presents the results of transcribing a database concerning waste packages stored at the first repository operated by Andra, producing a 464-page document instead of a million pages that plain text printing would have required.

Keywords – database, passive preservation, permanent paper, radioactive waste management
Conference Topics – Innovation; Resilience

I. INTRODUCTION

Instability is the rule over the course of centuries, with wars, natural disasters, economic crises or simply societal changes taking place usually unexpectedly but regularly. Radioactive waste management is concerned with very extensive timescales and has therefore to cope with such disruptions. Preserving information pertaining to radioactive waste and the repositories where it is stored for centuries or more significantly contributes to keeping generations to come and the environment safe from any possible adverse impact, as well as giving the opportunity to revise how the

waste is to be managed over time. The richness and volume of the information concerned makes the challenge even more difficult: it can be partially preserved in analog form, but this form is unsuitable to some of it for which digital preservation needs to be considered. While mainstream preservation strategies based on migration are useful to maintain continued access to digital resources, they become simply inadequate when it comes to ensuring that material will still be accessible centuries from now, and therefore a complementary, more durable approach needs to be leveraged.

Andra, the French radioactive waste management agency, has adopted permanent paper (ISO 9706) as the reference material for long-term storage of records, after having noted that no maintenance contract with a provider of electronic solutions would be able to reach the scale of several centuries. On the contrary, permanent paper does not require maintenance and studies show that its durability in normal conservation conditions may reach at least 5 centuries. Until now, only analog contents (text, tables, pictures, diagrams) were considered for archival on paper, but this excluded more complex and larger born-digital content. Therefore, tests were conducted over 2020 and 2021 at Andra to evaluate how the Micr'Olonys solution developed by Eupalia could extend the applicability of permanent paper to digital content. Micr'Olonys was designed as a self-contained software-on-paper technology to store digital content in the form of 2D barcodes together with implementation-neutral means of access. The following sections describe the context and process of testing as it was carried out on a database relating to radioactive waste packages stored in a repository operated by Andra in the north of France.

II. PRESERVING A DATABASE OVER CENTURIES

A. *The issue of information preservation for radioactive waste repositories*

Radioactive waste repositories are designed to isolate waste from the living environment without human intervention over extended periods of time. No need for human intervention does not mean that oblivion will be looked for, quite the opposite: preserving the information, data and knowledge on the repositories for as long as possible is a shared objective of the various national radioactive waste management programs [1]. This has been addressed namely by a joint initiative under the framework of the Nuclear Energy Agency (OECD/NEA) called Records Knowledge and Memory preservation across generations (RK&M), that lasted from 2011 to 2018 [2] [3], followed by a working party, Information Data and Knowledge Management (IDKM) launched in 2020 and still ongoing.

During the first centuries after closure, awareness of a repository is necessary in order to ensure that no involuntary intrusion will occur, at the time when the decay of radioactivity has not sufficiently reduced the dangerousness of the waste. Another objective of information and data preservation is to allow current and future generations the possibility to understand the repository system and its performance and make informed decisions about the repository, even long after repository closure. For example, preserving information, data and knowledge will offer future generations the possibility to review and reassess the repository performance regarding protection of Man and the environment, provide modifications or retrieve material from the repository, if they consider it is necessary, without undue costs and risks.

The RK&M initiative identified a toolbox of 35 mechanisms that may contribute to information, knowledge and awareness preservation [3], among them a dedicated set of essential records (SER), collection of the most important records for waste disposal, selected for permanent preservation during the lifetime of the repository. The SER provides sufficient information for current and future generations to ensure an adequate understanding of the repository system and its performance. This will enable responsible parties to review and verify the repository performance and the safety case, and to make informed decisions,

even long after repository closure. The SER should serve as a source of detailed data and information on the repository system primarily for specialists and researchers, as well as for decision makers, regulators and other authorities. Selection criteria for the records to be part of the SER have been proposed by the RK&M expert group [4], based on a combination of relevance level (not relevant/nice to have/should have/must have) and effort it would take for a future generation to recreate the information if it were lost (some effort/extremely high effort). “Must have” information plus “Should have” associated with “extremely high effort” would be recommended for SER. Information regarding the radioactive content of waste packages is of course considered as part of the SER, but is the radioactive content of each waste package necessary? Reflections on the SER continue in the framework of IDKM, and the question of which data sets should be part of the SER is an issue that is being addressed.

B. *Information preservation at Andra*

In France, Andra launched in 2010 the “Memory for Future Generations” program in order to address in a more ambitious and structured way the issue of long-term information, knowledge and data management for radioactive waste repositories. This concerns both the near-surface repositories operated by Andra: the Manche disposal facility (CSM), the Aube disposal facility (CSA) and the Cigéo project of deep geological repository for intermediate level long-lived waste (ILLW) and high-level waste (HLW). The CSM, situated in the northwest of France, is the first repository operated by Andra. It received waste packages from 1969 to 1994. According to the regulation, CSM is in the “dismantling and closure phase” that will still last a few decades, before entering the “surveillance phase.” The surveillance phase should last at least 300 years, allowing most of the dangerousness of the waste to disappear due to the radioactive decay. The successor of CSM, still receiving waste packages, is the CSA. Being the repository at the most advanced stage of its lifecycle, the CSM acts as a pilot for actions related to memory preservation.

For all repositories classified as nuclear facilities, the French environment code requires that a set of records will be prepared by Andra, prior to entering the surveillance phase. This set of records, similar to the set of essential records proposed by the RK&M initiative, is called “Detailed Memory File” (DMF). In order to allow for long term preservation of the DMF,

the memory provisions adopted by Andra state that the DMF will be printed on permanent paper and stored in at least two different places. A preliminary version of the DMF of the CSM has been tested by a panel of external experts in 2012, providing lessons and conclusions that are presently used for its improvement [5]. The DMF contains of course a comprehensive description of the radioactive waste inventory, but at this stage not the database for each individual waste package, because printing this database on permanent paper would require an excessive amount of paper and of storage volume in the archives. The data are kept electronically, which is also the most useful for present needs; they will be migrated regularly, for at least as long as Andra will exist and manage repositories, which is expected to last still for more than a century.

The “Memory for Future Generations” program [6] aims both at implementing robust memory preservation provisions, based on the regulatory requirements but not limited to what is explicitly required, and leading R&D activities to extend the robustness and timescale of memory preservation. Apart from archives-related activities, it has developed a wide range of societal interactions, including for example artist contests and residencies, in order to spread in the society as largely as possible the awareness of the repositories and related documents. R&D activities are multidisciplinary, dealing with social sciences and humanities (e.g. semiotics or socio-anthropology) as well as landscape archaeology or materials sciences. Not to forget reflections on long-term preservation of digital archives, including the tests of database transcription on permanent paper presented in this publication.

III. PRELIMINARY TEST

Preserving the database mentioned above for each individual waste package stored at the CSM repository in analog form would translate into a printed document of about one million pages, a volume too large to be practical for long term archiving, but also for efficient access to relevant information. Therefore, Andra considered the Micr’Olonys solution developed by Eupalia to transcribe this database into digital form on paper.

Micr’Olonys prints a simple self-contained primer together with 2D barcodes (called “emblems”) that contain a digital file. The Micr’Olonys primer gives future users the ability to restore the information

without needing any specific technology such as particular hardware, operating system or programming language.

A general overview of the passive digital preservation approach and Micr’Olonys is provided in [7], while the Micr’Olonys technology is introduced in more detail in [8].

A preliminary test of Micr’Olonys was conducted at Andra in September 2020. A subset, consisting of 4,042 lines of data related to CSM radioactive waste was exported from the complete database that contains around 1,5 million lines into a Microsoft Excel XLSX file that was 1.7 MB in size.

For this test, the Excel file was then converted to the FODS format: Flat OpenDocument Spreadsheet. This uncompressed, fully textual format is supported by the LibreOffice suite. Its very explicit XML structure would be an advantage to make readability and reinterpretation easy over the long term. The test database subset produced a 41.8 MB FODS file.

This file was then compressed using a prototype version of the integrated Micr’Olonys compression algorithm to strip its size down to 111 kB (a 99.7% downsize). The very high compression ratio could be reached because of the strong redundancy of both the FODS format and the data the file contained. This file was transcribed into 3 emblems over 3 A4 pages: 2 emblems for data, and 1 redundancy emblem for critical error correction such as one missing page.

The French advanced prototype version of the Micr’Olonys primer was appended to these emblems to form a complete self-contained 15-page paper document. The primer consisted of about 3 pages written in plain French, 4 pages of pseudo-code written in simple French algorithmic language, a list of encoded letters over 2.5 pages, and 2 system emblems over 2 pages.

The 2 system emblems contain, in compressed form, all necessary code to parse and decode the emblems that contain user data, including the appropriate decompression and error correction functionalities. This code runs within the computing environment that implementation of the pseudo-code creates.

The 15-page document was handed over to a computer engineer at Andra with absolutely no prior knowledge of Micr’Olonys or what this document was about. The engineer was tasked with “making

something” out of this document, by supervising a 2nd year co-op student working at Andra that would take care of actual implementation.

Together, they managed to fully restore the original FODS database file in about two weeks of discovery and software development. In the process, they faced a few minor challenges that they eventually solved autonomously within a few days.

Two main shortcomings caused these challenges that were unforeseen by the Micr'Olonys authors. One is that recent changes to the Micr'Olonys primer structure created uncertainty as to how the software was meant to behave at some point once implemented, and gave the impression to the testers that their implementation was incorrect, when it was actually fully correct. The other is that the end condition and guidance to make use of the restored data was insufficiently explained, again giving the testers the impression that “something more” was to be expected, when actually they had reached successful completion of the decoding and restoration process.

Both shortcomings – and some other minor ones – were addressed with a subsequent update to the Micr'Olonys primer. It is expected that this improved update would accelerate implementation for other testers, shortening the necessary time to about one week for a junior programmer – and probably also for a senior programmer in the distant future.

This test demonstrated the relevance of storing digital information in such documents that make it possible for a junior software developer to restore the original data bit-for-bit without needing any specific hardware or software. Indeed, such a solution, in addition to being independent from today's technology, should also not assume current advanced know-how will remain readily available among professionals of centuries to come. Submitting the document to a beginner in programming, who has not yet acquired today's best practice and conventions, is to our mind the best way to come close to what the process of implementation by a future programmer is likely to be.

With the restoration process validated on a small test case, implementation of the Micr'Olonys solution at Andra continued in 2021 with the transcription of the complete database.

IV. END-TO-END DATABASE TRANSCRIPTION TO PAPER

The complete database consists of a main table containing around 1.5 million lines, and a few other much smaller tables containing complementary information and metadata. It was decided to transcribe each table individually so that each produced document would remain simple in its structure. Focus was on the main table as its size allowed for convincing validation of the approach, should it succeed.

The question of the stored file format was of course important. Making sure the format of the restored data will not hinder its reinterpretation, either manually or automatically, is a key element in the more general requirement of ensuring the whole document remains meaningful over time, without any modification or update to it whatsoever. The native Oracle format of the database was not an option as it is complex and proprietary. The richness offered by the FODS format selected for the preliminary test, and other XML-based formats, didn't seem relevant for a single flat database table. Therefore, the CSV format (comma-separated values) was selected for its simplicity and relevance. Moreover, it is a widely used and supported format within the database community, and the French National Archives routinely use this format to archive databases in their custody.

Extraction of the main database table into the CSV format generated a 626 MB file. The complete compression and decompression feature was implemented into the Micr'Olonys solution, both in the transcriber software and in the primer, the latter incurring absolutely no additional complexity. The compression feature is fully transparent to the user, both when transcribing data today and when restoring it in the future.

Using Micr'Olonys built-in compression, the 626 MB CSV file was brought down to 23 MB, a 96.3% reduction. This reduction was in line with expectations following the compression ratio of the preliminary test. This compressed data was then transcribed into 444 emblems in as many A4 pages, including redundancy emblems (3 emblems in addition to every group of 17 emblems).

Regarding the primer, it was decided that a bilingual document, written in both French and British English, would improve chances of its meaning coming out clearly in the distant future, when the form and practice of natural languages will have changed to an extent that is likely to impede

comprehension. Indeed, it is widely known that the Rosetta stone which presents the same message in two different languages, Ancient Greek and Ancient Egyptian, the latter with two different transcriptions, hieroglyphic and Demotic characters, was key for Champollion to deciphering the Egyptian hieroglyphic writing.

The primer was structured so that both language versions integrate seamlessly, referencing appropriate page numbers and common sections (the list of encoded letters and the system emblems).

The bilingual primer being 20 pages in length, the complete document produced for the main database table amounted to 464 pages, a very satisfying volume when compared with the estimated one million pages of the same information written in human-readable form.

In addition, the “emblematic” form offers other advantages compared to the analog form. For one, built-in error correction can easily make for small physical alterations to pages, while the same alterations to the analog form would mean irremediable loss of information, unless the document is replicated. Moreover, restoration from the emblems immediately producing information in digital form makes it possible to automatically process the data, in a trusted manner since the restoration software built in the primer will indicate the presence of even the slightest error of one byte, or its absence thereof. A comprehensive search functionality is also made possible very easily for a human user to quickly find relevant information within the mass of data. On the contrary, an analog form would translate into a burdensome and error prone process of Optical Character Recognition, or the necessity to manually browse through a million pages in search for the useful information.

Printing the document of 464 pages was carried out on a standard professional Toshiba multi-function laser copier. Micr’Olonys prints each page at the most common 600 dpi resolution, each emblem being formed strictly from black and white elements. From this document, the 444 data emblems were scanned using the automatic document feeder (ADF) of the same multi-function copier. Scanning was done at 600 dpi as bitonal (black and white without greyscale), again a very common and widespread configuration. Using the ADF means quick scanning, but also some geometric distortions as the pages are moved imperfectly for

scanning by the mechanical device. Fig. 1 & 2 below show such distortions.



Figure 1 A squeezed line, visible in the middle, resulting from imperfect scanning. This emblem is still read without any error.



Figure 2 A major distortion resulting from imperfect scanning.

Using the integrated Micr’Olonys reader software provided by Eupalia, a first decoding pass resulted in 17 of the scanned emblems with errors. The corresponding pages were rescanned but 4 remained in error. However, since they were distributed within different groups of 20 emblems, complete error correction could have been achieved, and a bit perfect file could have been restored at this stage. To complete the testing process, the remaining emblems were rescanned using the flatbed scanner, making sure the emblem images would not be cut at the edges, and using greyscale instead of bitonal. As a result, all emblems were finally correctly decoded by the Micr’Olonys software, and the stored file was restored. A file comparison with the initial file validated that the process was successful.

Using the primer algorithm implementation in C# that is integrated within the Micr’Olonys reader software, restoring the initial file took 3 h 10 min, for an average of 55 kB / s. This was executed on a laptop computer equipped with an Intel Core i5-6300U processor running at 2.5 GHz.

V. A COMPLETE TECHNICAL CHAIN TO ENSURE DURABILITY

A passive digital preservation solution needs to ensure that each and every aspect of it will stand the test of time, without any human intervention. It is a chain where any weak link endangers the whole process.

For the experiment described here, we used permanent paper, a medium that is able to last at least 500 years in normal conditions of conservation. Film is another medium with a similar life expectancy.

The Micr’Olonys primer is a short document that only makes use of fairly simple and widespread

concepts of computer science and document imaging that have been in use for decades. The description parts in natural language have been written using common words rather than specific technical words, so as to mitigate the risk of meaning shifting over centuries.

The CSV format was selected to store the archived information, CSV being a textual format with a very simple and straightforward structure that can be easily described.

Finally, an opening part puts the document in context. A relevant UTF-8 table (i.e., one only presenting those characters actually present in the stored document) is the key to converting the digital information into information understandable by humans, while the CSV format specification is the key to making the data structure explicit.

Of course, the stored data should be put in context: what does it describe or represent, how should it be used and in which context, is it related to other documents or information, etc.

VI. CONCLUSION

Storing digital information in a passive way, one that does not require regular human intervention and physical refresh, is mandatory when considering multi-century timescales. The radioactive waste management sector is concerned with such timescales (and more extended ones) and is mandated to preserve and transmit critical information over to future generations.

In this paper, we presented how Andra, the French radioactive waste management agency, worked together with the Eupalia company to test and validate Micr'Olonys, a passive digital preservation solution that extends the applicability of permanent paper to digital content.

During the course of 2022, the collaboration will continue with the testing and validation of Micr'Olonys on microfilm. Indeed, information density significantly improves with microforms when compared to paper. Moreover, the completely different physical natures of film and paper has the potential to open up a more robust passive preservation strategy by using both media in combination.

Applying Micr'Olonys to other emerging passive digital preservation media, such as ceramic and DNA, is also being actively pursued. In particular, these

media have the potential to offer greater longevity and density over paper and film, and would further diversify digital preservation media to increase overall durability and accessibility, thus improving resilience of the approach.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231954, as well as from Bpifrance, and was supported by the technical industries financial instrument of the Centre national du cinéma et de l'image animée (CNC).

REFERENCES

- [1] J.-N. Dumont, S. Wisbey, S. Hotzel, A. Berckmans, and A. Claudel, "Analysis of the Needs for Long Term Memory and Knowledge Preservation Relating to Radioactive Waste Disposal Facilities," *WM2017 Conference*, March 5 – 9, 2017, Phoenix, Arizona, USA.
- [2] S. Wisbey, et al., "Development of an Integrated System for the Long-term Preservation of Records, Knowledge and Memory relating to Radioactive Waste Disposal Facilities," *WM2017 Conference*, March 5 – 9, 2017, Phoenix, Arizona, USA.
- [3] Radioactive Waste Management and Decommissioning, "Preservation of Records, Knowledge and Memory Across Generations: Final Report," OECD/NEA, 2019.
https://www.oecd-neo.org/jcms/pl_15088
- [4] Radioactive Waste Management and Decommissioning, "Preservation of Records, Knowledge and Memory (RK&M) Across Generations: Compiling a Set of Essential Records for a Radioactive Waste Repository," OECD/NEA, 2019.
https://www.oecd-neo.org/jcms/pl_15090
- [5] J.-N. Dumont and G. Martin, "The First International Decennial Appraisal of the Detailed Memory of the Manche Disposal Facility," *WM2014 Conference*, March 2 – 6, 2014, Phoenix, Arizona, USA.
- [6] J.-N. Dumont, P. Charton, and F. Boissier, "Andra's Strategy and Approach for Long-Term Memory Preservation of a Deep Geological Repository," *WM2015 Conference*, March 15 – 19, 2015, Phoenix, Arizona, USA.
- [7] V. Juguin, "Passive Digital Preservation Now & Later: Microfilm, Micr'Olonys and DNA," *iPRES 2019*, Sep 2019, Amsterdam, The Netherlands.
https://ipres2019.org/static/pdf/iPres2019_paper_139.pdf
- [8] R. Appuswamy and V. Juguin, "Universal Layout Emulation for Long-Term Database Archival," *CIDR 2021*, Jan 2021, Chaminade, USA.
http://cidrdb.org/cidr2021/papers/cidr2021_paper30.pdf

ROBUSTIFYING LINKS WITH ZOTERO

Martin Klein

*Los Alamos National Laboratory
USA
mklein@lanl.gov
[0000-0003-0130-2097](tel:0000-0003-0130-2097)*

Shawn M. Jones

*Los Alamos National Laboratory
USA
smjones@lanl.gov
[0000-0002-4372-870X](tel:0000-0002-4372-870X)*

Abstract – Referencing resources on the web has become an integral part of our digital scholarship. However, the long-term availability and accessibility of these resources has rarely been the focus of significant research and development efforts. In this paper we introduce the Zotero Robust Links extension, a tool that helps authors create archival copies of referenced web resources and offers robustified HTML code that can easily be copied into manuscripts. Our goal with this extension is to provide a tool that helps authors contribute to the integrity of the scholarly record.

Keywords – Reference Rot, Web-based Scholarly Communication, Zotero, Web Archiving

Conference Topics – creative solutions to shared challenges in daily practice; insight-oriented institutional and personal progress towards digital preservation goals

I. INTRODUCTION

References in scholarly articles are essential to provide supporting arguments, complementary information, and pointers to related work. Increasingly, these references have come in the form of HTTP links to resources on the web. A large body of research, for example [1-5], has provided ample evidence that the web is a very dynamic space with resources frequently being created, deleted, and moved. While missing resources and the often encountered infamous “404 - Page not found” error are a detriment to a user’s browsing experience, the issue is just as severe in the realm of scholarly communication. Broken references in scholarly articles are a hindrance to reproducibility and a risk to the integrity of the scholarly record. In previous work [6], we have shown the increasing use of links to web resources and quantified the ratio of links that are subject to link rot (scenario where the link

does not work anymore, often resulting in a 404 error). Additionally, we have demonstrated that, over time, a large percentage of links are subject to content drift (scenario where the content of a linked page has changed significantly) [7]. As part of the Hiberlink Project¹, we coined the phrase “reference rot” as the combination of link rot and content drift.

Persistent identifiers (PIDs)² such as Digital Object Identifiers (DOIs)³ have been introduced to address this problem. Jones et al. [8] provided an in-depth analysis as to why DOIs insufficiently address the issue of reference rot in scholarly communications. Klein et al. [9] highlighted inconsistencies with resolving DOIs. Van de Sompel et al. [10] quantified the observed lack of use of DOIs. Aside from these shortcomings, we have shown that authors of scholarly papers frequently reference web resources that do not have a PID assigned [6], such as blog posts, videos, and, more recently, source code [11].

Fortunately, the web archiving landscape offers services that support the long-term availability and accessibility of web resources. Specifically, the Internet Archive’s “Save Page Now” (SPN), archive.today, perma.cc, and megalodon.jp are examples of proactive web archiving services that allow authors to create archival copies of the web resource they intend to reference. Our prior work [8] notes that some authors attempt to address the reference rot issue by first creating an archival copy. These authors then use that archival copy’s URI in their reference rather than using the URI of the original live web resource. A popular example is the Internet Archive’s link rot bot for Wikipedia⁴. However, this approach relies on the permanent

¹ <http://hiberlink.org/>

² <https://www.dpconline.org/handbook/technical-solutions-and-tools/persistent-identifiers>

³ <https://www.doi.org/>

⁴ https://en.wikipedia.org/wiki/Wikipedia:Link_rot

existence and availability of one web archive, the one at which the archival copy was created - a scenario in which one link rot problem was merely replaced with another, as many types of disasters, technical, financial, and political, can befall individual web archives.

Analyzing this problem space motivated our creation of the “Robust Links” concept [12], which consists of two main steps:

1. proactive creation of an archival snapshot of the web resource to be referenced and
2. robustifying the reference by enhancing the HTML link with defined attributes.

The first step creates a version of the resource that is representative of what the author intends to reference and the second provides a number of fallback mechanisms in case the resource on the live web is not accessible (link rot), its content has changed (content drift), or its archival copy is unavailable from the one archive it was created in. For a more detailed overview of the Robust Links concept and its HTML attributes we refer to the Robust Links specification⁵.

To further the adoption of the Robust Links concept, we implemented a web service⁶ that supports users in conveniently executing the two steps outlined above [8]. The web service builds upon an API we designed and implemented⁷ to also allow for creating robustified links in bulk.

While the Robust Links web service targets individual scholars and authors of web-based manuscripts, it represents yet another system that needs to be included in the research life cycle and in the suite of tools researchers utilize in their daily work. We noticed from discussions with our target audience that there is a strong desire to incorporate such preservation and robust linking approaches into existing and widely used tools. Zotero, the popular open source reference manager, had been mentioned as an example of such tools⁸.

In this paper we introduce our Zotero Robust Links extension. We walk through its functionality and intended use and discuss some aspects of improvements planned for future work. With this paper and the introduced software extension, we hope to further raise awareness about the problem

of reference rot in scholarly communication and highlight one possible path to address the issue - by robustifying links with Zotero.

II. ZOTERO AND ROBUST LINKS

A detailed overview of Zotero’s functionality is beyond the scope of this paper but we refer to its documentation⁹ for a thorough overview. Generally speaking, Zotero comes as a desktop application, a web service, and a browser extension. Our extension is designed for the desktop application and we therefore focus our elaborations on this environment.

A. Adding Items

Zotero allows the ingestion of an item by its identifier, specifically a DOI, ISBN, PMID, or arXiv ID. Upon pasting the identifier into the search bar, Zotero connects to the hosting platform as well as 3rd parties such as Crossref to gather all available metadata about the to be ingested item. It then automatically populates the corresponding metadata fields for the created record, visible in the Zotero “Item Pane”. For a typical journal article this includes authors, title, date, DOI, and the URI. As is common with many reference formats, Zotero populates the DOI field with the DOI PID string and not the HTTP DOI. Unfortunately, in these cases, the URL field is populated with a paper’s landing page (final link of the DOI resolution redirect chain) rather than its HTTP DOI URL. As some reference formats only support either a DOI or a URL, this behavior can lead Zotero users to unknowingly prevent DOI adoption.

Zotero also allows users to add new items manually to a collection. For example, to add a blog post, the user can click the “New Item” button and select “Blog Post” as the type of item to be added. It is then up to the user to manually add appropriate and corresponding metadata to the Item Pane, including the URI of the resource.

The Zotero browser extension offers a third way to add items to a collection. By displaying a resource in the browser and clicking the “Save to Zotero”

⁵ <https://robustlinks.mementoweb.org/spec/>

⁶ <https://robustlinks.mementoweb.org/>

⁷ <https://robustlinks.mementoweb.org/api-docs/>

⁸ <https://trello.com/c/5Jk5bbxz>

⁹ <https://www.zotero.org/support/>

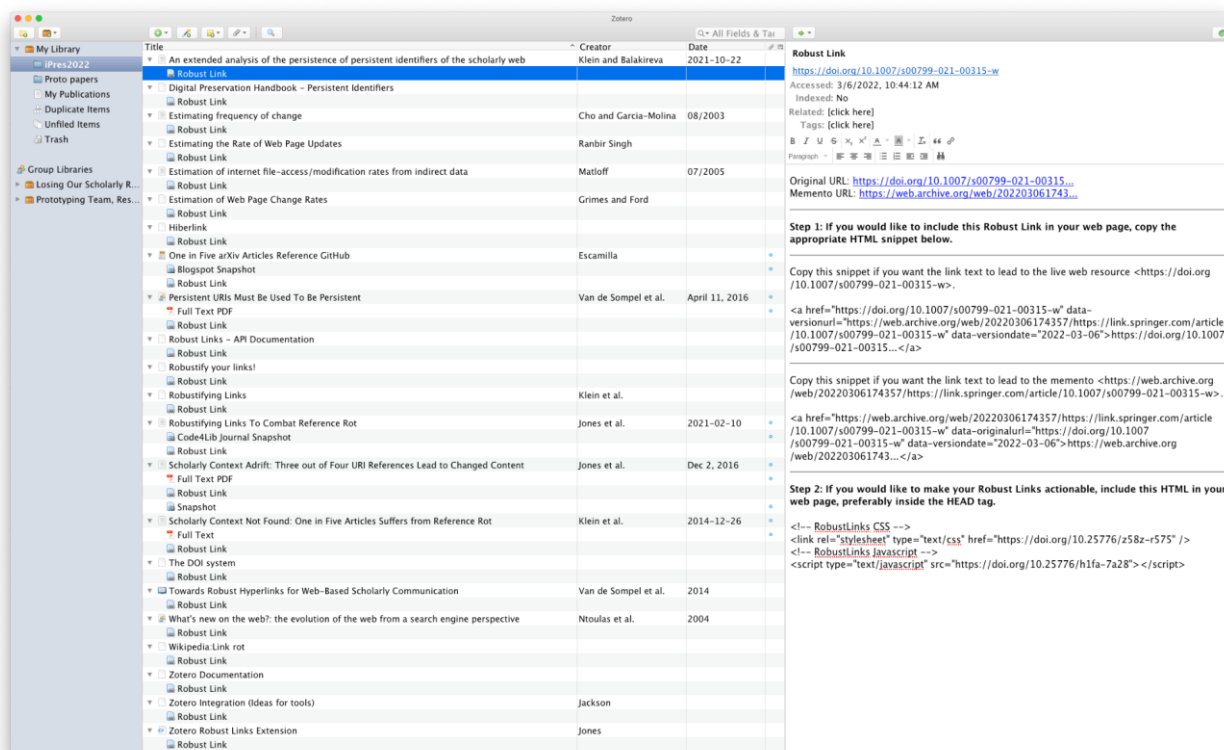


Figure 4 Zotero Collection (left), Items (center), Note (right)

button, a user can add the displayed resource to a selected Zotero collection. The ingestion process works automatically, very similar to the adding by identifier method, with all available metadata fields populated.

B. Robust Links in Zotero

By default, the Zotero Robust Links extension [13] automatically creates an archival copy of any item ingested via its identifier or via the browser extension, as long as Zotero had populated the URI field with a valid URI. It takes the URI and submits it to the Robust Links API, which, in turns submits the URI string to one of the web archiving services mentioned above. For manually created items, the Robust Links extension needs to be triggered manually. With a right mouse click on the item of interest, a user has the option to "Robustify this resource" and pick the default, any, or a specific web archive for the proactive archiving of the resource. In either event, the extension displays a notification with details regarding the URI and the web archive it is submitted to. The notification disappears after 5 seconds or after the user clicks on it.

C. Robustified Links

Once the archiving process is complete, the extension again shows a notification on the screen

with a message indicating success (or failure). This notification also disappears after 5 seconds or a mouse click. If an error occurs, the error message will remain and the user needs to click it to dismiss it.

Each item for which the extension has created a Robust Link now has a Zotero Note as a child node. The note, hierarchically aligned below the item, is named "Robust Link". It contains the original URI of the resource, the URI of its created archival record as returned from the utilized web archive, and instructions on how to robustify the HTML link in a manuscript, satisfying the second step of creating a Robust Link mentioned above. From these instructions, a user can simply copy and paste the robustified HTML code into a manuscript.

Fig. 1 shows a screenshot of the Zotero main window. All collections are shown on the left in the "Collection Pane". We have created a collection called "iPres2022" that contains all references from this paper. The items and their corresponding notes are shown in the center of Fig.1 and the content of one of the Robust Link notes is shown in the "Item Pane" to the right.

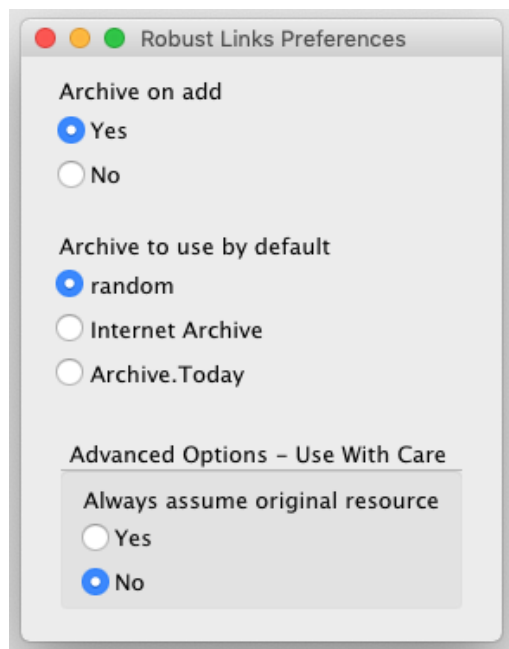


Figure 5 Configuration Panel.

D. Extension Configuration

The Zotero Robust Links extension comes with a number of default settings upon installation on a system. Via the “Tools” menu item, a user can access the configuration panel, displayed in Fig. 2. Recall that references added to Zotero will be automatically robustified if they contain URLs. The user can disable this functionality by toggling the corresponding setting on top of the panel. The extension’s default archive is “random”, which means the Robust Links API will randomly create a copy at either of the two currently available web archives (Internet Archive and archive.today). A user may prefer a specific archive, and thus can specify their preference. If this setting is left unchanged, the options of “default” and “any” web archive, mentioned in the manual processing above, are identical. The setting on the bottom of the configuration panel is aimed at advanced users. By default, the extension asks the Robust Links API to check if a submitted resource is already an archived version before creating a new one. With this setting, a user can ask the API to bypass this step and just create a new archival snapshot, potentially saving time.

E. Exporting Robustified Links

Zotero supports the export of all or individual items in a collection. For example, a user can right-click on a collection in the “Collection Pane” and

export the entire collection (all items in the collection). Alternatively, a user can select one or more items and export them with a right click. The export panel, shown in Fig. 3, prompts the user to pick a format. Behind each of these formats is a corresponding Zotero “Translator.” We maintain a special translator for Robust Links [13] that a user can install separately from our extension. This Robust Links translator allows Zotero to export items into an HTML file that contains each item’s corresponding robustified link. For demonstration purposes, we exported our entire iPres2022 collection containing all references of this paper, and made it publicly available¹⁰.

III. FUTURE WORK AND CONCLUDING REMARKS

Limitations to the extension remain. For example, Zotero allows for the creation of a bibliography of items with a chosen citation style, for example, that of the American Chemical Society. In the future we will analyze various citation styles and assess how to incorporate robustified links. Currently the extension only supports two web archives. We are negotiating the addition of other web archives, such as perma.cc, to increase the chances for each reference’s long-term availability. Lastly, Zotero Notes are currently the best way to convey robustified HTML. In future Zotero versions, it may be possible to actively edit metadata in the “Items Pane”, potentially offering better options to inform the user about the robustified links.

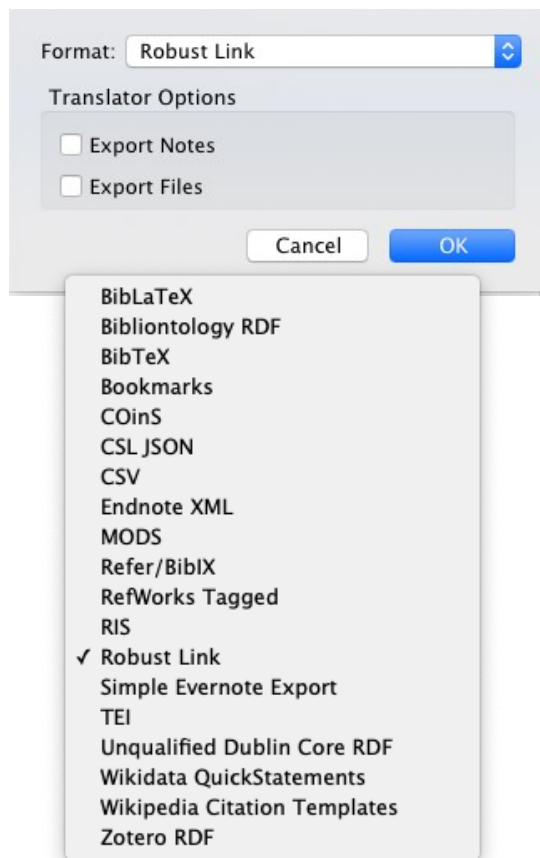
Our goal is a tool for manuscript authors that can help them be better stewards of their references and scholarly articles at large. With our extension to the popular reference manager Zotero, we aim to meet the researchers where they are rather than creating yet another tool for them to learn. We hope that adoption of our extension will help robustify more links to support the integrity of the scholarly record.

REFERENCES

- [1] Norman Matloff. 2005. Estimation of internet file-access/modification rates from indirect data. *ACM Trans. Model. Comput. Simul.* 15, 3 (July 2005), 233–253. DOI:<https://doi.org/10.1145/1103323.1103326>
- [2] Junghoo Cho and Hector Garcia-Molina. 2003. Estimating frequency of change. *ACM Trans. Internet Technol.* 3, 3

¹⁰

<https://robustlinks.mementoweb.org/zotero/iPres2022.html>



(August 2003), 256–290.

Figure 6 Export panel.

DOI:<https://doi.org/10.1145/857166.857170>

- [3] Alexandros Ntoulas, Junghoo Cho, and Christopher Olston. 2004. What's new on the web? the evolution of the web from a search engine perspective. In Proceedings of the 13th international conference on World Wide Web (WWW '04). Association for Computing Machinery, New York, NY, USA, 1–12. DOI:<https://doi.org/10.1145/988672.988674>
- [4] Carrie Grimes, Daniel Ford. 2008. Estimation of Web Page Change Rates. <https://storage.googleapis.com/pub-tools-public-publication-data/pdf/34428.pdf>
- [5] Sanasam Ranbir Singh. 2007. Estimating the Rate of Web Page Updates. <https://www.aaai.org/Papers/IJCAI/2007/IJCAI07-462.pdf>
- [6] Klein M, Van de Sompel H, Sanderson R, Shankar H, Balakireva L, Zhou K, et al. (2014) Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. PLoS ONE 9(12): e115253. <https://doi.org/10.1371/journal.pone.0115253>
- [7] Jones SM, Van de Sompel H, Shankar H, Klein M, Tobin R, Grover C (2016) Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content. PLoS ONE 11(12): e0167475. <https://doi.org/10.1371/journal.pone.0167475>
- [8] Jones, Shawn M., Martin Klein, and Herbert Van de Sompel. "Robustifying Links To Combat Reference Rot." The Code4Lib Journal, no. 50 (February 10, 2021). <https://journal.code4lib.org/articles/15509>.
- [9] Klein, M., Balakireva, L. An extended analysis of the persistence of persistent identifiers of the scholarly web. Int J Digit Libr (2021). <https://doi.org/10.1007/s00799-021-00315-w>

- [10] Herbert Van de Sompel, Martin Klein, and Shawn M. Jones. 2016. Persistent URIs Must Be Used To Be Persistent. In Proceedings of the 25th International Conference Companion on World Wide Web. Geneva, CHE, 119–120. DOI:<https://doi.org/10.1145/2872518.2889352>
- [11] One in Five arXiv Articles Reference GitHub. <https://ws-dl.blogspot.com/2022/02/2021-02-23-one-in-five-articles.html>
- [12] Van de Sompel H., Klein M., Shankar H. (2014) Towards Robust Hyperlinks for Web-Based Scholarly Communication. In: Watt S.M., Davenport J.H., Sexton A.P., Sojka P., Urban J. (eds) Intelligent Computer Mathematics. CICM 2014. Lecture Notes in Computer Science, vol 8543. Springer, Cham. https://doi.org/10.1007/978-3-319-08434-3_2
- [13] Jones, S., & Klein, M. (2021). Zotero Robust Links Extension: Create Robust Links from within Zotero (Version 2.0.0-20210419153723) [Computer software]. <https://github.com/lanl/Zotero-Robust-Links-Extension>

EVALUATING DIGITAL PRESERVATION CAPABILITY WITH LARGE AT-RISK COLLECTIONS

Lessons learnt from preserving the NVA Archive

Leo Konstantelos

University of Glasgow
United Kingdom
Leo.konstantelos@glasgow.ac.uk

Clare Paterson

University of Glasgow
United Kingdom
Clare.Paterson@glasgow.ac.uk

Emma Yan

University of Glasgow
United Kingdom
Emma.Yan@glasgow.ac.uk

Abstract – This paper presents the efforts of the Archives & Special Collections (ASC) unit at the University of Glasgow to preserve the large, at-risk collection of the NVA Archive. We discuss the nature of the collection, and the way it was used to evaluate our digital preservation capability.

Keywords – Digital preservation, at-risk collections, cultural heritage, University archives
Conference Topics – Community

I. INTRODUCTION

The Archives & Special Collections (ASC) unit at the University of Glasgow Library is responsible for “managing, promoting, enabling access and supporting engagement with the Library’s unique and distinctive collections”¹. ASC collects and provides access to archival records, manuscripts, rare books, and other primary and secondary sources to support teaching and research at the University and the wider community. The University has been an early adopter of digital records and processes [1] – and following an ongoing digital transformation as part of the University’s “World-Changing Glasgow Transformation” initiative² ASC has been increasingly collecting born-digital records of historical, cultural and business significance.

This paper will discuss our efforts to build a robust digital preservation service through the lens of our work to preserve one digital collection with continuing value to the Scottish arts community (and beyond), the NVA Archive. NVA – an acronym for ‘nacionale vita activa’ (roughly translated as ‘the right

to influence public affairs’) [2] – was an internationally renowned arts organisation based in Glasgow, which closed in 2018. NVA epitomized community participation through public art that “reconnects people to their built and natural environment” taking inspiration from the “Ancient Greek ideal of a lively democracy, where actions and words shared among equals bring new thinking into the world.” [3]

The size of the NVA digital records, and the impetus to secure their ongoing existence after the closure of the organisation, pushed our (rather nascent) digital preservation processes and systems to their limits; but also helped us evaluate our capability to preserve large, at-risk digital collections.

II. ABOUT THE NVA ARCHIVE

NVA was founded in 1992 by Angus Farquhar, and has produced a number of key public art projects including *The Hidden Gardens* in Glasgow, *The Storr* on the Isle of Skye, the international touring work *Speed of Light*, and the *Kilmahew/St Peter’s* project based around the rescue of the modernist ruin of St Peter’s Seminary³.

In September 2017, the Directors of NVA (Europe) Limited made the decision to withdraw from plans to rescue St Peter’s Seminary. Over the following months, attempts were made to develop an alternative proposal for the building and to stabilize the organisation, but it became clear that this was not possible. Additionally, in February 2018, NVA

¹ www.gla.ac.uk/myglasgow/archivespecialcollections/

² www.gla.ac.uk/myglasgow/worldchangingglasgow/

³ Information on NVA artworks available at: <http://nva.org.uk/artworks/>

received notification that its bid to Creative Scotland for core revenue funding had been unsuccessful. The scale of the financial challenges had become untenable and, in June 2018, the Board of NVA announced that it was to close.

ASC approached the Business Archives Surveying Officer at the Ballast Trust - a charitable foundation that provides a rescue, sorting and cataloguing service for business archives⁴ - for assistance in securing temporary storage and a records survey for the records, as the collection was deemed at-risk due to the NVA's imminent closure. In August 2018, the physical records along with a hard drive of digital records were transferred to the Ballast Trust, where they were box-listed by the Surveying Officer.

The digital records amount to approximately 800GB, of which 300GB are images. The remaining records consist primarily of word documents, excel spreadsheets, PDFs, and images in various file formats (see Table 1). The files themselves include correspondence, minutes, financial records, administrative records, staff records, images, video, audio, project plans, project evaluation, tender and funding bid documents, marketing, and promotional materials.

Table 1 Volume and composition of the NVA digital archive

Data volume	782GB	
File formats	40,000 image files	150 audio files
	11,579 PDF	359 video files
	19,501 documents	878 Adobe files
	5,820 spreadsheets	117 AutoCAD files
	248 presentations	337 zip files

III. COMMUNITY AND ARCHIVAL VALUE

NVA has been at the heart of the Scottish contemporary art world for over two decades. The organisation had a track record of producing large scale public performance artworks which thousands of people in Scotland (and internationally) have engaged with. Their last, unfinished project, Kilmahew/St Peter's, involves Scotland's best known modernist building and one of the first 'modernist ruins'. This, along with the enduring legacy of many of their past works, makes it highly likely that the records of NVA will be of significant interest to art and architectural historians, geographers and the art and cultural sector. Recognizing this value, our primary aim has been to preserve all that we can of the NVA archive. This approach will provide an

opportunity for a community of voices to inform decision-making around appraisal, description, and access. It has, however, also challenged our digital preservation capacity across staffing, resource, and technology.

IV. DIGITAL PRESERVATION IN ASC

To enable long-term preservation and continuing access to digital records, the University of Glasgow Library has been investing into digital preservation via the Digital Preservation Working Group (DPWG)⁵, a cross-University collaboration working to implement the University's digital policy and strategy. Established in 2015, the group oversees the delivery of digital preservation services, with representation from the University Library, IT Services and the Data Protection & Freedom of Information Office. The DPWG is also responsible for setting, maintaining and monitoring compliance with the University's Digital Preservation Policy [4].

For Archives & Special Collections, the drive to develop our digital preservation service is multi-faceted. There is, however, a strong connection to the communities we are part of. The University community – both as an organization and a community of researchers and students is key. However, we also sit within the wider heritage, cultural, and business communities. As a collecting institution, it is critical that we engage creatively and practically with the digital to ensure that we can continue to document, preserve and provide access to Scotland's economic, cultural, and creative heritage.

Within the business heritage community, ASC has worked for many years in partnership with the Business Archives Surveying Officer and the Ballast Trust to facilitate the survey, rescue, and preservation of business archives. Increasingly, as in the case of the NVA Archive, this work is centred around digital records. Reflecting on our work to preserve the NVA Archive, we will highlight the impact this Archive has had upon various aspects of our capacity to support the preservation of born-digital business records.

V. EVALUATING DIGITAL PRESERVATION CAPABILITY

The digital records of the NVA Archive provided an excellent case to evaluate the capability of our

⁴ <https://ballasttrust.org.uk/>

⁵ <https://bit.ly/3Cms6gN>

digital preservation system, a cloud-hosted instance of Archivematica 1.12.2 with 1TB cloud storage for processing and archival. In particular, we were keen to explore the system's performance with collections of high volume, file numbers and format diversity.

We used our digital archiving workflow⁶ (Figure 1) as the basis for evaluation. The workflow formalises and amalgamates existing ASC processes and practices around collections development, acquisition and appraisal, with digital preservation processing requirements.

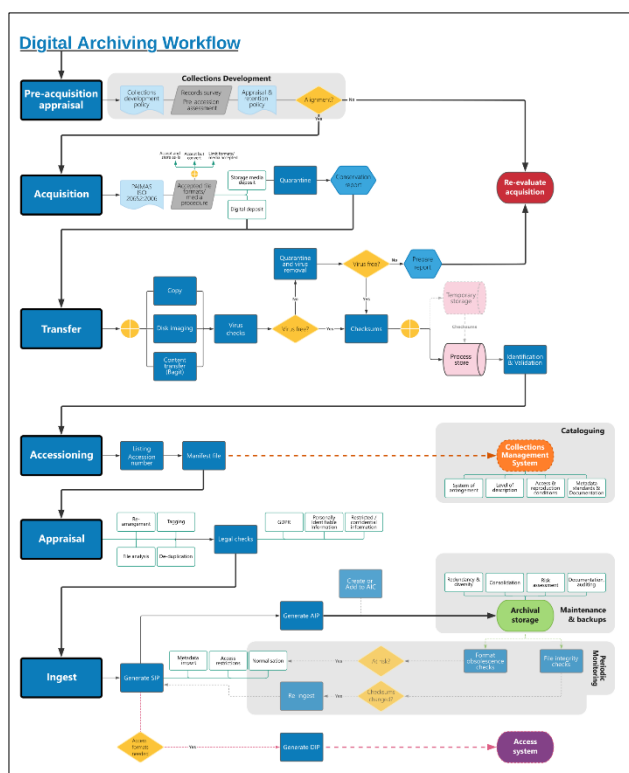


Figure 7 The ASC digital archiving workflow

PRE-ACQUISITION APPRAISAL

As a Glasgow-based limited company by guarantee, NVA fits the university's Collection Policy which states that "Archive Services primarily seeks to acquire records of Scottish business in the 19th, 20th and 21st centuries." [5]. The records survey conducted by the Ballast Trust, further highlighted a number of Intellectual Property and Data Protection issues which informed legal checks and the content of rights metadata.

EVALUATION METHODOLOGY

⁶ The latest version of the workflow is available on the Community Owned Workflows (COW) of COPTR: <https://bit.ly/3MsArUG>

We identified twenty test cases across six areas relating to preservation processing: cloud storage performance; system administration; transfer; appraisal; ingest; and archival storage. The latter five

Figure 8 Matrix of impact and likely for risk assessment

		Impact				
		Negligible	Minor	Moderate	Significant	Severe
Likelihood	Very likely	Low	Medium	High	High	High
	Likely	Low	Medium	Medium	High	High
	Possible	Low	Low	Medium	Medium	High
	Unlikely	Low	Low	Medium	Medium	Medium
	Very unlikely	Low	Low	Low	Medium	Medium

areas were meant to match the respective Archivematica tabs. Table 2 summarises the tests that were conducted per area.

Table 2 Summary of tests conducted per preservation processing area

Cloud storage performance	Test different methods of uploading files onto cloud storage Administer files on cloud storage
Transfer (Archivematica)	Complete automated and interactive transfers using default and custom processing configurations
Appraisal (Archivematica)	Use appraisal tools for content analysis and create SIPs from appraised/re-arranged content
Ingest (Archivematica)	Prepare and store AIPs for SIPs using preservation planning strategies for normalization
Archival storage (Archivematica)	Manage AIPs in the Archival storage tab
Administration (Archivematica)	Manage requests for AIP deletion and recovery Add, edit and delete Processing configurations View and manage failure reports View and manage Processing storage usage

Issues identified during testing were assigned risk levels, which were calculated using matrix of **Impact** and **Likelihood** of an identified Issue occurring and disrupting digital preservation processing (Figure 2).

The level of risk was assessed by the potential of an Issue to trigger a threat event, which in turn could give rise to one or more of the following:

Loss of Functionality

- The system does not perform one or more of its intended functions, either partly or entirely.
- The system exhibits unexpected behaviour(s) which affect or inhibit completion of operations.

Loss of Integrity

- The system and its data can no longer be trusted.
- The systems and its data are incomplete or incorrect.
- The security and confidentiality of the system and its data have been compromised (e.g. unauthorised access, wrong user permissions).

Loss of Availability

- The system can no longer be accessed.
- The system does not respond to valid queries and/or produces system fault errors.

The risk assessment was used to classify issues as either Low, Medium or High Risk (see Table 3).

Table 3 Risk assessment - levels of risk

Low	Medium	High
The Issue pertains to a non-critical or supplementary service/function/operation.	The Issue pertains to a core service/function/operation.	The Issue pertains to a critical and/or exigent service/function/operation.
The Issue does not prevent completion of operations.	The Issue may prevent completion of operations, but there are workarounds.	The Issue prevents completion of operations.
The Issue affects functions/operations that are easily recoverable or reproducible.	The Issue affects functions/operations, for which alternatives exist.	The Issue affects functions/operations that cannot be recovered or reproduced; and the related effects cannot be otherwise mitigated.
The Issue does not cause loss of integrity or availability.	The Issue does not cause loss of integrity but affects aspects of availability.	The Issue can cause loss of integrity and/or availability.

EVALUATION RESULTS

A total of 19 issues were documented across the six preservation processing areas. The risk assessment conducted showed that 45% of high-risk

issues related to cloud storage performance, followed by issues of Administration (22%). Similarly, 29% of all issues related to cloud storage performance; and 24% with appraisal (Table 4).

Table 4 Summary statistics of issues per risk level and preservation processing area.

	High	Medium	Low	% of total
Cloud storage	44%	0%	33%	29%
Transfer	11%	20%	0%	12%
Appraisal	11%	20%	67%	24%
Archival storage	11%	40%	0%	18%
Administration	22%	20%	0%	18%

With minor exceptions, the majority of the issues encountered during testing derived from the large volume and number of files in the NVA archive. We tested uploading files from the acquired hard drive onto cloud storage via both a web interface and a desktop client. For uploads via the web interface, server timeouts occurred in all tests where the file size was larger than 200MB. Uploads via the desktop client had a higher success rate, but sync speed was slow: in one test, it took 2 days to upload ca. 200GB of data. In other cases, file integrity checks failed due to file corruption during upload.

Archivematica performed more consistently across tests but struggled when transfer volume exceeded 1GB or when a transfer consisted of more than 1,000 files.

VI. LESSONS LEARNT

The experience of preserving the NVA Archive highlighted a number of areas where our digital preservation service needs improvement. Although the digital archiving workflow – and related ASC processes and policies – provided a good foundation for digital preservation processing, technical issues with cloud storage impeded operations.

Workarounds exist to mitigate the risks identified during testing. For instance, the number of files per transfer can be limited, and the collection can be preserved as multiple AIPs. However, problems persist with large files – e.g. a single video file in the NVA Archive was almost 3GB.

As we continue to develop our digital preservation service, and the more we engage with the digital preservation community, we expect to further our understanding of these issues and find robust solutions to preserve our at-risk collections.

ACKNOWLEDGMENT

As highlighted above, Archives & Special Collection's work to preserve the NVA Archive has been a partnership between the University and the Ballast Trust's Business Archives Surveying Officer. Chris Cassells, now of the National Library of Scotland, undertook a records survey of the NVA in 2018, and subsequently completed initial digital preservation actions for the collection across 2019 and 2020. His professional, and practical, input to the management of the Archive is greatly appreciated.

REFERENCES

- [1] E. Yan and C. Paterson, "World-changing Transformation: Collections at University of Glasgow Archives & Special Collections", 15th International Digital Curation Conference (IDCC20), 2020. <https://bit.ly/3MxZztj>
- [2] Fairweather, "DIANE WATTERS, St Peter's, Cardross: Birth, Death and Renewal", *Architectural Heritage*, vol. 27, no. 1, pp. 119-122, October 2017.
- [3] NVA, "Annual report 2015/2016", February 2017. <https://bit.ly/3pOiNkC>
- [4] University of Glasgow, "Digital Preservation Policy", v3.8, 2020. <https://bit.ly/3CDpMIP>
- [5] University of Glasgow, "Collections Development Policy", v1.0, February 2017. https://www.gla.ac.uk/media/Media_591723_smxx.pdf

ARCHIVEMATICA-EPRINTS INTEGRATION

Developing digital preservation capacity for open repositories

Tomasz Neugebauer

Concordia University
Canada
tomasz.neugebauer@concordia.ca

Sarah Lake

Concordia University
Canada
sarah.lake@concordia.ca

Abstract – Following three years of software development, requirements refinement, and testing, we released the integration of EPrints and Archivematica as a plugin to EPrints in 2021. This paper will explain how the integration evolved, how we implemented it in parallel to a new Archivematica instance at Concordia University, and how we are currently using it to preserve the contents of our institutional research repository. We will conclude with a discussion of some possible future enhancements envisioned for the integration.

Keywords – Digital preservation systems, digital repositories, integration, EPrints, Archivematica

Conference Topics – Innovation, Community

I. BACKGROUND

When Concordia University Library started planning to implement a digital preservation program in 2018, one of our first goals was to improve the digital preservation workflows for our institutional research repository (IR), Spectrum. Spectrum is built using EPrints, a free and open-source software package for creating open access repositories that are compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).¹ Like many repository systems, EPrints' native digital preservation functionality is limited [1]. To ensure the long-term preservation of the content deposited to the IR, we needed to implement a digital

preservation system that we could integrate with our EPrints repository.

After benchmarking a number of digital preservation solutions, we retained Archivematica as our preferred option. Archivematica is a project by Artefactual Systems that integrates a suite of open-source software tools allowing users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model.² Our decision was based on Archivematica's low relative cost, its high flexibility and scalability, and its active user community that has helped develop and sustain its open-source model since 2009.

With both Archivematica and EPrints being open-source projects, we saw an opportunity to collaborate with EPrints Services, Artefactual Systems, and other members of the open-source software community to create an EPrints-Archivematica export plugin. This integration would bridge a gap between two widely-adopted open-source systems and provide valuable digital preservation functionality for EPrints repositories. It would allow us and other open repository administrators to continually ensure that files entrusted to us are not lost or corrupted and sufficient information about the digital objects is collected to enable future preservation and access.

The integration plan was first presented in 2018 at an Archivematica Camp and the Open Repositories conference [2]. Development work to

¹ <https://www.eprints.org/>.

² <https://www.archivematica.org/en/>.

create an Eprints-Archivematica export plugin started shortly thereafter, with Concordia University leading the development. We released the integration as a plugin to EPrints in 2021, and we are currently using it in production to export digital objects and metadata from Spectrum to a dedicated Archivematica instance.

This paper will explain how the current version of the plugin is being used with our Archivematica pipeline and what we learned during the development. We will conclude with a discussion of some possible future enhancements envisioned for the integration.

II. PLUGIN DEVELOPMENT AND DESIGN

The GitHub repository³ has served as the common platform for refining the specifications, and iterations of the software development. Virtual meetings were held as the plugin was developed, with participation from the community, such as the University of Strathclyde [3]. A first official release of this work as a Bazaar Plugin was completed in December 2021.

One of the fundamental questions for the export is the structure of the transfer object. The current version exports out metadata and digital objects according to the standard Archivematica transfer structure with existing checksums⁴. It uses two folders at the second level: one for metadata files and accompanying checksum file, and an objects folder with the deposited documents and derivatives.

The plugin creates and tracks transfer records for each deposit. Instead of relying on a BagIt utility for generating checksums, the plugin itself generates a checksum.md5 file that includes all of the files in the objects folder, and compares these with what is already stored inside EPrints, looking to flag any checksum missing or mismatch problems during processing.

Instead of using an eprintID and date at the top level (the initial idea from 2018), the current version of the plugin uses an “Archivematica ID” assigned by EPrints to each transfer record as the top-level folder name. This was a practical decision in that

Archivematica needs to send this ID back to EPrints on completion, along with the UUID, and the callback functionality can easily retrieve it from this folder name, which is also the AIP (Archival Information Package) name inside Archivematica. In the latest release, we added the option to include a prefix for the repository name, for example “spectrum-999” instead of just the id, which is a useful configuration option for those institutions that have multiple EPrints instances exporting to the same Archivematica instance.

There are currently three command line scripts that ship with the plugin for creating, flagging, and processing. Each of these scripts can take arguments to limit its functionality to a specific deposit or a limited number of deposits. This was especially useful during development, testing, and batch processing of the existing backlog of deposits to process. After the backlog is completed, a periodic run of create and process transfers will be placed in the crontab, to export out only the newly published deposits, or ones whose files or metadata fields have changed in ways that match our specification for sufficient change to re-export. A deposit’s transfer can be flagged for export by a trigger configured to watch specific metadata fields and/or changes to the uploaded files.

When Archivematica successfully processes the transfer and stores the AIP in archival storage, it sends the UUID of this transfer back to EPrints using a Service Callback. The plugin stores the UUID of the AIP in archival storage as a part of the log for each transfer record, and changes the state to “archived”.

The metadata.json file generated by the plugin contains some basic descriptive metadata, including item title and date, and importantly, the eprintid/URL of the originating item. Unlike the EP3.xml file that is also included, when this is ingested by Archivematica, it is included in the METS file and indexed for searching, allowing for retrieval of AIP by eprintID, for example.

Typically, published items in repositories are either not modified post-publication, or when modified, re-published as new item versions with their own eprint IDs. However, some workflows (e.g.,

³ <https://github.com/eprintsug/EPrintsArchivematica>

⁴ <https://www.archivematica.org/en/docs/archivematica-1.13/user-manual/transfer/transfer/#transfer-checksums>

working papers, corrections to existing files) require that published documents change without creating a new eprint version. If the files in the item and/or metadata defined in the plugin configuration as constituting a sufficient change to merit a re-export, are modified post-export, the plugin would flag that item's archivematica record as in need of processing. It would be exported with the same top folder name again and ingested by Archivematica as a new AIP. The result would be an eprint with two separate AIPs in storage, and both the UUIDs would be associated with that item's archivematica record in EPrints.

III. PROCESSING IN ARCHIVEMATICA

In parallel to the plugin development, we deployed a self-hosted instance of Archivematica with an annual support contract with Artefactual Systems. Our instance currently has two pipelines, one dedicated to Spectrum, our EPrints repository, and the other to the Library's Special Collections, which also fall within the scope of the Library's digital preservation program. The following section outlines how we customized our Spectrum Archivematica pipeline to fully automate processing.

Archivematica integrates a suite of open-source tools to perform a host of preservation actions on the transferred items, including file format characterization, validation and normalization, and generating or extracting a large volume of technical preservation metadata, which is stored in a METS file. Archivematica then generates an AIP containing the transferred items and associated derivatives and metadata and places it in archival storage.

We use Automation Tools,⁵ a set of Python scripts designed to automate the processing of transfers in Archivematica, to automatically check a watched folder for new transfers every three minutes. Artefactual Systems helped us implement a script to clear completed transfers from the dashboard automatically, which allowed us to process batches of several hundred transfers at a time without impacting Archivematica's performance. This modification was necessary for scalability because the accumulation of hundreds of transfers typically makes the dashboard unresponsive and prevents us

from seeing the transfers in progress or troubleshooting failed transfers.

We implemented an automated processing configuration so that when a transfer is picked up by Archivematica, it is fully processed, packaged, and stored without any intervention needed on our part unless one of the microservices fails.

The only step of this workflow that isn't fully automated is deleting the completed transfers out of the transfer source folder, which we are currently doing manually through SFTP after each batch. Artefactual is currently developing a product called Enduro⁶ which is intended to eventually replace Automation Tools, and could potentially help us automate this step further down the line.

For our archival storage infrastructure, Concordia University recently subscribed to the Ontario Library Research Cloud (OLRC)⁷ a private cloud storage network for Canadian universities. The OLRC uses a modified version of Duracloud⁸ that can be configured as a storage location in Archivematica's Storage Service application. With Duracloud, three copies of each AIP are replicated on servers with periodic fixity-checking across a private network of geographically-dispersed university-owned and operated data centers. Should one copy of an AIP become unreadable, it is automatically replaced by a new one created from the two others.

Name	UUID	Size	Created	Status	Encrypted	Actions
2779	7367d957-4840-44d3-9f58-61d4ef081d54	3.5 MB	2022-02-02 13:11	Stored	False	View

Figure 1 EPrintsArchivematica export in Archival Storage through the Archivematica Dashboard. Showing results of search by "AIP Name", which is also the Archivematica object ID for this item inside EPrints.

Search	Filter	Search
986367	Transfer metadata	Search archival storage

Figure 2 It is also possible to search by descriptive metadata submitted in the metadata.json file in the transfer, such as the EPrintID. For this, the "Transfer metadata" index is used on the "Browse Archival storage" interface in Archivematica Dashboard.

⁵ <https://github.com/artefactual/automation-tools>.

⁶ <https://github.com/artefactual-labs/enduro>.

⁷ <https://cloud.scholarsportal.info/>.

⁸ <https://www.lyrasis.org/DCSP/Pages/DuraCloud.aspx>.

Once an eprint AIP has been created and stored, a post-store callback in Archivematica updates the item record in EPrints to include the Archivematica UUID of the eprint and a timestamp. This serves as a confirmation in both systems that the item has been preserved—a feature that we originally thought would be nice to have and that we ultimately decided was essential. A caveat to this is that the callback cannot be limited to the pipeline that needs it, which isn't ideal for institutions using one storage service for two pipelines like we are. We have proposed this as a feature on the Archivematica GitHub.⁹ Not being able to limit the callback to a specific pipeline means that on the EPrints side, we have to ignore irrelevant callbacks from the storage service when it places Special Collections AIPs in archival storage. It is relatively simple to enable or disable a specific callback using the Storage Service, so we can disable it during times when the other (Special Collections) pipeline is actively processing.

AMID	Dataset ID	DataObj ID	Needs update	UUID
2779	eprint	986367	No	7367d957-4840-44d3-9f58-61d4ef081d54

Timestamp	Action	Comment	Result
23 January 2022 01:00:33 UTC	Transfer Created	created via trigger	Success
2 February 2022 18:06:48 UTC	Transfer Processed	processed	Success
2 February 2022 18:11:14 UTC	Transfer Archived	UUID [7367d957-4840-44d3-9f58-61d4ef081d54]	Success

Figure 3 Archivematica Records Management Screen in EPrints viewing the details of a transfer. The UUID of the AIP in Archivematica is sent to EPrints using a Callback.

IV. ISSUES AND LIMITATIONS

OpenDOAR¹⁰ reports more than 630 instances of EPrints repositories worldwide, some of which might very well be running an out-of-date EPrints platform and in need of digital preservation. More testing of the plugin on out-of-date EPrints repository versions might prove valuable in the future. The plugin should work well for EPrints instances running any version of Eprints 3.3 or 3.4. It was developed and tested successfully on EPrints version 3.3.12 and 3.4.3. The authors are not aware of any tests of the plugin on EPrints repositories running earlier versions than 3.3.12. On the Archivematica side, the integration was developed and tested on Archivematica 1.12 / Storage Controller 0.16, but we have since upgraded to Archivematica 1.13 / Storage Controller 0.18.

As we started to use the plugin in production, we came across unanticipated issues and limitations

that prompted us to reflect on our preservation strategy as a whole and led us to make changes to the code and our workflow. We have been using this plugin to preserve more than a decade's worth of deposits from our institutional repository, and one of the discoveries we made in this process is that thousands of PDFs imported into the repository did not have a checksum in the EPrints database. The issue was so common that we decided to develop functionality in the plugin to add the checksums to EPrints during export, throwing only a warning rather than the usual "checksum mismatch" error in these cases.

Another lesson-learned about EPrints during the development of the plugin is that the file names stored in the the EPrints File Object¹¹ can in some cases differ from the corresponding file names on disk. The use of regular expressions was required in the plugin to replace some characters (quote and double-quote) in the value that is returned by the internal EPrints object call to get "filename". The file's URL is served over HTTPS with the quotes and double-quotes in the name, but stored on disk with a file name that replaces those characters with their ASCII addresses: =0027 and =0022.

We also found that any metadata-only eprints, i.e. items that did not contain a document or upload of any kind, caused the transfer to fail in Archivematica. This issue prompted a discussion about whether or not these items should be preserved in Archivematica at all, since they didn't contain any deposited digital objects. In the end, we added an option in the plugin to export the metadata-only transfers to a different folder location so that they could get skipped over by Automation Tools. The metadata would still get exported out to be stored along with any of the logs from the preservation batch jobs.

One limitation of our workflow is that we have found the process of resolving certain errors to be unnecessarily labor-intensive. For instance, if a normalization job fails due to an error with the tool, command, or file, our automatic processing configuration will approve the normalization anyway. In these cases, we have had to re-ingest the transfer with a manual normalization workflow, which involves adding the preservation derivative to the transfer folder through SFTP, generating a

⁹ <https://github.com/archivematica/Issues/issues/1325>.

¹⁰ <https://v2.sherpa.ac.uk/opendoar/>.

¹¹ https://wiki.eprints.org/w/File_Object

checksum, and adding it to the checksum file; otherwise, the transfer will fail due to mismatched checksums. We have only encountered this issue in a couple of transfers so far so it has not been a significant issue, but it would be helpful to have a more efficient way of handling this.

Most of the preservation issues we encountered as we processed our repository's backlog involved transfers containing unusual deposit formats. For example, in instances where a student had created software as their thesis project, the application bundle in their deposit often contained files that would cause the transfer to fail in Archivematica. In one case Archivematica was unable to assign UUIDs to symbolic links in a MacOS application bundle, which made the transfer fail at the "Generate METS" step. Our solution was to simply not extract the problematic packages, which isn't perfect, but it allows us to perform bit-level preservation as a first pass and leaves us the possibility to revisit content-level preservation in the future.

V. CONCLUSION

In spite of the occasional issues and limitations, we are very satisfied with the results of this integration project. As of the publication of this paper, we have successfully processed the entire backlog of roughly 18,000 live eprints from our repository, resulting in approximately 500,000 indexed files in archival storage. We are in the process of determining what our preservation workflow will be going forward. This will most likely involve running create and process transfers weekly and establishing a procedure for the removal of items from archival storage. This will be done in coordination with the Thesis Office who would communicate with the Digital Preservation Librarian if, for example, a thesis is withdrawn.

A future enhancement to the plugin is to include in the AIP the processing log generated from the results of each command on each eprint that was exported using the plugin.¹² We determined that since this log is a record of the preservation actions that were performed upon the objects, it should be encoded as preservation metadata using PREMIS. PREMIS is a de facto digital preservation metadata standard implemented in Archivematica and supported to some degree in many other systems such as: Islandora, RODA, Preservica, DSpace,

BitCurator, AtoM, HathiTrust [4]. Archivematica can parse an imported `premis.xml` file into the main METS file of the AIP, so that all of the preservation metadata about the objects is in one place and in a machine-readable and interoperable format.

We anticipate that feedback from the user community will inspire new enhancements that will allow this integration to flourish over time. For example, the community has already identified the need for a `delete_transfers` script that would make it easier to remove transfer objects from EPrints that are no longer needed. We see the release of this plugin as an important step towards bridging gaps in open-source digital preservation workflows for repositories, and we hope that it will empower repository managers to implement digital preservation practices at their institutions.

ACKNOWLEDGMENT

The authors would like to thank EPrints Services, for their development work and collaboration that made this project possible, and Concordia University Library for sponsoring the work. We would also like to thank the members of the EPrints open-source community for their contributions to and feedback on the plugin, and Artefactual Systems for the support and guidance they provided for the configuration of our Archivematica pipeline and Automation Tools. We would also like to thank Tessa Walsh for her thoughtful suggestions and contributions in the development of this plugin and Concordia University's digital preservation planning.

REFERENCES

- [1] T. Neugebauer, P. Lasou, A. Kosavic, and T. Walsh, "Digital Preservation Functionality in Canadian Repositories," Canadian Association of Research Libraries. 2019. Available: https://www.carl-abrc.ca/wp-content/uploads/2019/12/orwg_report2_preservation_repos_en.pdf.
- [2] T. Neugebauer, J. Bradley, and J. Simpson, "Digital Preservation through EPrints-Archivematica Integration," *International Conference on Open Repositories*, Bozeman, USA, 2018. Available: <https://spectrum.library.concordia.ca/id/eprint/983933/>.
- [3] G. Macgregor and T. Neugebauer, "Preserving digital content through improved EPrints repository integration with Archivematica," *UK Archivematica User Group*, 2020. Available: <https://strathprints.strath.ac.uk/73978/>.

¹² <https://github.com/eprintsug/EPrintsArchivematica/issues/37>.

- [4] M. Jordan and E. McLellan, "PREMIS in Open-Source Software: Islandora and Archivematica," in *Digital Preservation Metadata for Practitioners*. Cham., Switzerland: Springer, 2016, ch. 16, pp. 227-239. [Online]. Available: https://doi.org/10.1007/978-3-319-43763-7_16.

MAPPING THE LANDSCAPE OF DIGITAL PRESERVATION NETWORKS

The nestor Digital Preservation Community survey

Michelle Lindlar

TIB – Leibniz Information
Centre for Science and
Technology
Germany
michelle.lindlar@tib.eu
[0000-0003-3709-5608](tel:0000-0003-3709-5608)

Svenia Pohlkamp

DNB
Germany
s.pohlkamp@dnb.de

Monika Zarnitz

ZBW Leibniz Information
Centre for Economics
Germany
m.zarnitz@zbw.eu
[0000-0001-9229-1877](tel:0000-0001-9229-1877)

Thomas Bähr

TIB – Leibniz Information
Centre for Science and
Technology
Germany
thomas.baehr@tib.eu
[0000-0002-9337-7127](tel:0000-0002-9337-7127)

Stefan Strathmann

Göttingen State and
University Library
Germany
strathmann@sub.uni-goettingen.de
[0000-0001-5328-1174](tel:0000-0001-5328-1174)

Abstract – "Digital preservation is people" and "Digital preservation cannot be done alone" are often heard statements within our domain. Nevertheless, no exhaustive survey of digital preservation communities had been done. The nestor Digital Preservation Community survey closed this gap and the nestor working group "Community Survey" is currently working on a publication of the survey results. This short paper presents the survey design and process, gives an insight into some of the findings by using the survey results to answer questions about the landscape of digital preservation communities and gives a brief outlook on further work.

Keywords – digital preservation networks, digital preservation communities, survey

Conference Topics – Community; Exchange.

I. INTRODUCTION

"Digital Preservation is People" has become one of the most fervent slogans of our domain. It has been used to highlight the contextualization of our work within an institutional framework [1], the skills needed by people to do the job [2] and the relationship between people and technology [3].

Another universally accepted statement about our domain is that digital preservation is an enormous task - one that is too big to be tackled alone [4]. Therefore, it comes as no surprise that digital preservation networks have been around for almost as long as national programs addressing the risks of digital information loss. Within the first decade of the new millennium, networks like the Open Preservation Foundation¹ and the Digital Preservation Coalition² were built upon the momentum of EU funded projects and continue to grow and flourish until today. Other networks, like the US-based Digital Preservation Network (DPN) have ceased to exist [5].

It would be naive to assume that the "usual suspects" of networks contain all institutions worldwide who deal with digital preservation. Furthermore, we have to acknowledge that our knowledge of networks and communities is naturally limited by the geographical and domain-based framework that we ourselves interact in. Looking at iPRES, figures of authors [6] or

¹ OPF - <https://openpreservation.org/>

² dpc - <https://www.dpconline.org/>

attendees by country [7] show that digital preservation is tackled by institutions across the globe. It is therefore safe to assume that the landscape of digital preservation networks is larger than we know.

While surveys about digital preservation topics are not uncommon, they mainly target individuals and institutions, asking for input on resources or requirements. Examples for this are the OPF Community Survey [8] or the NDSA Storage Survey [9]. There had not been an extensive survey aimed at mapping the landscape of digital preservation networks³. Within nestor⁴, the German Competence Network of Digital Preservation, the working group "Community Survey" was founded and tasked with closing this gap. The survey presented here was drafted and conducted in 2019 - 2020, the results were analyzed in 2020 - 2021 and results are currently being finalized to be published in the second quarter of 2022⁵. Section II of this paper presents further background information about the design of the survey and the way in which it was conducted. Section III briefly showcases some of the survey results and how they can be used to answer questions about the digital preservation community landscape. We conclude this paper with an outlook to further work in Section IV.

II. SURVEY DESIGN AND PROCESS

The working group was kicked off in February 2019 and consists of 5 members from nestor partner institutions. Having had its first meeting just shortly before the global pandemic came into light, members' available resources frequently changed and subsequently the time plan had to be shifted and adjusted several times. Due to this, the overall process took over 3 years. The entire survey project can be broken down in 4 phases, which are as follows:

- *Phase 1: Definition and Preparation* February 2019 - May 2019
- *Phase 2: Survey* September 2019 - May 2020

- *Phase 3: Analysis* May 2020 - July 2021
- *Phase 4: Preparing Publications* August 2021 - April 2022

The following subsections briefly describe key issues that needed to be addressed organizationally as part of the project. These form necessary background information for the outcome presentation in Section III.

A. *Definition of "Community" and of the survey's goal*

The first step towards the preparation of the survey was finding a shared definition of "Community". The goal was to find a framework in which no network would define itself as "too small" or "too broad", thus feeling it does not fit into the boundaries of the survey. The intention of the definition was therefore to include rather than to limit. After much discussion, the working group reached the following shared definition for "digital preservation community" [10]:

- an open community of persons and/or institutions who engages with the subject of digital preservation as its sole or one of several subjects
- a community whose members are committed to digital preservation in a manner that goes beyond pure self-interest, in particular it goes beyond the sole or central purpose of supplying a product or providing a commercial service
- a platform for discussing digital preservation practice and research, including the development of tools
- a community can be
 - local, regional, national or international
 - large or small
 - product-related or not product-related

In parallel, we needed to formulate what we wanted to achieve by conducting this survey. Through discussion within the working group it

³ Communities" and "networks" are used interchangeably through- out this paper

⁴

<https://www.langzeitarchivierung.de/Webs/nestor/EN/Home/>

⁵ At the time of writing, the results have not been published yet. However, this is scheduled to happen before iPRES2022. In the case of acceptance, all references in this paper will be changed to the published versions.

became clear that we wanted to create a "map" of digital preservation communities - a map in a geographic as well as a subject-based sense of the word. The survey results should include a registry, which interested practitioners and researchers as well as other networks could use to identify networks that cover issues they are interested in. Such a registry could also allow for identification of targets for cross-community collaboration, hence creating synergies and making best use of our limited resources in digital preservation. Based on this, the working group formulated two types of output for the survey results: a registry of community profiles on the nestor website as well as a report on the survey data set, which summarizes anonymized results.

B. Questionnaire

The questionnaire⁶ was designed as an online questionnaire using the "Mailingwork" survey tool⁷. It consisted of 40 questions which were divided into the following categories:

- *Formal aspects* (10 questions)

Rationale: Understanding of where community is located, what it defines as success factors, how long it has been operating and how it can be reached.

- *Governance structure & financing* (5 questions)

Rationale: Understanding of community's legal status, financing sources and internal governance bodies (e.g., Board).

- *Organizational structure* (12 questions)

Rationale: Understanding of membership types, membership numbers and distributions across organization types; Understanding of geographic and subject scope as well as key services; Understanding of personnel resources (FTEs) and collaborations with other communities.

- *Communication* (10 questions)

Rationale: Understanding of outreach activities in width and depth; Understanding of collaborative work spaces used.

- *Events* (3 questions)

Rationale: Understanding of events organized for members / other target groups

C. Distribution of the survey

In a first step, the working group collected a list of known digital preservation communities as well as of mailing lists via which the survey announcements were circulated. Contacts from known communities were contacted directly and asked to take part in the survey, but also asked to suggest networks that they thought should be included in this survey. These named candidates were then also approached directly. Two follow-up emails were written if no response had been received. In addition to the direct contacts and mailing list distribution, the working group members used their social media channels and international practitioner networks asking to amplify the project. The survey ran for 8 months. While this may seem like a long time, it seemed necessary to receive the best amount of responses during global lock-downs.

III. RESULTS

Overall we received 73 responses. After deduplication and data cleansing of entries that did not match the given community definition, 54 valid responses formed the basis for all result analysis.

The data set presents a unique information resource about digital preservation communities. In this section we briefly describe the structure of the community profiles and showcase how the survey result presented in the forthcoming nestor publication can be used to answer questions about the current digital preservation community landscape.

A. Community Profiles

As described above, one of the targeted outcomes of the community survey is a registry of digital preservation communities. For this, a community profiles template was created, which includes 32 criteria that can be extracted from the survey questions. These criteria are grouped into the sections "General characteristics", "Mission and scope", "Governance structure and financ-

⁶ The full questionnaire will be made available in March 2022 as part of the nestor materials publication [10]

⁷ <http://mailingwork.de/software/features>

ing", "Organizational structure", "Cooperation", "Modes of communication - scope", "Events organized by the network". For those survey respondents who had indicated that they would be willing to include their data in a publicly available registry, profile sheets were generated and sent to the named contact asking for corrections and approval of the profile as well as for a logo to be included in the registry. At the time of writing this paper, 33 networks have agreed to be included and have approved their profile.⁸

B. *Aggregated results for the nestor material publication*

While the published community profiles give an in-depth insight into single communities, there is not a profile for every respondent to the survey. In contrast, the nestor material publication [10] includes the anonymized results of all 54 valid responses, making it an excellent resource for quantitative analysis. Since a discussion of the entire data set is not possible within the limits of a paper, we will showcase the results using 5 sample questions that can be answered using the data presented in the report.

1. *Where are digital preservation networks located?*

Despite the working group's efforts to spread the survey as wide as possible, the majority of the responding communities (80%) are located in Europe or North America. Table I shows the distribution of all respondents by geographic region. However, it needs to be noted that 7.4% (4 cases) listed "International" or "World" as their location. Other respondents stressed that their membership is indeed international, their offices, however, are all located in Europe or North America. Table I can therefore only be used to make a statement about the main location of the community, not of its geographic reach.

Table I
Surveyed communities clustered by geographic regions

Region of the world	% of answers
Asia	3.7%
Australia	5.6%
Europe	51.9%
North America	29.6%
World	7.4%

⁸ At the time of writing this paper, the profiles are not yet published via the nestor website. They are scheduled to go

2. *Is there a correlation between a community's founding year and the size of its membership?*

One might think: "The longer the network has been around, the more members it has". But is that really true? While "founding year" allows for a comparable answer, membership number is not quite as straightforward to interpret. This is largely due to different types of memberships that exist, such as personal or institution based categories. One respondent may have 20 organizations as members, whereas another community counts 1,000 individuals as members. Nevertheless, a quantitative analysis of founding year against membership numbers can offer some insights. Figure 1 shows the distribution of founding year and number of members for all respondents who supplied data for both questions (n=47). Especially the older networks, which were founded between the years 1945 - 1995 may come as a surprise as digital preservation was not really a topic back then. These responses can be contextualized by cross-checking the results in Figure 1 against answers to the question whether digital preservation is one of several topics of the community (10 cases). Since digital preservation research only dates back to the 1990s, it is safe to assume that communities found prior to that cover digital preservation as one of several topics. This, in return, needs to be taken into consideration when looking at the membership numbers of these communities - broader services and fields of interest, such as several library-relevant topics in addition to digital preservation, may have an impact on the overall number of members.

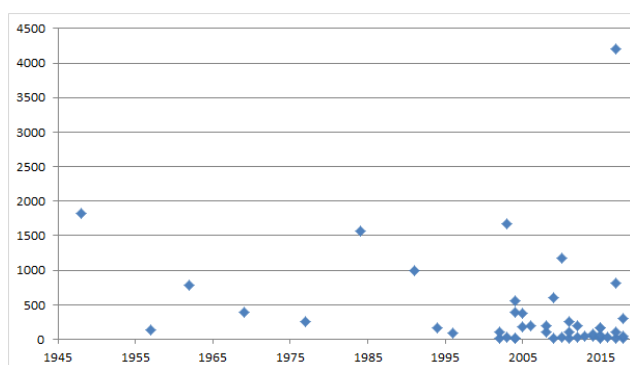


Figure 1 Number of members by networks' founding year

online later in March. In case of a successful, review the link will be added to the publication.

The clustering of communities found between 2000 and today underlines the previous statement that digital preservation research only dates back to the 1990s. It is thus no surprise, that the majority of responding communities were found after 1995. Three interesting observations can be made by looking at the numbers of communities found between 2000 and today: there is no decrease in intensity, suggesting that new communities have been created at a somewhat consistent rate over the course of the past 20 years; while there appears to be a peak in member-rich communities around 2005, the overall numbers suggest that comparatively young communities can still reach high numbers of membership; new communities are still found at a rather consistent rate, in other words: there appears to be no "over-saturation of digital preservation communities".

3. Are digital preservation communities just for archives?

When asking "who is involved in digital preservation" a first answer is often "archives", followed by "libraries". But is that all? And who is actively involved in digital preservation communities? To understand this, the survey asked about members' organization types. Out of the 54 respondents 46 supplied answers to this question. Some of those who didn't provide answers stressed that their community has no official membership model, making it hard to estimate organization types.

Table II shows the different organization types that were mentioned including how often they were mentioned and what percentage of communities have members of this organization type. It may come as a surprise that "Universities" ranks highest in the list of mentioned cases, however, we need to keep in mind that often a university library, archive, research institute or a computing center is the direct benefactor of the membership, but the university itself signed the membership agreement. Overall, the listing shows that the need for digital preservation communities exists in a broad organization base - over half of the communities have universities, libraries, archives and research institutions as their

members; over 40% additionally have museums and enterprises as members. The high number of "Others" is surprising. While one respondent classified their entire membership base as "others", 13 respondents made use of that category in addition to the named categories. The report [10] gives further information on the breakdown of different organization types across communities' membership basis. It is interesting to note that amongst the survey respondents were also "specialized networks", where one organization type makes up for 100% of the members. Such "specialized" digital preservation communities exist for libraries (n=2), archives (n=3), universities (n=1), enterprises (n=1) and government (n=1).⁹

Table II
Members' organization types across surveyed communities

Member type	Cases	% of communities with member type
Universities	39	72.2%
Archives	37	68.5%
Libraries	34	63.0%
Research Institutions	28	51.9%
Enterprises	23	42.6%
Museums	23	42.4%
Government agencies	4	7.4%
Broadcasting	2	3.7%
Individuals	2	3.7%
Others (unnamed)	15	27.8%

Table II also highlights where there is room for growth within communities. Why, for example, are broadcasting companies only mentioned twice? What can we do to get those organization types engaged in more networks?

4. What services do digital preservation communities offer?

After the previous subsections gave insight into who the communities are, another vital question is what services they provide to their respective members. Survey participants were asked to indicate which service they offer for their members and, if applicable, non-members. 10 possible services incl. the option "Other" were given. 51 participants provided answers regarding

⁹ Not included in this listing is 1 case of 100% "Others" and 1 case of 100% "Individuals", as those do not allow for organization matching.

services. The answer options as well as the number of times they were chosen can be seen in Table III.

Table III
Members' organization types across surveyed communities

Services	Cases	% of communities offering service
Knowledge transfer	44	81.5%
Community building	37	68.5%
Technology / Tool development	21	38%
Technology Watch	13	24.1%
Standardization	13	24.1%
Digital-preservation-as-a-service	12	22.2%
Lobbying	12	22.2%
Offering technical solutions / digital preservation software	10	18.5%
Certification	6	11.1%
Other	8	14.8%

The services can be grouped together in three "blocks". The two top ranking services are two classic "community" items - knowledge transfer and community building. These are offered by 68 - 81.5 % of the responding communities. Several communities who responded offered only one of these two services and no others.

A second block of services deals with technology in form of facilitating (joined) open source tool development, technology watch services or offering technical solutions such as (end-to-end) digital preservation software or even full digital-preservation-as-a-service. Between 18 - 25 % of the communities offer one or several of these technology-themed services. While this seems low in comparison to knowledge transfer and community building, it still stresses the high importance of community support around technology in digital preservation.

While the first block connects members to each other and the second block connects members to technology, the third block can be described as outward facing services for the members. These are: lobbying, standardization and certification as well as fundraising, which was the only entry made in the additional free-text field for "other". As can be seen in Table III, these types of services are less frequently offered than technology services, ranging between 11 -25%.

5. How are digital preservation communities financed?

The last question we would like to showcase in this paper is how the surveyed communities are financed. 53 communities provided information for this. We did not ask the participants to weigh their financing sources, i.e. indicated how many percent of overall funding a specific category makes up for, but to just list those that are applicable. Approximately 40% (n=22) of the responding communities are (partially) financed through membership fees, 39% (n=21 for each of the three categories) listed revenues from services, sponsoring or in-kind contributions as funding sources. The categories sponsoring (n=14), which was mentioned by 26% of the participating communities, as well as government funding, which was mentioned by 14.8% (n=8), received fewer mentions.

It is therefore safe to say that digital preservation communities are largely funded by the community members themselves – either directly in form of membership fees or fees for services or indirectly in form of in-kind contributions.

IV. OUTLOOK AND FURTHER WORK

As demonstrated in the Results section, the data gathered in the survey is a valuable resource. It can be used to answer questions about digital preservation communities such as how they are financed and what services they offer to their members. Since such structured information about digital preservation communities was previously not available, the nestor community survey has closed a gap in digital preservation discourse. Nestor has included the Digital Community Survey as a line-item in its product matrix and the working group is planning to re-run the survey in regular intervals, currently looking at every 3 years. Valuable lessons-learned in this first run of the survey will be reflected upon and fed into the next version of the survey.

A key issue the working group would like to improve in the next run of the survey is the time plan. Reflecting upon the time needed for the four phases as described in Section II, the time for "Definition and Preparation" (4 months) as well as for running the survey itself (8 months) seems reasonable. We may consider to keep the survey open for a short time frame – however, the longer period allowed us to individually chase known communities and ask them to participate. Already

having a contact list of networks to build upon, we may consider shortening the time the survey is open slightly while still being considerate of the high work load that many community managers face and the additional burden that a survey might pose. Without a doubt, the time needed for phases 3 “Analysis” (14 months) and 4 “Preparing Publications” (9 months) is too long and needs to be improved upon. We are hoping that the workflows we have established in this first run of the survey, in particular the templates for the community profiles and automated mechanisms to populate them as well as overall decisions regarding presentation forms, will allow for significantly shorter phases 3 and 4 in the future.

In addition to a stricter time schedule, the working group is in particular hoping to reach more communities, especially in currently non- or under-represented regions (see Table I). Within the working group itself, a higher awareness towards international communities and networks exists and communities identified throughout the year are kept track of to include in future surveys. Presentations and publications, such as this paper, help the working group in spreading the word about the value of the survey outcome and we are hoping that this will encourage more communities to participate in the future. We are currently identifying target channels to publish the results through to heighten the visibility of the community profiles and the report.

Lastly, the in-depth analysis during phase 3 has provided some feedback which will be fed into the next questionnaire to make it more concise and universally understandable. A number of questions have offered re-occurring answers in “others” free-text fields. A particularly high number (n=6) of named “other” categories could be found in the question regarding the legal status. All named “other” categories will be considered for inclusion as fixed categories in the next instance of the questionnaire. The working group will ensure that any changes made to the survey structure will still allow for comparability of the results across different survey instances over the years. Additionally, the working group will be happy to receive any feedback and comments on the survey and is hoping for the wider digital preservation community to shape this survey into a useful tool to keep on mapping our global landscape.

ACKNOWLEDGMENT

The authors would like to thank all networks that took time out of their busy schedules to participate in the survey and/or helped spread the word about the project.

REFERENCES

- [1] T. Owens, *The Theory and Craft of Digital Preservation*. Johns Hopkins University Press, 2018.
- [2] G. Hurley, Digital preservation is people: Thinking about digital skills for archivists, Personal Online Blog, 2018. [Online]. Available: <https://www.granthurley.ca/blog/digital-preservation-is-people-thinking-about-digital-skills-for-archivists/>.
- [3] J. Ranger, Digital preservation is people, AVP Online Blog, 2014. [Online]. Available: <https://blog.weareavp.com/digital-preservation-is-people>.
- [4] Libraries are facing big challenges in digital preservation: We cannot do it alone, Online IFLA Blog, 2017. [Online]. Available: <https://blogs.ifla.org/lpa/2017/11/30/libraries-are-facing-big-challenges-in-digital-preservation-we-cannot-do-it-alone/>.
- [5] DPN Nodes, The digital preservation network (dpn) to cease operations, Message published in LYRASIS online blog, 2018. [Online]. Available: <https://duraspace.org/the-digital-preservation-network-dpn-to-cease-operations/>.
- [6] M. Lindlar, Ipres authors 2004 - 2017 by country (infographic). Zenodo, 2018. [Online]. Available: <https://doi.org/10.5281/zenodo.1462500>.
- [7] iPRES2019, “Conference attendees,” in *Proceedings of the 16th International Conference on Digital Preservation*, 2019, p.20. [Online]. Available: <https://ipres2019.org/static/proceedings/iPRES2019.pdf>.
- [8] Open Preservation Foundation, “2019 - 2020 digital preservation community survey,” Tech. Rep., 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.4066912>.
- [9] National Digital Stewardship Alliance, “2019 storage infrastructure survey,” Tech. Rep., 2020. doi: DOI10.17605/OSF.IO/UWSG7.
- [10] T. Bähr, M. Lindlar, S. Pohlkamp, S. Strathmann, M. Zarnitz, “Results of the nestor community survey 2019-2020,” in *ser. nestor Materials*, nestor, 2022 (Forthcoming).

DEVELOPING AN APPROACH FOR ARCHIVING DIGITAL AUDIO WORKSTATION PROJECTS: *A pilot study*

Dr Michael Brown

*Alexander Turnbull Library
Te Puna Mātauranga o Aotearoa
National Library of New Zealand
Aotearoa New Zealand
michael.brown@dia.govt.nz*

Valerie Love

*Alexander Turnbull Library
Te Puna Mātauranga o Aotearoa
National Library of New Zealand
Aotearoa New Zealand
valerie.love@dia.govt.nz*

Abstract – This paper concerns a current pilot study relating to contemporary popular music created on digital audio workstation (DAW) software, being undertaken at the Alexander Turnbull Library (part of the National Library of New Zealand Te Puna Mātauranga o Aotearoa). For the pilot we have collaborated with New Zealand music artist Luke Rowell to archive the production components for two albums. The study addresses the reality that, where music production once involved physical media such as magnetic tape, for the last 25 years it has largely shifted to the digital domain. While preserving the final musical product released to the public remains technically straight-forward, documenting the processes which artists now employ in digital production is far more challenging.

This paper will begin with some background about Luke Rowell's music, then consider DAW software and the archival challenges it presents. We will then cover the approaches taken by the Library and our progress to date.

Keywords – digital audio workstation, music archiving

Conference Topics – Innovation; Exchange

I. INTRODUCTION

Born in 1983, Luke Rowell is one of New Zealand's foremost synthpop and vaporwave musicians [1]. Over the last 22 years, he has released 15 albums and performed hundreds of gigs around the world. His track 'Gravy Rainbow', released under the alias Disasteradio, was a YouTube hit in 2010, with over one million views [2], while his albums as Eyeliner are

considered among the essential works of the global vaporwave movement. Much of his music is available

for free download and reuse under Creative Commons licenses [3].

Music writer Martyn Pepperell notes that Rowell "grew up in a family with interests that met at the intersection of art and technology" [4]. He adopted computer-based music technology at an early age, exemplifying recent generations of artists whose practice is wholly born-digital. His initial albums were composed using the freeware tracker-sequencer program Jeskola Buzz. Around 2010 Buzz ceased being actively supported and Rowell adopted the Steinberg DAW Nuendo as his main creative tool.

II. DIGITAL AUDIO WORKSTATIONS

Since the development of non-linear audio editing in the late-1980s, digital audio workstations such as Pro Tools, Logic Pro, Audacity and Nuendo have become central to music production. Garageband, which comes bundled with Apple devices, is a consumer-level example that many people will be familiar with. These applications replicate, streamline, and extend techniques first developed in recording studios, such as multi-track recording, tape editing, sound mixing, and signal processing (manipulation of audio signals). They also offer new creative tools relating to digital sampling (reuse of sections of audio), sequencing (programming of note or sample sequences), and sound synthesis. DAWs are now utilized in most music production contexts, from home-recording to being embedded in professional studios.

Why is it desirable to preserve the information represented by a DAW production? Research

libraries such as the Turnbull Library have traditionally sought to collect analogue items which contribute to the study of music as a creative process. Such sources might include manuscript scores and sketches, live recordings, and studio tapes, through which one can trace the genesis and refinement of musical works, including how recordings have been constructed in recording studios. The project materials generated by DAWs are the digital equivalent of what music archives have previously collected in analogue form. If these files are not preserved, we inadvertently create a born-digital void for future understanding of how most music is now created.

The complexity of DAW applications and general risks pertaining to digital material, however, present a daunting challenge for digital preservation. The objective of our pilot study, which we have come refer to as the Disasteradio Project, has been to develop sustainable workarounds for archiving the information contained in Rowell's DAW productions. We aim to complete the study in late 2022 and hope our findings might prove useful to music archivists, as well as artists wanting to secure long-term access to the digital componentry of their music.

III. DIGITAL PRESERVATION

As Riccardo Ferrante writes, a key challenge for the preservation of digital content is technical obsolescence:

“Born-digital objects require particular hardware and operating systems in order to be accessed and rendered with integrity. As technology evolves, parts of these technical environments are replaced with newer, faster, smaller components. The hardware and software necessary for a 10-year old object to operate as designed is often no longer readily available; neither is the skill set required to maintain that environment. Innovation in the marketplace is itself a risk to the cultural heritage it produces” [5].

Standard audio files such as WAV or MP3 are currently deemed as having a low risk of obsolescence. DAWs are another story entirely. When audio is loaded into, generated, and manipulated in a DAW, each decision is stored in a project file. This aggregate of editing, mixing and effects metadata refers to audio assets and/or coded audio sequences being used for the production. The production's integrity is thus dependent on both project file and source files. Adding further complexity is the almost ubiquitous use of plugins:

third-party software components that extend a DAW's functionality with virtual instruments and various forms of signal processing. Signal paths, the flow of audio signals from source to output, may also be routed through an external hardware environment including sound modules. Another challenge is that most DAWs are proprietary applications, each using its own project file format that cannot be opened on other platforms. Many lack open-source code that would offer greater interoperability. This poses real challenges to ensure that a DAW session can still be accessed into the future.

There are four basic digital preservation strategies that the Library considered in archiving DAW files:

- 1) *Maintain the original technical environment (hardware, software, plugins, etc.)*
- 2) *Replace the original software with a backwards-compatible application*
- 3) *Emulate by creating a virtual version of a suitable environment*
- 4) *Migrate the digital content into a new format that can be accessed [6]*

The first two of these were deemed impractical by the Library, especially given the varied permutations of DAWs, plugins, OS and hardware which would need to be maintained or replaced. Since the establishment of the National Digital Heritage Archive (NDHA) in 2008, the Library's digital preservation strategy has been based on migration [7] as the methodology for long term preservation of and access to digital collections. There are increasing numbers of tools available for virtualization of digital content, and the Library did consider emulation as a possibility. However, for the purposes of the pilot project, there seemed to be far less risk in expanding our current migration strategies to include DAWs, rather than attempting to build a new virtual environment for the collection. Given the complexity of the DAW files, even migration itself was no simple task.

At this point, we should note that issues with obsolete DAW projects have been recognized in the music industry. Artists have reported slowly but surely losing access to older DAW sessions, thus being unable to remix or rearrange their past work [8]. Losing the ability to carry out such iterative creative processes highlights another reason why strategies to allow ongoing access to DAW

productions are needed. As it happens, the music and video industries have developed several file formats to enable better operability across platforms, most notably the OMF (Open Media Framework) file or the newer AAF (Advanced Authoring Format) file [9]. With these container formats, source assets are preserved alongside editing metadata. Although primarily designed to enable projects to be operated across different DAW (and video editing) platforms, AAF files could allow for ongoing migration of projects from DAW to DAW to ensure future access. However, from a digital preservation perspective, such complex formats pose daunting challenges similar to those presented by native DAW project files. Nor does their use overcome issues arising from incorporation of multiple third-party plugins, each with its own OS and hardware dependencies [10]. While AAF files were not used in this pilot project, as they do not have the ability to adequately record plugin data that is paramount to Rowell’s music, preserving AAF files may be considered for other digital music collections.

IV. THE DISASTERADIO PROJECT

The Library’s collaboration with Luke Rowell began following the 2018 announcement that the cult music label Flying Nun Records was donating its master tape archive to the Turnbull Library [11]. Rowell contacted the Library to asked whether we were also preserving DAW project files. We didn’t have any DAW files in our digital music collections at the time but flagged this as an area for further research. The Library proceeded to develop a proposal with Rowell to archive two albums as a pilot study. The Disasteradio Project was approved in 2020 and work on the pilot study began. The two albums chosen for archiving were the Disasteradio album *Visions* (2007), created on Jeskola Buzz, and the Eyeliner album *Buy Now* (2015), created on Steinberg’s Nuendo. Our first discovery was that the Buzz projects for the tracks on *Visions* could no longer be opened correctly, a mere 13 years after being created. So, another Nuendo album, *Charisma* (2010), was substituted.

The Library’s basic approach to digitally preserving Rowell’s DAW projects has been a special form of migration. For each album track, Rowell has himself created a package of production components to give to the Library. If the Nuendo project file (.npr) represents the native digital object,

then we have found a range of alternative ways for the information it contains to be expressed and preserved. The Library considered the significant properties of Rowell’s electronic music files, and created a standardized file manifest for each track across the entire collection (see Table 1). Using the UK National Archives PRONOM registry as a guide, only low-risk formats are represented in the packages [12].

Table 1
Types of material within the Luke Rowell Digital Music Collection and file formats for each type of material.

Materials	File format
Track stems (dry) without effects and automation	WAV
Track stems (wet) with effects and automation	WAV
MIDI (Musical Instrument Digital Interface) types 0, 1 and General	MID
Audio samples (if used)	WAV
Working mixes	WAV, MP3
Final mixes, mastered and unmastered versions	WAV
Spreadsheet of technical information	XML
Screenshots of session settings	JPEG
Screencast with commentary	MPEG-4
Music videos and promos	MPEG-4

Certain components of these packages are relatively straight-forward. Musicians will be familiar, for instance, with exporting sets of individual tracks (for example, vocals, bass, drums) from a multitrack production as separate audio files, or stems. Such stems are readily preserved as digital objects and can be imported into any DAW for remixing. Mixdowns, created from mixes of individual tracks, are another standard output from a DAW project.

More novel approaches have also been taken for providing researchers with other avenues for investigating the fine details of Rowell’s music. He has compiled spreadsheets, for instance, that document the settings, plugins and signal paths for individual stems within each project, cross-referenced to other assets in the collection packages. Screenshots show the settings for every plugin used for every stem. For *Buy Now*, over 700 such images have been created and preserved. Rowell has also created screencasts of the Nuendo projects for all 11 tracks on this album. In these videos he gives a guided tour of the relevant session, disclosing his

creative decisions and documenting the user experience of a digital audio workstation for future researchers.

As a form of migration for digital preservation, the approach taken for the Disasteradio Project is new for the Library. Normally, digital archivists would manage the migration process following receipt of the original files, converting material into stable formats for preservation and access. The Disasteradio Project deviates from archival orthodoxy, migration being undertaken prior to transfer and by the donor themselves. In this case, Rowell is clearly the person who best understands his technical environment and creative process. He can therefore accomplish the migration task most effectively. In recognition of the work required by the donor to ensure future access and understanding of his work, the Library paid him for time spent preparing the collection.

V. DISCOVERY, RESEARCH AND RE-USE

In 2021, Rowell and the Library completed work on archiving *Buy Now*, which was primarily composed using Musical Instrument Digital Interface (MIDI) sequences. MIDI files aren't audio files themselves, but rather are performance instructions containing notes, tempo, instruments, volume, and other information to explain how the music should be played by a software program, electronic instrument, or other device. While the Library hasn't preserved the original DAW project files, we hope that by preserving the MIDI, stems, and mixes, and the ways in which we have preserved and documented the packages retains the significant properties of the collection and expresses the underlying abstract form of the files [13]. Through these packages, the collection affords the ability to remix and understand Rowell's technical process. The resulting collection material was released in May 2021 and is now available through the National Library of New Zealand website [14]. The Luke Rowell Collection is open access and available worldwide. Rowell has generously made the material available for online download under a Creative Commons BY-NC-SA license, which means the music can be remixed, resampled, mashed-up, and rearranged. We see the collection as having potential interest for researchers in the music studies and education sectors, in the music community, and for the wider public. As a promotional adjunct we invited other artists to contribute to a remix album based on the collection.

This album, *Free Buy Now Remixes*, was released on Bandcamp as a free download, alongside a blog providing further examples of potential reuse [15].

VI. CONCLUSION

Luke has since deposited materials relating to the Disasteradio album *Charisma*, with description and ingest almost completed by the Library. Following completion of the pilot study, we anticipate presenting a fuller analysis of the Disasteradio Project. *Charisma*, which includes a range of hybrid audio-MIDI compositions, has presented a more complex archival proposition than *Buy Now*. This recent experience has highlighted the variety of challenges that may still be faced using the method outlined in this paper, even within a single artist's oeuvre. We acknowledge that one of the main drivers of success of the project so far has been the close collaboration between the artist and the Library. Our interim conclusion, though, is that the approach holds considerable potential for digital preservation of DAW projects and may have some applicability for work in adjacent fields such as digital video and architecture.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for helping improve this iPres paper. We also thank the rest of the Disasteradio Project team (Flora Feltham, Zach Webber, Jessica Moran) and our National Library colleagues Kirsty Smith, Jay Gattuso, Mary Hay, Jay Buzenberg, Sholto Duncan, and Chris Szekely. Appreciative thanks also, of course, to Luke Rowell and the artists involved in the remixes album.

REFERENCES

- [1] Pepperell, Martyn. Disasteradio profile, *Audio Culture* website. <https://www.audioculture.co.nz/profile/disasteradio>
- [2] Disasteradio, 'Gravy Rainbow' music video, YouTube (2011). https://youtu.be/d-LKa1Y9_ok
- [3] Disasteradio on Bandcamp, <https://disasteradio.bandcamp.com/>
- [4] Pepperell, Martin. Disasteradio profiles, *Audio Culture* website: <https://www.audioculture.co.nz/articles/disasteradio-part-1> and <https://www.audioculture.co.nz/articles/disasteradio-part-2>.
- [5] Ferrante, Ricardo. 'Care of Born-Digital Objects' in Lisa Elkin and Christopher A. Norris, eds, *Preventive Conservation: Collection Storage* (Society for the Preservation of Natural History Collections et al), p.832.

- [6] Slade, Sarah, David Pearson and Steve Knight, 'An Introduction to Digital Preservation' in Lisa Elkin and Christopher A. Norris, eds, *Preventive Conservation: Collection Storage* (Society for the Preservation of Natural History Collections et al), pp.823-824.
- [7] Mosely, Sean, Jessica Moran, Peter McKinney, and Jay Gattuso. 'Conceptualising Optimal Digital Preservation and Effort', conference paper given at iPRES 2016, Bern, Switzerland.
- [8] McGuire, Colin. 'The Concrete and the Ephemeral of Electronic Music Production', *Dancecult* 6/1 (2014): <https://dj.dancecult.net/index.php/dancecult/article/view/457/460>
- [9] The open AAF standard is maintained by the Advanced Media Workflow Association, see: <https://www.amwa.tv/aaf/>; cf. <https://www.loc.gov/preservation/digital/formats/fdd/fdd000004.shtml>. For documentation of the Open Media Framework, see: <https://web.archive.org/web/20110713212349/http://www.linuxmedialabs.com/Downloads/LSI/omfspec21.pdf>
- [10] For a discussion of the practical advantages and difficulties of OMF and AAF, see: <https://www.pro-tools-expert.com/home-page/2020/8/27/aaf-and-omf-from-video-editors-how-to-make-sure-what-they-provide-works-for-us>
- [11] Brown, Michael. 'The Flying Nun Project: Tally Ho!' National Library of New Zealand blog post, 8 May 2019. <https://natlib.govt.nz/blog/posts/at100-new-collections#flying-nun-records-collection>.
- [12] The National Archives UK PRONOM database: <https://www.nationalarchives.gov.uk/PRONOM/Default.aspx>
- [13] Lee, Christopher, and Margaret Hedstrom. 'Significant properties of digital objects: definitions, applications, implications'. Proceedings of the DLM-Forum 2002 Parallel session 3.
- [14] See Luke Rowell Collection (ATL-Group-00554): <https://tiaki.natlib.govt.nz/#details=ecatalogue.1033472>
- [15] See: <https://disasteradio.bandcamp.com/album/free-buy-now-remixes>; and <https://natlib.govt.nz/blog/posts/download-now-free>.

GOING FOR GOLD OR GOOD ENOUGH?

Observations on three years of benchmarking with DPC RAM

Jenny Mitcham

Digital Preservation Coalition

UK

jenny.mitcham@dpconline.org

[0000-0003-2884-542X](tel:0000-0003-2884-542X)

Paul Wheatley

Digital Preservation Coalition

UK

paul.wheatley@dpconline.org

[0000-0002-3839-3298](tel:0000-0002-3839-3298)

Abstract - The Digital Preservation Coalition's Rapid Assessment (DPC RAM) was launched at the iPres conference in 2019. This digital preservation maturity model was developed with community input as part of a collaborative project with the UK Nuclear Decommissioning Authority. The DPC's hope was that it would be of broad utility to the wider digital preservation community, providing a flexible framework for assessing current capabilities and future goals. This short paper reflects on the three years since the publication of DPC RAM, discussing progress that has been made and lessons that have been learned since it was first launched. Future directions for the maturity model are also articulated.

Keywords - maturity modeling, assessment, benchmarking, community, DPC RAM

Conference Topics - Community; Environment

I. NUCLEAR BEGINNINGS

The Digital Preservation Coalition's Rapid Assessment Model (DPC RAM) was developed in the summer of 2019 as an output of a collaborative digital preservation project between the Digital Preservation Coalition (DPC) and the UK Nuclear Decommissioning Authority (NDA). The Digital Preservation Coalition is an international membership organization and global community. It enables its members to deliver resilient long-term access to digital content and services, helping them to derive enduring value from digital assets and raising awareness of the strategic, cultural and technological challenges they face. The Nuclear Decommissioning Authority has a remit to clean up the UK's earliest nuclear sites safely, securely and cost-effectively with care for people and the environment. The NDA is a member of the DPC and has a commitment to ensuring that the information and data that underpins its work and supports its

mission is effectively preserved and accessible for as long as it is required.

Reliable, Robust and Resilient Digital Infrastructure for Nuclear Decommissioning was originally planned as a two-year project which was to draw upon the experience of the DPC and its members to leverage good practice and to advise, guide and develop policy for the NDA. A key element of the project was to establish a methodology for measuring progress as the NDA established digital preservation policies and workflows. There was also a strong interest in benchmarking against the wider community, in order to help inform the NDA's future digital preservation goals.

There is no shortage of maturity models in digital preservation, and the project team at the DPC spent some time researching existing models with a view to adopting one to use with the NDA, however, it was eventually concluded that the available models didn't quite suit the task at hand. This seemed to be primarily down to the nature of the organization. Existing maturity models seemed more typically directed at libraries, archives and academic institutions and the language and concepts used did not transfer so easily into other sectors. Whilst the NDA certainly has an urgent need to preserve its digital information for very long periods of time, it is far from being a traditional 'memory institution'.

II. RAM IS BORN

Under the auspices of its work with the NDA, the DPC decided to develop a new maturity model that would be applicable to organizations of any size and sector, and suitable for all digital content of long-term value. The DPC were keen to develop a model that could be used not just with the NDA, but with all of its members, regardless of sector or context. The

model would also be made freely available to the wider digital preservation community.

The DPC felt that the model that was developed should be quick and easy to use. Completing a maturity modeling exercise should not be too onerous or off-putting a task. It was also agreed that the model should be based on existing good practice, without being too prescriptive about a particular method or approach to digital preservation. Given that there are often a variety of different approaches that can be taken to implement digital preservation, flexibility within the model was important.

The DPC chose to build on and update an existing model rather than starting from scratch. Adrian Brown's Digital Preservation Maturity Model [1] provided a flexible framework to build on and develop further. The original author offered support and encouragement to the DPC for developing this model further and provided valuable input into the revision and review process.

After a period of development and a number of rounds of community feedback from the wider DPC membership and Supporter organizations, the first version of the new maturity model was complete. DPC members were supplied a name for the model and visualizations for the worksheet provided for logging results. The resulting model, launched during the Lightning Talks at iPres 2019 in Amsterdam, was truly a collaborative community effort.

III. RAM: A QUICK GUIDE

DPC RAM provides a simple framework for carrying out a self-assessment of digital preservation capability [2]. There are eleven sections of RAM. The first six being organizational capabilities (covering issues such as organizational viability, policy and strategy, and legal issues) and the last five service level capabilities (focusing on more 'hands on' areas of digital preservation such as acquisition and ingest, metadata management and bitstream preservation). For each of the eleven sections of RAM (illustrated in table I), users of the model must pick one of 5 levels which best represents their current situation. At level 0, an organization would have no awareness of an issue, and at level 4 they would be fully optimized and managing that issue in a proactive way.

Examples are included within the model to indicate what types of activities *might* be in place for an organization to reach a particular level, but these examples are not intended to be prescriptive or a list that an organization must systematically check off. There is an ethos of flexibility built into the model -

Table I
The eleven sections of DPC RAM

Organizational capabilities		
A	Organizational viability	Governance, organizational structure, staffing and resourcing of digital preservation activities.
B	Policy and strategy	Policies, strategies, and procedures which govern the operation and management of the digital archive.
C	Legal basis	Management of legal rights and responsibilities, compliance with relevant regulation and adherence to ethical codes related to acquiring, preserving, and providing access to digital content.
D	IT capability	Information Technology capabilities for supporting digital preservation activities.
E	Continuous improvement	Processes for the assessment of current digital preservation capabilities, the definition of goals and the monitoring of progress.
F	Community	Engagement with and contribution to the wider digital preservation community.
Service capabilities		
G	Acquisition, transfer and ingest	Processes to acquire or transfer content and ingest it into a digital archive.
H	Bitstream preservation	Processes to ensure the storage and integrity of digital content to be preserved.
I	Content preservation	Processes to preserve the meaning or functionality of the digital content and ensure its continued accessibility and usability over time.
J	Metadata management	Processes to create and maintain sufficient metadata to support preservation, discovery, and use of preserved digital content.
K	Discovery and access	Processes to enable discovery of digital content and provide access for users.

it doesn't so much tell you what you need to put in place to implement digital preservation, but it does make some suggestions as to what may be appropriate in order to reach a particular level.

As continuous improvement is at the heart of DPC RAM, users of the model are encouraged to revisit their self-assessment on an annual basis to log progress and reframe targets. DPC members are encouraged to share these results annually with the DPC to facilitate community benchmarking opportunities and targeted support.

IV. RAM: THE EARLY YEARS

Post-launch it was encouraging to see enthusiastic use of RAM from many community members, both within and beyond the DPC. Wider usage of the model led to a modest accumulation of comment and feedback and the decision was made to address this with a new version of DPC RAM.

Version 2.0 of DPC RAM was released in March 2021. Avoiding any changes to the basic structure of the model, the revisions focused instead on clarifying the language and adding new examples.

One of the new themes that was addressed in version 2.0 was environmentally sustainable digital preservation. Inspired by the work of Keith Pendergrass, Walker Sampson, Tessa Walsh and Laura Alagna [3] on this topic, new examples within the model encourage practitioners to bring environmental considerations into their decision making on digital preservation issues alongside other factors such as financial cost, risk and user requirements.

As well as being a maturity modeling tool, DPC RAM is also providing a useful foundation for other DPC tools and initiatives, helping provide a shared reference point to map other resources to. Examples of this include the Novice to Know-How learning pathway, a new digital preservation skills framework and a set of core digital preservation system requirements.

Since its launch, DPC RAM has continued to be used by the DPC to inform a number of interactions with members. The DPC offers support and advice on any aspect of a DPC RAM assessment to their members - whether this be by answering questions about the model, providing anonymous benchmarking information, reviewing a self-assessment, or helping an organization to consider priorities and next steps. Through collating member self-assessments, it is possible to gain a basic understanding of some of the broad themes from DPC RAM assessments, some of which are discussed in the next section.

V. ANALYSIS

DPC members are encouraged to share their RAM assessments with the DPC on an annual basis. The DPC is committed to ensuring that the confidentiality of this information is respected. Aggregated information is shared with members to enable comparison and benchmarking, but the information shared does not include the individual scores of any identifiable organization.

Looking at information gathered from members in 2020 and 2021 some observations can be made with regard to the specific sections of RAM.

1) Organizational capabilities of RAM typically score higher than service capabilities: This was apparent on both years of data collection, with results for sections A-F typically being slightly higher than sections G-K

for most organizations. This is not unexpected given that foundational work on the organizational areas would most likely need to be in place before investment is made in digital preservation processes and procedures.

2) DPC members score highly at 'Community': Results have demonstrated how strongly DPC members score for the 'Community' section of DPC RAM. This section of the model is all about engagement with and contribution to the wider digital preservation community and this is clearly something that DPC members already make a firm commitment to. The DPC were keen to ensure that the value of this outward facing aspect of working in digital preservation was captured and recognized in some way within the framework of the maturity model.

3) Progress in 'Continuous improvement': This was one of the lower performing sections of the model in 2020 and the results for this section improved most strikingly in 2021. This was not an unexpected result given that one of the ways to move forward in this section is to carry out a regular self-assessment and benchmarking exercise, set targets and create a plan to move towards those goals. Getting a commitment to continuous improvement in place and an agreed schedule for check in and review will hopefully stand members in good stead for continuing to demonstrate progress in other areas of the model.

4) Lower scoring sections: 'Acquisition, transfer and ingest' and 'Content preservation' were the lowest scoring sections, and the sections with the biggest gap between current and target levels in 2021. In 2020 they were the lowest scoring after 'Continuous improvement'. There is a huge amount packed into the 'Acquisition, transfer and ingest' section of RAM. It is certainly one of the fullest sections in terms of examples included and it encapsulates a huge amount of practical action that needs to be taking place to move up the levels. 'Content preservation' relates to preserving the meaning or functionality of the digital content and ensuring its continued accessibility and usability over time. This section of RAM covers perhaps some of the most complex challenges of digital preservation and it is noted that practitioners may be choosing to focus their attention on other areas of the model at present.

VI. WHAT ELSE HAVE WE LEARNED?

After three years of supporting the community to use DPC RAM, the authors have several additional observations to make:

1) *Implementing digital preservation takes time (especially in the midst of a global pandemic)*: It was already suspected that implementing digital preservation was more of a marathon than a sprint even when times were good, but the first two years of data collection with DPC RAM coincided with the Covid-19 pandemic which led to a shift of priorities for many organizations. However, it was clear from the first few sets of results that members shared that although progress could be demonstrated using DPC RAM, it was typically small increments of improvement in one or two areas rather than sweeping changes across the board. In some cases it was also noted that scores went down rather than up. Again, this is perhaps not surprising given the upheaval and shifting of priorities that occurred during the Covid-19 pandemic.

2) *Not everyone needs to aim for the top*: When DPC RAM was first launched, the accompanying guidance suggested to users that they should carefully consider the target level that was most appropriate for them, and not necessarily assume they should aim for the top level across the board. Although 'gold standard' digital preservation is something that some DPC members will certainly be striving for, it is unrealistic to expect all organizations to aim for the top level (level 4). It has been encouraging to see the community embrace this approach. Members who shared results with the DPC in 2021 set their target levels on average at a level 3.3 (this figure is slightly lower than the previous year). It is recognized that an approach or target that is appropriate for one organization may not be realistic or achievable for another.

3) *Targets can change*: There have been several examples of organizations adjusting their targets over time. Adjustments can go either way, both up and down. Some organizations have realized that a lower level is actually more realistic for them (and also perfectly appropriate to meet their needs). Others have reached a previous target level, and have moved on to set their sights higher for next time. The DPC encourage those using the model to revisit both current and target levels on an annual basis. The opportunity to reflect on and refine goals, as well as measure progress towards them, appears to be valuable.

4) *Sustaining current levels also needs resource*: When DPC RAM was first introduced there was an obvious focus on supporting the community to move forward with RAM and move up through the levels. More recently, conversations have also touched on

how to sustain or maintain a particular level. As noted earlier, it is possible to slip down as well as move up levels with DPC RAM, and it is likely that some effort may be required to maintain current levels if organizations are not actively pushing forward in a particular area.

5) *The approach used to complete a RAM assessment can be significant*: RAM was designed to be quick and easy to use. It is possible for a digital preservation practitioner to complete a RAM assessment in an hour if they have all the information at their fingertips. This however might not be the most impactful way to proceed. Feedback from DPC Members suggests that there can be additional benefits when RAM is applied collaboratively with a group of colleagues. Not only does it balance out some of the inevitable subjectivity that is introduced when one person goes it alone, but it is also a helpful way of bringing colleagues on board with digital preservation goals and enabling them to become more invested in the challenge. Even an assessment that is done as a solitary exercise may still be shared and socialized with colleagues through any number of channels.

6) *RAM can be an effective communication tool*: Although RAM was designed primarily as a means of measuring progress, use with DPC members has demonstrated a wider utility. RAM can be applied as a powerful tool for advocacy and communication as it breaks down the complex topic of digital preservation into a simple set of metrics that can be quickly and easily shared and communicated with colleagues. Having a simple visualization showing where you are, where you would like to be (and perhaps even where others in the community are) is a powerful way of illustrating capability gaps and the need for further resourcing.

7) *There is more than one way to do digital preservation*: This point was recognized when the model was first developed, and it has been a theme throughout the life of DPC RAM. Maturity models and certification frameworks by their very nature tend to point the user in a particular direction regarding 'the right way to do things'. This approach can disguise some of the complexity around digital preservation decision making, despite being helpful to users who would like to know what should be put in place to implement digital preservation. For example, if a model states that three copies of the digital content should be maintained, this doesn't allow for local priorities or variations to be considered. Perhaps an organization has valid reasons why three copies are

not required in specific circumstances (particularly when other factors such as economic or environmental cost are factored in or if the digital content is considered to be lower value).

Rather than being prescriptive, DPC RAM focuses on the elements that go into a decision-making process around how digital preservation is enacted. This may involve many different variables such as resource (human and financial), value of the content, needs of the users, perceived risks and impact on the environment. Wrapping this within the rather rigid framework of a maturity model is challenging but worth the effort.

VII. WHAT NEXT?

It has been encouraging to see how DPC RAM has been adopted by the international digital preservation community over the last three years. Thinking forward to the next three years, the authors hope to see the following:

1) *Better metrics*: Already it is possible to see some broad trends from RAM assessments that have been shared with us by our members, but this is only from a proportion of the membership and represents a very short time period. The DPC are keen to continue to gather and collate DPC RAM assessments to gain a fuller overview of trends, sticking points, and speed of progress. By understanding our members better, it will be possible to provide appropriate support and guidance in the future.

2) *Better support*: work has recently been carried out to enhance advice and guidance on the DPC website on how to move up the levels of RAM. A 'RAM Jam' workshop was also held, which enabled members to share tips and experiences about how they moved up to the 'basic' level of RAM. Further workshops such as this will be held over the next few years and online guidance will continue to be developed and enhanced as a result. The DPC will continue to provide direct support to their members with their annual RAM assessments where this is requested.

3) *More case studies*: A number of case studies have been published [4] that describe how DPC RAM has been used within different types of organization. These insights are useful points of reference for others who are considering using the model. There is an intention to publish more of these in the future.

4) *Further translations*: DPC RAM has been translated into Spanish, French, Italian, Portuguese, and Japanese by community volunteer translators [5]. This is an encouraging step towards greater

accessibility to a wider international audience. The DPC hopes to see the number of translations continue to grow over subsequent years.

5) *A new version*: There is an intention to revise RAM again within the next three years. Digital preservation is an evolving field, and it is important that DPC RAM continues to be responsive to community feedback as good practice further evolves and develops. Feedback on DPC RAM can be submitted at any time via the DPC RAM website. All feedback is welcomed and will be carefully considered for inclusion in a future version.

VIII. CONCLUSION

Maturity models such as DPC RAM can enable organizations to progress more effectively with their digital preservation development by facilitating better awareness of their current capabilities, enabling more realistic and relevant targets to be set and the construction of development roadmaps. The flexibility of DPC RAM in particular enables organizations to not only decide how they will carry out a particular aspect of digital preservation but also define what 'good enough' looks like for them. Though the DPC regularly stress that RAM is about continuous improvement, improvement should only continue so long as it is necessary. Focus should also be placed on maintaining capability at an appropriate level where a suitable target level has been reached.

The first three years of supporting RAM has been an invaluable learning experience for the DPC. Using RAM as a consistent framework within which to have conversations about digital preservation with members has enabled a greater depth of understanding about digital preservation capabilities across the community and a more quantifiable base of evidence relating to its strengths and weaknesses. This increased understanding will be beneficial to the DPC in planning future activities to help address knowledge gaps or barriers to progress that have been highlighted. It is anticipated that DPC RAM will continue to provide structure to the member support services delivered by the DPC going forward as well as being freely available for the whole community to benefit from.

REFERENCES

- [1] A. Brown, *Practical Digital Preservation: a how-to guide for organizations of any size*, Facet Publishing, 2013.

- [2] Digital Preservation Coalition, *Digital Preservation Coalition Rapid Assessment Model (DPC RAM)*. Version 2.0, 2021. <http://doi.org/10.7207/dpcram21-02>
- [3] K. Pendergrass, W. Sampson, T. Walsh, and L. Alagna, "Toward Environmentally Sustainable Digital Preservation." *The American Archivist*, vol. 82, no. 1, pp. 165–206, 2019.
- [4] <https://www.dpconline.org/digipres/implement-digipres/dpc-ram>
- [5] <https://www.dpconline.org/digipres/discover-good-practice/translations>

MONITORING BODLEIAN LIBRARIES' REPOSITORIES WITH MICRO SERVICES

Edith Halvarsson

*Bodleian Libraries,
University of Oxford
United Kingdom
edith.halvarsson@bodleian.
.ox.ac.uk*

James Mooney

*Bodleian Libraries,
University of Oxford
United Kingdom
james.mooney@bodleian.o
x.ac.uk*

Sebastian Lange

*Bodleian Libraries,
University of Oxford
United Kingdom
sebastian.lange@bodleian.
ox.ac.uk*

The Digital Preservation Micro Services and Reporting (DPMS) service is Bodleian Libraries' monitoring system for digital collections. DPMS interacts with the storage layer of the Libraries' existing repository systems. This method is an alternative to the monolithic systems model for digital preservation [1]. The decoupled nature of DPMS has meant that preservation tools could be incorporated into the Libraries' existing repository services, without needing to migrate assets to a separate digital preservation platform.

This paper outlines how the DPMS service was developed. It describes the components of DPMS's technical framework and the process of implementing it in Bodleian Libraries' digital repositories

**Keywords – Micro Services, Open Source, Reporting
Conference Topics – Innovation**

I. BACKGROUND

Bodleian Libraries is a group of 28 libraries that serve the University of Oxford (England). It is the largest academic library service in the United Kingdom and one of the largest library services in Europe [2]. In addition to its sizable physical collections, the Libraries have collected born-digital content for over 15 years. During this time the Libraries have developed specialist repositories for digitized content, born-digital archives, research data and research publications. The Libraries' digital collections are now primarily managed in three core services: 1) Digital.Bodleian, 2) the Oxford University Research Archive (ORA/ORA-Data), and 3) Bodleian Electronic Archives and Manuscript (BEAM).

In 2017 Bodleian Libraries and Cambridge University Library (CUL) undertook a joint market review of digital preservation systems to assess their

suitability for the Libraries' existing repositories. The review was completed as part of a research project called the Digital Preservation at Oxford and Cambridge (DPOC) project [3]. At the time of the review, none of the assessed systems met all the different repositories' essential requirements. The option of migrating all digital collections to one joint monolithic digital preservation system was therefore deemed unsuitable, as it would have resulted in the repositories being unable to undertake some of their core administrative activities. A micro services architecture provided the Libraries with an alternative approach for monitoring and reporting on digital collections, where digital files could still be retained in their respective heterogeneous repositories.

In collaboration with Dave Thompson, Digital Curator at the Wellcome Library, a digital preservation solution which could work with existing digital repositories was scoped. This resulted in a paper for Library Hi Tech in 2018. The paper introduced a design proposal for a new architecture to work with Big Data volumes of preserved digital resources [6]. This paper came to form the basis for a business case to realise a micro services based digital preservation approach at Bodleian Libraries. In 2018 funding was secured for an initial proof of concept from the University of Oxford's IT Capital fund. The funding enabled the creation of a micro services reporting platform which has been in usage in the Libraries for the past two years, with new micro services being added to the platform on a regular basis.

II. MICRO SERVICES ARCHITECTURE OVERVIEW

The novel features of micro services architecture are a focus on small size and interdependency. Micro services are typically created to address a single function/capability and operate independently of each other, only communicating with other micro services via their published interfaces [5]. While micro services can introduce additional administrative overhead compared to a monolithic system, they can also bring great benefits. Among these are scalability (each micro service can be scaled individually to meet service needs) and portability. Micro services can be deployed across different heterogeneous platforms [5]. As illustrated by the micro services approach proposed by California Digital Library in 2010, micro services can bring great benefits to diverse digital curation infrastructures as they “can be deployed in the context in which it makes most sense, both technically and administratively” [1]. In Bodleian Libraries, the usage of micro services has for example enabled the Libraries to scale individual micro services to meet collection growth needs over the past two years.

III. MICRO SERVICES UPTAKE

While micro services have been discussed in the digital preservation field for over 10 years, it is not yet a commonly adopted institutional approach. Based on the experience of the Libraries, it is possible that the more modular costing model involved with a digital preservation micro services approach (potentially covering multiple individual support contracts and inhouse staff costs) is prohibitive for organizations needing to cover all digital preservation activities within a single support contract. This could however shift in the future if a micro services reporting framework was provided as a commercial solution.

IV. THE DPMS PROJECT

The Digital Preservation Micro Services and Reporting (DPMS) project begun in 2018 and is scheduled to complete in early 2023. During the DPMS proof of concept (2018-2019), the service’s technical framework was developed. The platform, based on Elastic Stack (ELK), is described further in the technical overview section below. The technical framework from the proof of concept formed the basis for all micro services which were subsequently added to DPMS.

Development of the DPMS service is running in sequential stages. Each new stage looks at a particular category of digital preservation tools/micro services. These categories are as follows:

1. File integrity
2. Virus checking
3. Backup and restore analysis
4. Characterisation
5. Validation
6. Digital preservation copy monitoring

Once a stage is completed, the new micro services are deployed in production so that they can be actively used by the Libraries core repositories. So far, micro services relating to categories 1-4 have been developed. Micro services relating to the final categories (5-6) will be developed in 2022.

V. THE DPMS FRAMEWORK COMPONENTS

The DPMS framework is built on open-source tools. The Libraries’ preference is for using supported open-source projects, rather than writing custom tools [4]. DPMS has incorporated the outputs from several open-source tools and utilities (such as Siegfried, MediaConch, ExifTool, Zabbix, rsync and others). These tools and utilities make up the individual micro services offering. Each was assessed during the project to ensure that it was well supported and (where relevant) could provide metadata as JSON output.



Figure 1: Example of a Grafana dashboard

The DPMS platform itself is built on Elastic Stack (ELK). It was chosen as the preferred framework, as an instance of ELK was already actively maintained by the University of Oxford’s IT Services on behalf of the wider University. Using the existing stack took some of the overhead of running the service away

from the Libraries. The ELK stack comprises of Elasticsearch (a search engine), Kibana (a data visualisation tool), Beats (for centralizing log data) and Logstash (a processing pipeline). In addition to the Kibana software, Grafana (another open-source visualization tool) is also used for creating graphs which provide DPMS users with a more high-level overview of their data.

VI. THE DPMS WORKFLOW

DPMS interacts with the Libraries' existing repository systems via their back-end storage. Files held in the repositories' storage areas are scanned by the micro services. Each repository can mix and choose which micro services are most relevant to their collection profile. Metadata output from the micro services scans are aggregated into JSON log files via a log merging tool.

A copy of the JSON log file is sent to preservation storage and another copy is sent to Oxford University IT Services for indexing in ELK. Elasticsearch provides the search engine for interrogating metadata gathered by the micro services. This metadata can then be visualised in Kibana and Grafana for the end user.

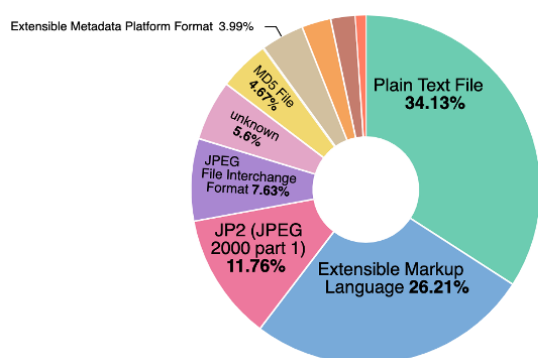


Figure 2: Example of a Kibana visualization (breakdown of file formats)

From the repository user's perspective, DPMS consists of dashboards which aggregates statistics about digital files and the storage they are held on. Grafana provides the main high-level overview of the repositories' digital collections and is generally how users choose to first access their statistics. The Grafana dashboards then link out to Kibana for more detailed statistics about the collections. Users have the option to create more advanced queries and searches in Kibana, or to drill down into the detail about particular files. Where issues have been highlighted by DPMS (such as a virus or deleted files), these are highlighted in the high-level Grafana

dashboards for further investigation. A ticket for the issue is also raised in GitLab, which is the Libraries' central location for technical documentation.

Micro services scans are repeated every 1-3 months, with fixity monitoring always undertaken on a monthly basis. DPMS is currently tracking approximately 58 million files (300TB of storage) across the Libraries' repositories. As ELK retains historic logs, DPMS can also illustrate changes to digital collections over time.

VII. SUPPORTING THE DPMS SERVICE

The DPMS service is run by the Libraries' digital preservation team, with support from the University of Oxford's ELK service. The digital preservation team are responsible for scheduling scans, aggregating JSON logs, onboarding new collections, providing training, and assisting with user queries. In total, the Oxford ELK team and the Libraries' digital preservation team dedicate 0.8 FTE to running the DPMS service.

However, the Libraries' repository staff are by necessity also actively involved in the preservation of their digital collections. Repository service owners are responsible for reviewing and (where possible) investigating preservation alerts from DPMS. As experts on their content and workflows service owners are often best placed to interpret 'unusual' activity in their repository, such as large-scale deletion of files or unexpected colour profiles in images. The time spent by service owners on investigating preservation alerts vary from each repository (with more active repositories requiring more staff engagement) and on how long the repository has been scanned by DPMS. As service owners become more familiar with the DPMS reports and findings each month, the time spent on investigating issues generally decreases.

VIII. ONBOARDING NEW REPOSITORIES TO THE DPMS SERVICE

Four of the Libraries' repositories currently use the DPMS service. When onboarding new repositories, the digital preservation team follows a standardised onboarding workflow. It can take several scans to fine-tune the Grafana and Kibana dashboards, to ensure that they meet the needs of the new repository. The onboarding workflow consists of the following steps:

- The repository owners provide initial set-up information for configuring the DPMS dashboards. Including:
 - Overview of the repository directory structure
 - How/when digital collections are updated
 - User credentials
- The repository owners are provided with training on using Grafana, Kibana, and GitLab
- DPMS completes an initial scan of the files in the repository
- The digital preservation team reviews the scanning results together with the repository service owner
- DPMS scanning patterns are updated to exclude certain files and directories if appropriate
- The digital preservation team creates a generic reporting dashboard for feedback
- The repository service owner provides feedback, and the dashboard is updated accordingly
- A configured JSON metadata log is sent to ELK for indexing

Once a repository has been onboarded it is added to DPMS's regular monthly scans.

IX. NEXT STEPS

Bodleian Digital Libraries Systems and Services department will take on the full system support costs following the completion of the initial DPMS project, covering software support contracts and inhouse staff (the Libraries' digital preservation team). The creation of the DPMS service is only an initial step towards enhancing the Libraries' understanding of its digital collections. Going forward, the Libraries will look at using the DPMS generated metadata to address areas of preservation concern. Such activities will include decreasing the percentage of unknown file types in the collections (by submitting new file signatures to PRONOM) and assessing gaps in available viewers and file conversion tools.

X. CONCLUSION AND SUMMARY

DPMS is an example of an alternative model for digital preservation monitoring using a micro services approach. This paper has illustrated how

Bodleian Libraries could utilise existing digital preservation tools and open-source frameworks to improve its monitoring capacity. The DPMS service model may also have applicability to other organizations who do not want to migrate from their current repository systems, or who have chosen to not implement a DAMS like system. As illustrated by the DPMS service, the knowledge which can be gained from implementing only a few micro services can greatly improve an organization's understanding of its collections.

ACKNOWLEDGEMENTS

Many thanks to Kristian Kocher (University of Oxford) for his continued guidance and expertise on ELK. Thank you also to Bram Hauer and Matt Thorpe (Bodleian Digital Library Systems and Services) for their work supporting the DPMS project. Last but not least, we would like to acknowledge the contributions of the late Dave Thompson and thank him for his support and forward-thinking spirit.

REFERENCES

- [1] Abrams, S., Kunze, J., Loy, D. (2010). 'An Emergent Micro-Services Approach to Digital Curation Infrastructure', *The International Journal of Digital Curation*, vol. 5, no.1, pp.172-186
- [2] Bodleian Libraries (2022). 'About the Libraries'. Available at: <https://www.bodleian.ox.ac.uk/about/libraries> (Accessed: 26 January 2021)
- [3] Bodleian Libraries (2020). 'Holding Page: Digital Preservation at Oxford and Cambridge'. Available at: <http://www.dpoc.ac.uk/> [Accessed: 26 January 2021]
- [4] Bodleian Libraries (2018). Digital Preservation Policy. External policy. Available at: <https://web.archive.org/web/20211121141920/https://www.bodleian.ox.ac.uk/sites/default/files/bodreader/documents/media/digital-preservation-policy.pdf> (Accessed: 26 January 2021)
- [5] Dragoni N., Lanese I., Larsen S., Mazzara M., Mustafin R., et al. 'Microservices: How To Make Your Application Scale'. A.P. *Ershov Informatics Conference* (the PSI Conference Series, 11th edition), Jun 2017, Moscow, Russia. ffh1-01636132f
- [6] Gerrard D.M., Mooney J.E, Thompson D. (2018). 'Digital Preservation at Big Data Scales: Proposing a Step-change in Preservation System Architectures', *Library Hi Tech*, vol. 36, no. 3, pp.524-538, <https://doi.org/10.1108/LHT-06-2017-0122>

FROM RAY CATS TO DPC RAM:

HOW BEST TO PRESERVE A DIGITAL MEMORY OF THE NUCLEAR DECOMMISSIONING PROCESS

Michael Popham

Digital Preservation Coalition
United Kingdom
michael.popham@dpconline.org
[0000-0002-6842-4294](tel:0000-0002-6842-4294)

Jenny Mitcham

Digital Preservation Coalition
United Kingdom
jenny.mitcham@dpconline.org
[0000-0003-2884-542X](tel:0000-0003-2884-542X)

This paper describes the work and outputs of the joint NDA-DPC project “Reliable, Robust and Resilient Digital Infrastructure for Nuclear Decommissioning”. This four-year project has produced a number of important deliverables that have been widely adopted by the international digital preservation community, such as DPC RAM. As a result of this project the UK’s Nuclear Decommissioning Authority is now much better placed to face the very real digital preservation challenges of the future.

Keywords – DPC, NDA, RAM, knowledge, sharing
Conference Topics – Resilience.

I. INTRODUCTION

The civil nuclear industry in the UK has already been in operation for over six decades but the legacy of this work, and of future developments in energy production, will have an impact for many thousands of years to come affecting hundreds of generations. Unlike most other sectors, the nuclear industry has clear and demonstrable use-cases that will require digital preservation planning for millennia.

The Nuclear Decommissioning Authority (NDA) is charged with the complicated task of decommissioning and cleaning the seventeen principal nuclear energy plants in the UK, a task accurately described as the largest and most important environmental restoration program in Europe. The extended life cycle of the program, set alongside robust commitments to security, integrity and safety, means the NDA approaches its work with a profound commitment to long-term information management, ensuring the right information is available to the right people in a format they can use and with the confidence that it can be trusted. Therefore, amongst its many challenges, the NDA is

by default required to become a trusted leader for information management and digital preservation.

In November 2018, the NDA and DPC began a dedicated project: “*Reliable, Robust and Resilient Digital Infrastructure for Nuclear Decommissioning*” [1] with three broad aims, namely to enable the NDA to:

- access and secure critical legacy data and systems
- adapt current data and systems to ensure their long-term viability
- commission data and systems with long term resilience from the outset

Initial findings were presented in a panel session at iPRES 2019, “*Achieving Criticality of Preservation Knowledge: Sustaining Digital Preservation in the Nuclear Field*” [2], and the project fostered a number of activities (notably the development of the *DPC Rapid Assessment Model* [3]) which have been taken up by the wider digital preservation community. The project was subsequently extended by a further two years – until November 2022 – with the goal of producing some specific deliverables and guidance, which will be described in this paper.

II. SCALE AND SCOPE OF THE NDA’S DIGITAL PRESERVATION CHALLENGE

Operating at an industrial scale on the cutting edge of a highly technical and complex area of scientific activity for such an extended period, has meant that the seventeen sites for which the NDA is responsible have an extensive legacy of digital data, applications, and systems. For much of the time vital records and data have been held on paper form, managed in line with prevailing records

management and archival practices. However, in recent decades, ever-increasing volumes of information and data have been created, managed, and kept in digital form. Nowadays there are immense swathes of information, such as mapping data held in GIS systems, in Building Information Management systems (BIMs), and virtual simulations, which will only ever exist in digital form but which play a vital part in the operation and management of the UK's nuclear industry, and will need to be securely preserved for the future.

In 2011, the Nuclear Energy Agency (NEA) of the Organisation for Economic Co-operation and Development (OECD) sponsored an international initiative "*Preservation of Records, Knowledge and Memory (RK&M) Across Generations*" [4], which produced an extensive set of recommendations in its final report in 2019. The focus was particularly on documenting the final disposal of radioactive waste and protecting humankind and the environment against the effects of ionizing radiation, noting "...it is not just a question of handing down a message, but of keeping that message interpretable, meaningful, credible and usable over time". (*ibid.* page 13). Earlier efforts around RK&M had focused on "avoiding inadvertent human intrusion [at disposal sites] through messages and methods focusing on danger and promoting aversion" (*ibid.*). One such effort is described in the short section on "The Ray Cat" (*ibid.* page 24):

Philosophers F. Bastide and P. Fabbri also responded to the 1984 poll asking how to communicate across 10,000 years Their proposal consisted of two steps:

- 1. Engineer a cat that changes color in response to radiation.*

- 2. Create a culture around this cat, such that if your cat changes color, everybody knows you should move someplace else.*

Ray cats would be genetically modified as to change color when coming near to radioactivity, thus serving as living indicators of danger... The choice for a cat was due to their long history of cohabitation with humans. In order to transport the message, the importance of the cats would need to be set in collective societal awareness. To this aim, Bastide and Fabbri proposed storytelling and myths, which could be transmitted through poetry, music and painting. As such, the meaning of the "feline Geiger counter" should spread and become culturally embedded over time.

Nearly 40 years later, our thinking around the appropriate documentation of nuclear waste

disposal sites has moved on somewhat, although it is much more prosaic – concentrating on topics like suitable long-term digital file formats and appropriate metadata fields i.e. the typical concerns of digital preservationists and archivists. However, these aspects seem far more likely to ensure that messages (aka information) remain "...interpretable, meaningful, credible and usable over time" (*ibid.*) when compared to poetry or tinkering with feline DNA, although perhaps some of our colleagues working on DNA-based digital storage might disagree?

The initial work of the NDA-DPC project involved DPC staff visiting several of the nuclear sites around the UK, and having discussions with colleagues across the NDA group on how best to manage and preserve their digital records and data.

III. DEVELOPMENT OF THE RAM

It was as result of efforts to assess the preservation readiness of the NDA that the DPC developed its Rapid Assessment Model (DPC RAM), a new digital preservation maturity model for organizations with a need to preserve digital content for the long-term. This tool was publicly released in September 2019, and was enthusiastically received by many digital preservation practitioners as both an easy-to-use self-assessment tool, and a helpful aid in discussing digital preservation capabilities and aspirations with colleagues.

The DPC RAM has been used by many organizations both large and small, and in response to feedback from the wider digital preservation community, a revised version of RAM was released in March 2021. Version 2 of RAM retains the 11 sections and 5 maturity levels of the original model, but enhancements were made in a number of areas, including an increased emphasis on user needs, and greater attention given to the legal aspects of digital preservation activity. Both versions of the RAM were tested with the NDA, and information managers across the NDA group were actively encouraged to undertake their own RAM assessment to gain insights into their preparedness for digital preservation.

IV. DEVELOPING KNOWLEDGE AND SKILLS

The NDA employs around 300 people but has overall responsibility for a large and diverse workforce, almost 20,000-strong. It was envisaged from the outset of the project that in order to make

progress towards its digital preservation goals, the NDA would need to develop the knowledge and skills of its people. Given the limited time and resources available, the NDA asked the DPC to commission the production of a number of data type guidance notes, based on the successful series of Technology Watch reports already published by the DPC.

Released in July 2021, the Data Type Series of Technology Watch Guidance Notes were written by experts at Artefactual Systems, working in collaboration with staff from the DPC. Each of these short documents is designed to provide a primer on the current state of community knowledge about preserving common types of data, such as documents, spreadsheets, and moving images. Ten such reports are available at the time of writing but their successful reception has encouraged the DPC to consider them a useful model for future publications on a specific theme.

Publications have also been accompanied by some face-to-face activities, albeit slightly hampered by the advent of covid-19. The NDA asked the DPC to establish a dedicated taskforce to develop some advice and guidance on the preservation of records held within an Electronic Document and Records Management System (EDRMS). Eighteen people from across the DPC's membership contributed to the work of the taskforce, which resulted in a Briefing Day (Unbroken records: A briefing day on Digital Preservation and EDRMS) and an online booksprint to produce the EDRMS Preservation Toolkit.

The project has also enabled the DPC to redevelop and enhance its thinking around workforce development, by defining the knowledge, skills, and capabilities required to fulfil particular roles in the digital preservation lifecycle. Although earlier models exist (such as the DigCurV framework), which have been widely adopted within the preservation community, they have not been maintained and updated to reflect the developments in digital preservation that have emerged within the past decade. The collaboration between the NDA and the DPC provided the impetus to develop a new framework and audit tool, which the NDA could use to identify and fill gaps in the knowledge and skills of its workforce. At the time of writing, this framework and audit toolkit were scheduled to be publicly launched at the iPRES2022 conference.

V. SO WHAT ELSE IS NEW?

The joint NDA-DPC project has also prompted activities in two other key areas that are of

widespread interest to the international digital preservation community. In the spring of 2022, the DPC publicly released a statement of the ten high-level function requirements for a digital preservation system. This document was originally developed to inform an assessment of the capabilities of the NDA's core digital preservation infrastructure, addressing aspects originally identified by the NDA's own RAM exercise. It soon became apparent that this document could be adapted to simplify and enhance digital preservation system procurement for both the procurer and for 3rd parties responding to procurement exercises – as challenges around this process had been previously identified in a workshop attended by DPC Members and Supporters. Although the statement of requirements is closely coupled to the DPC's Member-only Procurement Toolkit, it was agreed that given the NDA's public mission and the potential benefit to the wider digital preservation community, this document would be made publicly available.

Discussions with information managers across the NDA also confirmed that they recognized the challenges of preserving databases, of which a huge variety exist across the 17 sites managed by the NDA. They shared the challenge faced by many large, complex, and long-lived organizations of having important records languishing in legacy databases, as well as databases which are still being actively consulted despite having reached end-of-life and no longer being actively supported. In response to the challenges and risks presented by these data, the NDA and Sellafield Ltd. commissioned the DPC to undertake a focused one-year sub-project to develop some good practice guidance for the preservation of databases. This work began in February 2022, and we anticipate that initial findings may be presented at iPRES 2022.

VI. CONCLUSION

Whilst it seems unlikely that the preservation of the nuclear industry's digital records, knowledge and memory will ever depend on color-changing Ray Cats, colleagues within the NDA are now relying on a new maturity model called the RAM and the combined knowledge of their staff and the wider digital preservation community to help ensure their vital digital records are retained and maintained through future generations. Through its collaboration with the DPC, the NDA has been able to both draw upon the expertise of the wider digital

preservation community, and also share its own knowledge and findings for the benefit of all.

ACKNOWLEDGMENT

The authors are indebted to colleagues across the NDA for their time, energy, and support with the work of this project.

REFERENCES

- [1] <https://www.dpconline.org/digipres/collaborative-projects/nda-project>
- [2] <https://ipres2019.org/program/?session=46>
- [3] <https://www.dpconline.org/digipres/dpc-ram>
- [4] https://www.oecd-neo.org/jcms/pl_15088

CARING FOR BORN DIGITAL VIDEO CAMERA ORIGINAL FORMATS

CONSIDERING INTENTIONAL CHANGE

Crystal Sanchez

Smithsonian Institution

United States

sanchezca@si.edu

Over time, we have seen the exponential growth of born-digital files, specifically those created by consumer, prosumer, and professional cameras. This paper is about a new issue that is specifically rooted in today's digital workflows: the maintenance of born-digital camera original video formats in an archival setting and the intentional irreversible change that may be required during processing to stabilize them for future access.

Keywords – born digital; digital preservation; file-based; video production

Conference Topics – community

I. INTRODUCTION

The Smithsonian Institution (SI) has been working steadily towards understanding and building approaches to responsibly care for born digital video files created by production teams. Born digital file care has been an active area of the field for many years but approaching files that originate from video cameras brings its own unique challenges, and the archive is beginning to see their growth due to a variety of reasons rooted in changing production tools and workflows. This paper is about a new issue that is specifically rooted in today's digital workflows: the maintenance of born-digital camera original video formats in an archival setting and the *intentional irreversible change* that may be required to stabilize them for future access.

The Smithsonian Institution centralizes digital collection asset storage with an enterprise Digital Asset Management System (SI DAMS) [1]. As part of the Smithsonian DAMS team, we work closely with archivists, collection managers, registrars, and

conservators across the Smithsonian's many units, as well as advancement, communication, and video production professionals. Over the past ten years, the SI DAMS team has worked hard to support video in the system with implementation of tools, metadata extraction, and technical documentation for all files in our care. Over time, we have seen the exponential growth of born-digital files, specifically those created by consumer, prosumer, and professional cameras.

A. *A Shift to Camera Original Files*

Work with all of these stakeholders has shifted over the years. At first, video files were predominantly received as ProRes wrapped in Quicktime (MOV). It was most common for creators using Final Cut Pro to transcode files from camera originals into ProRes codecs during transfer to computers for editing. In this workflow, ProRes [2] is the Master file; camera originals were deleted when cards were re-formatted for reuse. Jonah Volk, writing in 2009, but anticipating the future, expresses this, "While current workflows generally involve transcoding media to QuickTime for use in editing software, it is not inconceivable to imagine that Apple might in the near future allow for native MXF editing in QuickTime, as Adobe Premiere already does." [3]

As the Smithsonian Institution rapidly shifted to Adobe products around 2015, video producers also shifted to Adobe Premiere for editing. Adobe supports a wide variety of codecs, allowing users to edit with most camera original formats without transcoding. This saves producers time in the workflow but creates new challenges for archivists.

As Adobe themselves say in their guide, “If you could only edit native formats, you probably would” [5].

As a result of this shift in production practices, we surveyed the video production practices at the Smithsonian in the Spring of 2018. The survey asked staff to describe their current production tools, practices, and pressing archival needs.

Sixteen participants from ten Smithsonian museums discussed their current practices. Overall, most producers reported maintaining a variety of camera original file formats as raw masters, only exporting lower resolution derivative edit masters for delivery. We also saw trends in the cameras and tools they used; producers referenced the same handful of cameras, and most reported using Adobe products to create produced pieces. We anticipated receiving a variety of camera original born-digital files in the near future. The survey results highlighted the need to develop a risk-assessment approach to analyze specific, common, born-digital, camera original video formats to develop some recommended practices for our community [6].

Technical Data Research & Analysis was conducted in 2019 on SI DAMS repository’s 42,823 video files, findings of which suggest this growth trend. In analyzing video codec data, ProRes codecs made up 12.5% of the total in 2019 (39.5% AVC/h.264; 22% unknown; 9.5% DV; 7% MPEG-2; 6.5% Motion JPEG; 2% Uncompressed 10 bit; 5% misc) [4]. In looking at file format data, totals showed 73% QT(MOV); 17% MPEG; and just 7.5% MXF.

The preceding years indicate an increase in total video files to 76,520 files at the beginning of 2022, up 44% from 2019). MXF wrapped files increased from 7.5% in 2019 to 10% in January 2022. This indicates to us that the shift in production practices will likely increase the deposit of raw footage to the archive from edit masters formats (ProRes) to camera original formats.

B. Researching Camera Original Formats

After this analysis, my colleague Taylor McBride and I began a crowd sourced project to inventory & research a dozen formats encountered at SI and connect them to cameras, encouraging shared documentation of format specifics through open tools (ie, Google Sheets) [7]. We presented this work to the field at AMIA 2018 to gain community feedback and encourage participation [8].

The SI DAMS team turned this work into a Supported File Formats Guide for our internal audience (SI), documenting a handful of specific wrappers and video codecs our community encountered and what actions we recommend for their long-term care [9].

For example, AVCHD, a camera original format developed by Sony and Panasonic in 2006, stores the video content using a commonly used AVC video codec, but wrapped in proprietary MTS wrappers split into 2GB files (spanned clips). Video clips and other camera files are stored in a proprietary packaged directory structure that is only natively accessible on a Mac OS. It is a risky structure, with a good video codec in a largely inaccessible wrapper. We recommend rewrapping and combining any spanned clips but maintaining the original video codec data when possible.

P2 is a Panasonic camera original format with a proprietary and complex directory structure, separating audio and video streams into separate directories and spanning larger clips into multiple files. Since the format is highly dependent on retaining directory structure, we recommend processing the files to concatenate any spanned clips and output a master asset as close to the technical specs of the original as possible.

Other formats come off the camera as single files with common codecs and wrappers, and they are extracted from their sidecar camera files and saved as is (XAVC, XDCAM).

Considerations on how to build a model for approaching the long-term care of camera original born-digital file-based video formats has led to more questions. Normalizing files to one defined codec or format often creates larger files than the originals. Saving everything as native formats as deposited doesn’t address the inherent risks some formats embody. Stabilization and processing of specific formats requires one to address certain key technical aspects.

C. Key Aspects to Consider

The 2014 Federal Agencies Digitization Guidelines Initiative Guide to Creating and Archiving Born Digital Video is a good framing of key aspects archivists may encounter [10]. The Advice for File Archivists section lists important concepts like “Document the Original Order” (RP 2.1) [11], especially camera created file structures, and

"Identify the file characteristics at the most granular level" (RP 2.3). The recommendation to "Determine and document criteria for when (if ever) it is appropriate to change the video's technical properties (RP 2.5) asks you "Is the file "at risk" in its current form?"

Other recommendations in the guide, although noteworthy and important, are not feasible for the Smithsonian's scale and tools, for example: "Retain the original file when transcoding" (RP 2.6). Taking into account our resources, if the data is not deemed critical, we do not keep original files after processing or even "Retain all the data from the original files if the video file structure has changed" (RP 2.10). Some technical and origination data is changed and lost as a result of processing. The recommendation "Select appropriate technical characteristics for the video encoding if transcoding, normalizing or otherwise changing the video stream to meet business needs" (RP 2.7) highlights the reality that as archivists and digital preservationists, our business needs include changing files when they are deemed "at risk".

The FADGI recommendation "Determine and document criteria for when (if ever) it is appropriate to change the video file's technical properties" (RP 2.5) might be our most difficult challenge, and over the years we have been grappling with what this criteria might consist of. If we determine the files are at risk, is it irresponsible not to act, even when decisions we make create irreversible change to our collections. The process has led to the following insights in our growing efforts to build this criteria.

D. Some Insights

BUILDING A SHARED LANGUAGE

In talking with video producers at the Smithsonian, we discovered we were approaching the same questions with different experiences and assumptions. Producers referenced cameras and wrappers when discussing formats, ie the GoPro makes mov, while we were talking in the language of codecs and structure, ie this file is AVC split into multiple clips in this subdirectory. Working to build a bridge where cameras and codecs are mapped is an important step in understanding the technical aspects of the products shared between these two stakeholders. As archivists, it is our responsibility to build those relationships into a more permanent and sustainable solution.

UNDERSTANDING PRODUCTION WORKFLOWS

Starting these conversations with producers led to expeditions to production shoots. How do they choose settings on the camera? What aspects of the workflow are important for them in their work? How do they transfer from camera to computer to editing suite? Are there any data management practices in place, even if not formalized, that we can use to understand file history as they move through the production workflow and into the archive? This process is an important component of building those relationships and allows for not only a better understanding of the files but also the goals and priorities creators have for them during and after archival deposit.

DEFINING RISK TOLERANCE BY ANALYZING THE STABILITY OF FORMAT STRUCTURE

File format research for camera original video formats is not a well-developed field, and we must depend on our own examples from the cameras we encounter or from commercial sites listing out technical information. In addition to using the Library of Congress File Format Sustainability Factors [12] as a guide, we also include the structure of the camera original format to weigh stability, as seen in P2's separate audio and video sub directories [13] and AVCHD's [14], spanned clips. These camera original format structures were deemed too unstable to retain, taking into account a risk analysis matrix drawing from the Sustainability Factors, internal tool requirements, structural format dependencies, and future management and access needs.

REMEMBERING ARCHIVAL PRACTICE

We are always bound by the ethical code of our archival training, especially as cultural heritage professionals, which include concepts like original order, provenance [15], minimizing loss, and do no harm [16]. The SAA Core Values Statement and Code of Ethics lists out many bullet points, one of which is "Develop and follow professional standards that promote transparency and mitigate harm" [17]. Andrea Shahmohammadi's Smithsonian Institution Archives 2011 paper details these archival approaches with great clarity [18].

EMBRACING A NEW ARCHIVAL PERSPECTIVE – INTENTIONAL CHANGE

Staying true to archival core concepts might mean new models within the frame of digital formats

that require processing where intentional change occurs, ie not retaining original camera structures and not maintaining a copy of the original after processing. Can we define intentional irreversible change essential to stabilize these collections? We must take action that creates change that will lead to irreversible loss but essential to current stability and future access.

REQUIRING MINIMAL PROCESSING

And this leads us to the action we must take. Our format research on specific camera original formats has led us to recommend action for formats that we deem too unstable. We must minimally process these files to stabilize them according to the risks we determine they have, and we have normalized this at the Smithsonian for all of our stakeholders. Some formats are simply extracted, others are re-wrapped, some are concatenated, some are flattened and re-transcoded, some are completely changed to new formats and streams [19]. FADGI AV Group's Significant Properties for Digital Video, now in draft, serves as a guide to define the most significant technical properties to retain when migrating files with intentional loss [20].

SHARING FILE FORMAT RESEARCH

At-risk and stable and processing are all variable parameters defined according to organizational needs and tools and resources, but file formats created by video cameras and copied off card, with all their various sidecars and technical data and packaging standards, can be inventoried and documented and shared for all of us to access as a field. In looking forward, allow for a call to action to find a space to gather all of the happenings already occurring in research and practice [21].

In conclusion, the body of work that has been done in the last 10 years in building frames of thought and documentation for approaching born digital video is key to moving forward with the consideration of born digital camera original formats. It may take a new approach to video that allows archivists to face the riskiness of these formats, and it will take irreversible action towards intentional change. As Claire Fox wrote in 2020, archivist-driven research on born digital camera original formats "aims to shed light on what the ingredients of these formats are, the conditions of their creation and use, a look into historical context, and – most importantly – what responsibility

archivists have to preserve them, whether to the highest standard, or *maybe something different*" [22].

ACKNOWLEDGEMENTS

The discussion here is impossible without all of the work that has come before and all of the people working on video and digital preservation, but notably: Isabel Meyer for critical edits, support, and confidence. Dave Walker who helped fill in some research holes. Claire Fox for her careful review, edits, and excellent questions for me on expanding concepts. Taylor McBride's research and collaboration. The SI DAMS Team and all of my SI OCIO colleagues. The countless SI video creators who let me follow them around and chatted with us about cameras. The Smithsonian AV Archivists, formalized in the AVAIL group.. All our FADGI AV colleagues, led fearlessly by Kate Murray and Carl Fleischhauer. The Smithsonian Center for Folklife and Cultural Heritage Folklife Festival Documentation Team, for so much on hands experience, notably Stephanie Smith, Cecilia Peterson, Dave Walker, and Charlie Weber. Digital archivist colleagues who I call up randomly and ask codec questions. And all of the great work cited in this paper.

REFERENCES

- [1] Smithsonian DAMS. <https://si.edu/dams>. Accessed March 2022.
- [2] ProRes, as referred to here, is a family of codecs, all proprietary formats owned and developed by Apple. Apple ProRes white paper. https://www.apple.com/final-cut-pro/docs/Apple_ProRes_White_Paper.pdf
- [3] J. Volk. "Digital Preservation Workflow: Wrapper Formats",. 2009. Accessed March 2022. <https://zdoc.pub/digital-preservation-workflow-wrapper-formats-jonah-volk-dec.html>
- [4] C. Sanchez. *What's in your Repository? Facing Legacy Data in Smithsonian DAMS*. Crystal Sanchez. AMIA 2019. <http://www.amiaconference.net/poster-presentations-2/>
- [5] Adobe. *Best practices in working with native formats*. Last updated on Dec 22, 2021. Accessed March 2022. <https://helpx.adobe.com/premiere-pro/using/best-practices-formats.html>
- [6] Smithsonian Video Production Discussion Notes. Led by Crystal Sanchez. April 4 & April 11, 2018. <https://docs.google.com/document/d/1R0qk9Bi8ymQil5hiEwCYvG6w-bC1asy-K9YAJe-qnnE/edit?usp=sharing>
- [7] Born-digital camera-original video format inventory. Presented to AMIA 2018. Taylor McBride and Crystal Sanchez. https://docs.google.com/spreadsheets/d/1OvZkGkizNnx_nZ9QVDokJlIVFuIMK_7FYKC77YhoUac/http://www.amiaconference.net/wp-content/uploads/2019/01/2018-Program-Web.pdf

- [8] Harvard did this beginning in 2016 with common digital video formats to be preserved in the library's repository; not primarily file-based camera originals but serves as a nice model for this work. Accessed March 2022. <https://wiki.harvard.edu/confluence/display/digitalpreservation/Video+Formats>
<https://docs.google.com/spreadsheets/d/1rR7HNoQswcOrl66yeRRI2qMGDKzYQxitrOmD7nfVFGQ/>
- [9] *SI DAMS Supported File Formats Guide for Digital Still Images, Audio, and Video Files*. Authored by the SI DAMS Team. Published 2019. Accessed March 2022. https://www.si.edu/sites/default/files/unit/OCIO/si_dams_supported_file_formats_2019.pdf
- [10] *FADGI. Creating and Archiving Born Digital Video Part III. High Level Recommended Practices*. 2014. Accessed March 2022. Author note: she was a contributor to this report. http://www.digitizationguidelines.gov/guidelines/video_born_digital.html
- [11] Author disclosure here that the SI DAMS system does not support complex nested file structures well so initially this discussion took into consideration tool requirements but led to a more in-depth tool agnostic diagnosis.
- [12] *Sustainability of Digital Formats: Planning for Library of Congress Collections* includes the factors Disclosure, Adoption, Transparency, Self-Documentation, External Dependencies, Impact of Patents, & Technical Protection Mechanisms. Accessed March 2022. <https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml>
- [13] Panasonic. How to handle data recorded on P2 cards. Accessed March 2022. [https://pro-av.panasonic.net/manual/html/VARICAM_35\(VQT5K88A-10\(E\)\)/chapter04_04_07.htm](https://pro-av.panasonic.net/manual/html/VARICAM_35(VQT5K88A-10(E))/chapter04_04_07.htm)
- [14] AVCHD Format Specific Overview. Accessed March 2022. <http://www.avchd-info.org/format/>
- [15] Society of American Archivists. "Core Archival Functions." Accessed March 2022.
- [16] Society of American Archivists. "Preservation". Dictionary of Archives Terminology. Accessed March 2022. <https://dictionary.archivists.org/entry/preservation.html>
- [17] Society of American Archivists Core Values Statement. Accessed March 2022. <https://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics>
- [18] Andrea Shahmohammadi. *Born Digital Video Preservation: A Final Report*. 2011. Accessed March 2022. <https://siarchives.si.edu/sites/default/files/pdfs/bornDigitalVideoPreservation2011.pdf>
- [19] For more details on specific format choices, see the Video section *SI DAMS Supported File Formats Guide for Digital Still Images, Audio, and Video Files*. Authored by the SI DAMS Team. Published 2019. Accessed March 2022. https://www.si.edu/sites/default/files/unit/OCIO/si_dams_supported_file_formats_2019.pdf More work is in development to define significant properties for these specific formats for our internal audiences.
- [20] *FADGI. Significant Properties of Digital Video*. 2019. In Draft. Author note: she was a contributor to this report. <https://www.digitizationguidelines.gov/guidelines/sigpropvideo.html>
- [21] Promising work is happening with IASA's TC-07. *Guidelines for the Preservation of Born Digital Video*. <https://www.iasa-web.org/tc-07-guidelines-preservation-born-digital-video>
- [22] Claire Fox. *Not Normalized: Born-Digital Camera Original Video Formats in the Archives*. May 2020. Accessed March 2022. Emphasis mine. https://miap.hosting.nyu.edu/program/student_work/2020spring/20s_thesis_Fox_deposit_copy_v.pdf

DO WE REALLY KNOW OUR DATA?

Assessing File Format Policy Compliance and Digital Preservation Tenability via a New Software Tool

Tom J. Smyth

Library and Archives Canada

tom.smyth@bac-lac.gc.ca

[0000-0002-3829-2980](tel:0000-0002-3829-2980)

Abstract – Library and Archives Canada's (LAC) has developed a new software tool that compares the DROID report of a target dataset (e.g., material for acquisition, or material being migrated to a new media) to our newly revised File Format and Data Migration Policy. The policy now takes the form of a PRONOM-oriented Local Digital Format Registry (LDFR) database. The software tool compares the DROID report to the LDFR and outputs comprehensive file counts on policy compliance, and flags any file format non-compliance, migration, capacity or data management issues that would require elevated resources to manage. In the context of pre-transfer and PAIMAS, this enables LAC to assess preservation tenability, and estimate the investment and potential cost of ownership of any given fonds prior to transfer or making acquisition decisions. It also provides the objective data with which to negotiate with stakeholders, and/or to manage client expectations and adjust PAIMAS agreements in the light of evidence as needed. The tool is a major development for LAC, leading to better overall digital preservation, capacity management, migration, planning, program efficiency and sustainability.

Keywords – file format policy, preservation planning, risk assessment, preservation sustainability, PRONOM.

Conference Topics – Innovation; Resilience.

I. INTRODUCTION

In 2021-22, the Digital Preservation (DP) area of Library and Archives Canada (LAC) undertook a major revision of its File Format and Migration Policy. An institution's file formats policy is typically a statement of its ability and capacity to manage and

migrate data for preservation, based on the ideal file or data formats it wants to receive for particular content types. Some formats are easy for the organization to manage, while others are not for various financial and capacity reasons. By definition, this is an individualistic policy piece based on the details of an organization's mandate, perhaps its legal context, its dedicated resources, and to a degree, its DP program maturity. It also tends to be a snapshot, a moment in time of when it was written or published – which can be problematic, as such a core policy of the DP Archive evolves quickly under the lens of operations, so should be communicated to stakeholders at equal pace where possible (LAC can stand to improve on this too!)

This paper explores how LAC analyzed migration capacity and began assessing digital preservation tenability and sustainability at the point of pre-transfer. It also discusses LAC's advocacy strategy for describing capacity challenges and building objective data on which to articulate the requirements of file format compliance, data quality, and the operational impact of the costs of ownership – while delivering better and strategic digital preservation services to our clients.

II. PROBLEM STATEMENT

While LAC has maintained a paper-based policy since at least 2009, we knew we were missing a means of applying it in a machine-readable and dynamic manner to everyday digital preservation operations. Thus, our goal was to construct a method to assess any data under scrutiny for its 'compliance' with the

File Format Policy in the contexts of 1. incoming data for acquisition, 2. collections on legacy media being migrated, and perhaps also for 3. archival backlogs and re-appraisal.

In so doing, we wanted to address two main issues: 1. Creating a means of describing and mitigating digital preservation risk inherent to poor data quality or unmanageable file formats before they came in the door, and 2. Building the data on which to do DP planning, risk assessment, and migration for the DP Archive. Ultimately, we wanted to begin the process of articulating and estimating the costs of ownership for any target data – based on its degree of compliance with or variance from our File Format and Migration Policy.

Many great tools exist for file format characterization (DROID [1], Brunnhilde [2], Jhove [3], etc). However as a next step in LAC's Digital Preservation Program maturity, we found ourselves asking how an organization 1. Analyzes and understands its total and current-state capacity to migrate file formats (in terms of expertise, tools, and resources), 2. How it compares its preferably evergreen file formats policy against .e.g., incoming data for potential acquisition in an automated and machine-readable manner, and 3. How this process could assess and describe target data variance from the File Formats Policy – both in terms of its preservation tenability and the resources it is likely to absorb -- initially and in the long-term.

Put another way, how do we estimate the overall sustainability of a given acquisition, its risks and its costs of ownership – at the point of pre-transfer, with an eye to the core principles of OAIS sustainability – after we have started PAIMAS, but before an acquisition decision needs to be made? Where we know we must acquire a given collection, how can we articulate the cost of ownership in its current state, and calculate what resources and capacity will be necessary for its management? How should that information impact PAIMAS-oriented client negotiations or agreements, where applicable?

Going further, in financial, priority, and capacity management contexts, how do we build a framework that ensures we are applying our finite resources to the digital documentary heritage content that warrants high investment? How do we gather objective data with which to brief stakeholders on what the costs of ownership are likely to be? (and then use it to augment acquisition

decisions and/or build business cases for more resources and capacity?) For us at LAC, this was also the beginning of a vision for an institutional DP services catalogue, where the services could be added or removed based on available funds, capacity, and priority. Not all funds can receive a 'gold' or 'platinum' level of DP service, but how do you convince each business owner of that?

We therefore produced a methodology and tool that would generate real-time data, and enable the relevant conversations with easy-to-understand reports.

III. ANALYZING FILE FORMAT CAPACITY

An issue LAC identified over the years was how quickly our document-based File Formats Policy became outdated, and how this affected ongoing acquisition operations since clients were constantly referring to and acting upon the outdated web-based copy. DP unit needed a new means of 'keeping it evergreen', and we also wanted to make it capable of specific query (so we were not referring to a paper document on a digital matter).

To address this and to ensure the data could be easily updated in operations contexts, we generated a relational object and populated it with formats and Persistent Unique Identifiers (PUIs) based on the UK PRONOM database [4]. We also ensured the database was capable of exporting its information in an eye-readable HTML format. I have the habit of referring to this as the LAC LDFR – Local Digital Format Registry, which also tends to encompass the paper policy document.

LAC utilizes Preservica [5] as its core digital preservation module in its suite of systems and services that we refer to as the "Digital Assessment Management System" (DAMS). A next step in analyzing our institutional capacity for migration was to compile the list of file formats that were preferred for acquisition as-is (i.e., are best for LAC for their content type in the present state of 2022, based on our experiences and in comparison with the file format policies of other organizations). We then added formats to the LDFR database for which we had an existing, or a Preservica migration pathway (i.e., those formats that are not ideal, but whose migration can be automated). Next, we added formats we knew we could migrate but which were dependent on human intervention (i.e., format migration that had to be conducted manually with a

desktop application or otherwise needed DP staff expertise). The previous LDFR policy and staff knowledge informed the compilation of these latter “must be manually migrated” formats.

By process of elimination, this created a pool of remaining file formats from PRONOM or the previous policy that required elevated resources for LAC to manage. This was due to 1. An absence of known or deployed tools to handle the formats (e.g., M365!), 2. Those which had been previously flagged as demonstrating some issues (e.g., geospatial, AutoCAD), 3. Formats we knew existed from PRONOM but which have never entered our workflows (which would trigger DP analysis and policy decisions the first time they appear), or 4. Formats known to us but which are absent from the PRONOM database and thus do not have PUIDs (i.e., formats and notes we should contribute on an ongoing basis!)

By arranging these file formats into categories within the LDFR database, we can move, edit, or update their status on the fly. The LDFR thus shows at-a-glance (or through remote query) what our internal capacity is to handle any particular version of a file format – since each category reflects an increasing scale of complexity, which requires elevated resources, and must enter a particular ‘swim lane’ in the context of our DP unit operational workflows (i.e., absorbs capacity by requiring the time of specific subject matter experts).

IV. COMPARING LDFR TO TARGET DATA

Having now defined and established what our capacity is to handle individual formats via LDFR database categorization and file format characterization, we needed a means to compare it against incoming acquisitions data (or data in backlogs, or being migrated off legacy carriers), to assess the target data for risk, digital preservation tenability, capacity, likely work package cost, and thereby sustainability.

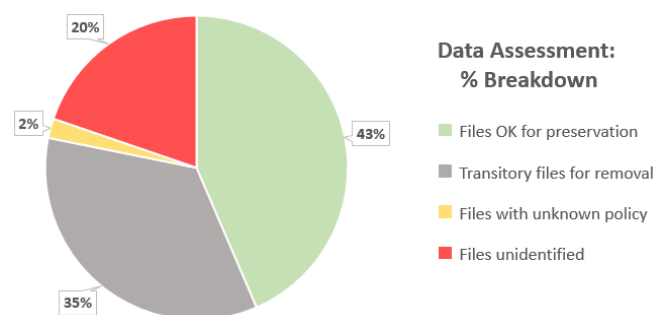
Using PowerShell, the DP area created a tool that accepts DROID reports as input (“the LDFR Tool”). Dropping a DROID report into a particular network directory would execute a script that runs a comparison between the DROID report and the LDFR database. The LDFR Tool then outputs an “LDFR Tool Report”, that arranges the data for acquisition into the LDFR’s categories by PRONOM ID (per the chart above), and graphs its degree of compliance with the

institutional File Format Policy with comprehensive file counts.

DIM

A-PREFERRED
- A0-No Action Needed
B-ACCEPTABLE AUTO MIGRATION
- B1-ACCEPTABLE / Tested Automated Migration
C-TOLERABLE MANUAL MIGRATION
- C1-TOLERABLE / Requires DP staff to Migrate
D-NOT ELIGIBLE FOR DP
- D1-Non-Archival content files
- D4-Zero Bytes Files
- D5-Else
X-ASSESSMENT REQUIRED
- X1-Has PUID - Missing in LDFR
- X1-HAS PUID / UNKNOWN requires analysis
Z-CANNOT ASSESS
- Z0-NO PUID Unknown File Format
- Z1-failed characterization (by extension)

For the first time, the digital preservation practitioners at LAC can now generate dynamic reports on the fly that are colour-coded and show at-a-glance how compliant the target data is against the File Formats Policy (LDFR) – or how problematic it will be. Where the analyzed data is non-conforming, the report articulates the total percentage of the payload, a precise file count, and paths to those objects. This makes the LDFR reports highly granular and objective (being based on PRONOM and our current DP migration capacities as reflected in LDFR!). A second, simplified version of the report is then generated for use in briefing e.g., archivists or clients about their data.



This has already triggered decisions in consultation that a high-effort but low-value component of a given dataset was not worth acquiring, or conversely that a high-value component warranted extra work.

These LDFR categories also depict the relative degree of effort, cost and/or accepted or inherent risk required to manage the tagged formats (“Unidentified” files require a DP practitioner’s time to analyze, whereas files that are “OK for preservation” can flow through the system in an automated fashion). In this way, we can show the business client that it is in everyone’s best interest to minimize costs and processing effort to the institution via pre-transfer LDFR Tool assessment, adjusting PAIMAS agreements, and engaging the donor to ensure file format policy compliance prior to transferring and approving an acquisition.

V. DEPLOYING THE TOOL AND ASSESSING SUSTAINABILITY

The LDFR Tool is intended for the DP practitioners at LAC, primarily the DP unit and the “Digital Integration” team (responsible for digital archival records, PAIMAS, and the stages of the DP workflow up to and including the generation of the Submission Information Packages (SIPs)). At the time of writing, the LDFR Tool has been deployed to these two areas and is now in everyday operational use.

The LDFR categories will eventually be linked to different levels of treatment in our institutional digital preservation service catalogue¹, which in turn reflect an increasing scale of required effort and duration in processing. Since incoming data that is highly non-compliant with the institutional file format policy will require more FTE effort from the DP practitioners (due to being unmanageable with automated workflows), acquiring such fonds will cost much greater annual, operational DP capacity than those that are highly compliant (and can be automated). This makes incoming data quality and compliance everyone’s concern, in order to maximize the rate of data and fonds acquisition, and minimize institutional risks and the financial and capacity impact on the DP area and its staff.

Rolling all these details up into a briefing for a given client for a potential acquisition (or a particular digital collection in need of management) would culminate in a designation that their collection will require a “Bronze, Silver Gold, or Platinum” level of DP treatment and the use of XYZ services from the DP catalogue. A service for manually migrating thousands of files might be “Gold” level DP

treatment, as this would obviously take a great deal more capacity, potential cost, and time than automated migration. Thus, based on the LDFR Tool’s findings and output report, negotiation can then occur with the client on what DP services to add or subtract, given client budgets, goals, and DP area available capacity. The LDFR Tool’s output report also thereby shows clients exactly what activities are indicated for their data, and precisely *how their data will be managed* for processing and preservation at LAC – which justifies required costs, time, duration and capacity.

Where the implicated collection is at the pre-transfer stage, this information might influence the decision on whether to acquire that fonds at all (e.g., where cost versus value is low), could contribute a great deal to the PAIMAS-oriented pre-transfer submission agreement, or trigger a negotiation and collaboration on how to improve data quality prior to transfer. In one specific case, this led to a client delaying transfer to enable combined resources to be applied to the unanticipated but essential DP work.

The LDFR Tool would also be used in the near future on all data being migrated to new DP archival media. Running the contents of an LTO tape through the LDFR Tool will describe exactly what data by PUID is on each of our 10,000+ magnetic tapes, leading to a risk-managed, order of priority for migration from LTO4-6 to LTO-9 in our Tape Library. We would then capture these details in our DP master database (where all our digital objects in our archive are logged), which in turn provides the raw data on which to base digital preservation planning, while globally managing the Canadian national digital collections.

VI. LESSONS LEARNED

The DP area will begin including the DROID and LDFR Tool reports in future Archival Information Packages (AIPs), thereby transferring the file format analysis, capacity context, and file format migration decisions forward into our DP archive master database – as additional context and provenance information, that may elucidate preservation decisions made now for the benefit of our successors in the future.

The File Format Policy will now remain evergreen, since the full content of the LDFR database can be adjusted on demand, but also exported in easily-

¹ Our vision for a DP services catalogue was inspired by the NDSA Levels of Digital Preservation Treatment.

readable document format. Thus, we can provide a fresh copy whenever clients and other departments in the Government of Canada request. While legacy data is legacy data and a File Formats Policy will not always be consulted, applying the thinking as far upstream as possible is our goal, to enable Government of Canada clients to conduct self-assessment, leading to early analysis and warning for future federal record transfers.

Given the framework that is now in place, we can also begin building the data necessary to estimate our bandwidth requirements for the use of Preservica's pathways for performing migration work in the Cloud.

Not all collections were created equal in terms of data quality or sustainability, and so not all of them can objectively warrant high investment or can support advanced access options (emulation, and so on). Capacity and resources are always at a premium. This is perhaps not an easy thing to socialize, even when the digital preservation practitioner deeply understands and shares the collection specialist's sense of intellectual responsibility for acquiring historically important but not always great quality data.

To ensure digital preservation tenability and sustainability, it is essential to bring business partners on board with core international digital preservation principles. Digital advocacy is indicated for securing the needed resources to address the prescribed and required DP work, and to improve pre-transfer and pre-acquisition conditions, leading to greater efficiency, capacity and improved (data) quality of life for the collections. If we can improve the state of data prior to transfers via this characterization methodology, we can avoid future surprises and impact on the institution's reputation years down the road. For these reasons, I thought this work was important to publish, as this approach would be easy to re-purpose anywhere. Our LDFR Tool is driven by DROID and PRONOM, so it is simple to update and roll-out.

Acquiring data we cannot adequately preserve or provide access to is contrary to the basic tenets of digital preservation. My vision would be for this work to enable formulaically calculated effort, duration, and costs for projects and communication to stakeholders, which in turn also assists in DP capacity

and migration planning and increased DP program efficiency and maturity.

ACKNOWLEDGMENT

The highest praise is due to the staff comprising the Digital Integration unit at LAC. I remain humbled and privileged to work with you, as has been the case since 2013.

The software development work underpinning this paper was conducted in tandem with Maxime Champagne and the LAC Digital Preservation unit, who do a great deal of heavy lifting on a daily basis. I am honored to have worked with you.

REFERENCES

- [1] The National Archives. *Digital Record Object Identification Software Tool* (DROID). <https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/>
- [2] Canadian Centre for Architecture and Tessa Walsh. *Brunnhilde: Siegfried-based characterization tool for directories and disk images*. <https://github.com/tw4l/brunnhilde>
- [3] Harvard University and Open Preservation Foundation. *JHOVE Harvard Object Validation Environment*. <https://jhove.openpreservation.org/#:~:text=JHOVE%20is%20a%20file%20format,A%20command%20line%20interface>
- [4] The National Archives. *PRONOM*. <https://www.nationalarchives.gov.uk/PRONOM/Default.aspx>
- [5] National Digital Stewardship Alliance (NDSA). *Levels of Digital Preservation*. <https://ndsa.org/publications/levels-of-digital-preservation/>

THE 2022 REVISION OF THE PREMIS RIGHTS ENTITY

Karin Bredenberg

*Kommunalförbundet
Sydarkivera
Sweden*

karin.bredenberg@sydarkivera.se

[0000-0003-1627-2361](tel:0000-0003-1627-2361)

Bertrand Caron

*Bibliothèque nationale
de France
France*

bertrand.caron@bnf.fr

[0000-0002-3433-9518](tel:0000-0002-3433-9518)

Leslie Johnston

*National Archives and
Records Administration (NARA)
United States of America
leslie.johnston@nara.gov*

[0000-0001-9908-0183](tel:0000-0001-9908-0183)

Michelle Lindlar

*TIB Leibniz Information Center
for Science and Technology
Germany*

michelle.lindlar@tib.eu

[0000-0003-3709-5608](tel:0000-0003-3709-5608)

Jack O'Sullivan

*Preservica
UK*

jack.osullivan@preservica.com

[0000-0002-0306-761X](tel:0000-0002-0306-761X)

Sarah Romkey

*Artefactual Systems
Canada*

sromkey@artefactual.com

[0000-0003-3833-7648](tel:0000-0003-3833-7648)

Marjolein Steeman

*Netherlands Institute for Sound and Vision
the Netherlands*

msteeman@beeldengeluid.nl

[0000-0002-1506-1581](tel:0000-0002-1506-1581)

Abstract – As digital preservation practice has matured, our understanding of what is covered by the Rights landscape broadened significantly. However, the Rights entity in the PREMIS data model has not kept pace with these changes, undergoing only minor revisions. In 2019, the PREMIS Editorial Committee formed a working group on Rights to review how the entity could be changed to better reflect the evolving use cases of our community. The initial phase of this work has involved gathering use cases in order to perform a gap analysis with the current definition of Rights within the PREMIS model. Ahead of an official White Paper describing the scope of the revisions to be considered, this paper presents a discussion of the use cases that have informed this work, and the gaps we have identified, before briefly outlining the next steps to be taken.

Keywords – PREMIS, Rights, Datamodel

Conference Topics – Community

I. INTRODUCTION AND BACKGROUND

With the revision in version 3.0 in 2015 [1], Rights in PREMIS underwent a minor revision to align it more with community needs. During the work to create the PREMIS OWL Ontology [2], which was

released in 2018, the Editorial Committee determined that additional work was needed for the Rights part of PREMIS to better serve the community needs.

The Rights entity in version 2 was designed to support the “assertion” of the rights basis for the repository to perform preservation actions, and therefore offers 4 mutually exclusive options for a *rightsBasis*, in combination with a semantic unit to define the actions for which rights have been granted (possibly with restrictions). As digital preservation practice has matured over time, the usage of Rights has broadened.

Within digital archives it is often necessary to capture many different types of rights as well as their changes over time. In addition to the right to preserve an object, aspects such as copyrights and usage rights form relevant information for OAIS-based processes. Knowledge such as “am I allowed to migrate this object” needs to be connected to each object in an archive in order to allow for automated preservation processes. In support of provenance and chain of custody it also becomes essential to

capture the history of rights information, and there may be a need to document rights to perform actions that are not strictly defined as preservation. An example of this is access to archived objects on the basis of pre-defined trigger events. Here, the archive might be obliged to prove during which time a trigger event took place and that access to an object was granted under this right. There is also a need to cover not only granted rights but also obligations that the repository is committed to.

As seen in these examples, the rights landscape for digital objects is not a simple one. While PREMIS allows for a connection of rights, agents and events to objects, questions have been raised about whether the data model sufficiently supports the complexity of objects which are subject to multiple restrictions and rules, more than one rights holder, internal users that have different roles, uncertain expiration dates, a range of different trigger events, documenting the outcomes of rights reviews, and so on.

In September 2019, a working group was formed to address the issue of Rights in PREMIS. The first assignment for the working group was to define the scope of the project and the issues to be resolved. For this, the group collected use cases from their respective institutions. These use cases form the basis of current work and are presented in the next section before briefly concluding this short paper with an outlook on further work. The use cases articulate new assertions that a future revision of the Rights Entity would express more precisely. These will be discussed within the PREMIS Editorial Committee and with the digital preservation community on whether they have to be adopted within the scope of a future version of PREMIS.

II. USE CASES GROUP 1: DOCUMENTING OBLIGATIONS AND 'TARGET' AGENTS

The first version of PREMIS limited the Rights Entity to expressing **permissions held by an Agent and granted to the digital preservation system**. Thus the only two different assertions that the Rights Entity allowed in PREMIS 1 were:

“Agent A holds this right to Object B” and
“Agent A grants [the repository] this permission related to Object B.” [3].

PREMIS 2 introduced the option to define prohibition as a restriction in relation to a granted

right, e.g., an embargo is seen as a restriction on dissemination for X years before it is allowed. Nevertheless, restrictions were limited to a free-text, human-readable description in a restriction semantic unit. Moreover, the term “restriction” seems very specific for some use cases where a condition has to be in place in order for the repository to be allowed to do something.

The following examples illustrate the needs of some users to express:

- An “obligated condition” or in other words an obligation;
- The ‘target’ Agent of a Rights rule (to whom the rule applies). In particular, the Agent would not always be implicitly the repository. It could be a human or software Agent. As such it might be part of the repository, but doesn’t have to be;
- The restrictions / conditions of a rule in a machine-actionable way, be it a permission or an obligation; possibly in a metadata schema other than PREMIS.

A. Example 1: Service Level Agreements Between Producer and Archive

BnF and its Producers negotiate service level agreements, defining among other things the storage media type, the transfer mode, the SIP’s maximum size, etc. These commitments can be defined in terms of permissions (ingest) under a certain restriction (e.g. only to submit packages from Monday to Friday, 8 am to 8 pm). Part of the agreement may well be certain overall obligations (e.g., to perform yearly fixity audits). The PREMIS schema should in this case not only support restrictions, but also obligations that go with the granted right.

To support this example, PREMIS should be able to express with sufficient precision the following assertions (non-comprehensive list):

- SLAs (*rightsStatement*), based on a contract (*rightsBasis*), allow Agents P (‘target’ Agents) to transfer Objects [to Agent Z] provided that the maximum number of files is N files, whose format should be either F, G or H.
- SLAs (*rightsStatement*) obliges Agent Z (‘target Agent’) to perform preservation actions such as annual audits, characterization and filename change, on Object O, and to store three copies of Object O, one of them on

disk, and the two others on magnetic tape.

B. Example 2: Format and Tool Policies

BnF needs to specify the relationship between the format and its associated tools (characterization, validation, rendering, migration, etc.). Describing and publishing its format policies (e.g., this Object must be analyzed by this software Agent, migrated by this other tool, etc.) would help sharing and discussing good practices of preservation operations. This provides another example of obligations in relation to the permission to a preservation action on a certain format.

Note that there is currently no way to specify a relationship between the Object format and the Rights Entity in PREMIS. This means that the obligation to perform an action on a certain Object cannot be related in a standard way to the fact that the Object is of format X.

To support this example, PREMIS should be able to express with sufficient precision the following assertions:

- Format policies (*rightsStatement*) based on an institutional policy (*rightsBasis*) obliges Agent A to perform the following actions:
 - Characterizing Object O' with tool T [because of its format I];
 - Normalizing Object O' with tool U [because of its format I];
 - Validating Object O' with tool V [because of its format I].

C. Example 3: Access and Use Embargo

At the U.S. National Archives (NARA), a "72-Year Rule" restricts access to United States decennial census records to all but the individual named on the record or their legal heir. Congressional/ Legislative records can have 20, 30, and 50-year embargos. The repository requires item-level metadata indicating these access and use restrictions relating to permanent records based on federal statutes.

This example shows the impact of extending the scope of the Rights entity from granting a preservation action to giving access for reuse. To support this example, PREMIS should be able to express with sufficient precision the following assertions:

- Access rules (*rightsStatement*) based on a statute (*rightsBasis*) grants the repository to perform the following actions:
 - Give access to the census record to any other individual provided (*Restriction*) the expiration of the 20/30/50/72-year embargo periods;
 - Give access to the census record to the individual Q named in the record ('target' Agents).

III. USE CASES GROUP 2: DIFFERENT LEVELS OF CONDITIONAL RIGHTS

Rights can be tied to specific temporal or spatial conditions. And at a certain point in time separate *rightsStatements* may have a different outcome on whether an action is granted or prohibited. These different *rightsStatements* have to be assessed on their mutual outcome. And this may need to be re-assessed whenever a change in conditions is triggered.

The following examples illustrate the needs of some users to express:

- the order in a stack of Rights;
- the way a change in conditions can cause one rule to overtake another rule.

A. Example 1: License to lift copyright, given a specific time or context

At the Netherlands Institute for Sound and Vision, the audio-visual collection is largely copyrighted, and the rights have to be cleared by the producer (often a broadcast company) before giving access for re-use to the materials. In that case the broadcast company provides a license for re-use for a specific period in time in a specific context.

To support this example, PREMIS should be able to express with sufficient precision the following assertions:

- Copyright law (*rightsBasis*) prohibits any representation, reuse and copy of Object O.
- A license (*rightsBasis*) allows re-use under some circumstances and takes precedence over the copyright law.

B. Example 2: Privacy Rights

At the Netherlands Institute for Sound and Vision, regardless of the fact that most material is

protected under copyright, Dutch national legislation has given the repository the rights basis for showing the material to the public on the premises of the repository. It has therefore created a public museum. Even so, all programs that are shown in the museum need to be screened on possible privacy and ethical issues. This is to prevent claims from persons that are being depicted on the program and may in some way be offended by it. This adds a fourth level of rights, adding up: copyright, licensing, statute and privacy (other rights).

In addition to the assertions mentioned in III. A. above, PREMIS should be able to express with sufficient precision the following assertions:

- National legislation (statute *rightsBasis*) allows the Archive to visualize Object O in the institution's premises (condition).
- The national legislation takes precedence over the copyright law.
- Privacy laws and institutional policy (statute *rightsBasis*) prevent the Archive to visualize Object O. and takes precedence over any other applicable rights basis.

C. Example 3: Changing rights on trigger

TIB - as well as many other institutions or services who archive materials like e-journals - gives access to some content only when specific conditions are met. The submission agreement and legal contract between the digital archive (in this use case: TIB) and the depositor (in this use case: a publisher) define trigger events which need to be met in order for the materials to be made available under different usage rights. An example can be a publication that is archived under all-rights reserved. If the publisher's website, through which subscription based access is usually granted, becomes unavailable for e.g. 90 days, TIB has the right to make the content available through its own website. In some cases content may only be triggered for a specific period (e.g., while the publishing website is down for a specific duration). Each trigger event is tied to a set of rights which are either active or inactive based on whether conditions connected to the rights are met or not. It is the digital archive's responsibility to keep an audit trail of which right was active during which time periods and for what reason - this includes past and present rights as well as the documentation of future rights in their connection to trigger events.

To support this example, PREMIS should be able to express with sufficient precision the following assertions:

- Usage right 1 (*rightsStatement*) based on contractual agreement with publisher (license *rightsBasis*) prevents the Repository from publishing Object O;
- Usage right 2 (*rightsStatement*) based on a contractual agreement with publisher (license *rightsBasis*) goes into effect when trigger event conditions (*Restrictions*) are met. It allows the repository to publish Object O as long as the condition is met;
- The history right 1 and 2 being effective over time is been documented;
- An Event of type 'unavailability report' is recorded which relates to rights 1 and 2.

IV. USE CASES GROUP 3: REVIEWING RIGHTS STATUS AND DOCUMENTING OUTCOMES

Cultural heritage institutions are familiar with restrictions put in place by contract (donor agreements) or under Copyright protections, but restrictions may also be put in place by executive order, legislation, government regulations, or security classification. These can affect not just access and use, but the ability for staff to view objects and take preservation actions on them.

Regardless of the source of the restrictions, those restrictions must be reviewed, documenting the results of those restrictions to make them machine-actionable for integration into preservation systems. PREMIS should be able to link an Object to multiple review Events, each with a review type, a review date, and properties that specify what triggered the review, the outcome of the review, and which rights have been changed and applied to all or part, which also link to rights statement in the Rights entity. The rights basis may reflect multiple statutes or classifications that apply simultaneously to all or part of the content; reviews correspond to the restrictions put in place by those statutes or classification. Review outcomes may refer to subsections of an object, and the review events will accrue over time, resulting in a situation where the current status must be programmatically determined.

This may also potentially apply to the review of Copyright or contractually restricted objects, although such objects are less likely to have multiple

reviews or partial restrictions. These use cases have a strong dependency on the use cases on accumulating rights.

A. Example 1: Restriction by Regulation

Local, State, or Federal government records may have restrictions put in place by executive order, legislation, or regulation, where the rights basis is Statute. This is not limited to access rights, but can also apply to the staff of the custodial organization, limiting the staff who are allowed to perform preservation actions or even view the records. These are restrictions on preservation storage (restricted, secured servers), as well as restrictions on staff who may interact with the records, which are in force simultaneously with the granting of preservation action rights in rights granted. The U.S. National Archives (NARA) must enforce restrictions for the records in its holdings that can fall under several simultaneous statutes, and may be reviewed for release several times.

To support this example, PREMIS should be able to express with sufficient precision the following assertion:

- Access right 1 (*rightsStatement*) granting access to Object O, to user U ('target' Agent), provided user U has role K

B. Example 2: Restriction by Security Classification

At The U.S. National Archives (NARA), an object that is restricted due to national security concerns can have both Classified and Unclassified content at the same time, with the status applying to different sections of the object. The classification may also be assigned by multiple agencies as Agents. When the NARA reviews objects for declassification and release, they may be released in part or in full, or not released at all. Redactions to the objects when released are marked with a Redaction Code: additional metadata that identifies the statute(s) under which the information remains redacted. Multiple review Events will be performed over time, the history of which must be retained.

To support this example, PREMIS should be able to express with sufficient precision the following assertion:

- Multiple Rights Entities, each representing a classification (*rightsStatement*) and linking the agency M that has imposed it;
- An Event of type 'review classification' is recorded for Object O, specifying the classifications that have been under review;
- The outcome of the review for Object O is documented with dates and restrictions that are fully or partially lifted.

V. OUTLOOK AND FURTHER WORK

The next step for this working group will be the publication of a draft White Paper accompanied by a request for feedback from the community. This will help us to define the final scope of the project.

All proposals and changes for the Data Dictionary within scope will be considered by the PREMIS Rights Working Group. The Working Group will put forward the proposals to the Editorial Committee for review of major decisions. These will be published and open for comments

After closing the community review, all comments will be evaluated and a final proposal will be presented to the Editorial Committee.

ACKNOWLEDGMENT

The authors of this paper would like to thank their respective organizations for the use cases described within this paper. We also want to thank PREMIS Editorial Committee member Rebecca Gunther for her contributions to the work with rights since PREMIS was created. In addition, we would like to thank former PREMIS Editorial Committee member Kevin DeVorse of NARA, who was responsible for the development of the 72-year Use Case, as well as Richard Dancy of Simon Fraser University Archives for contributing to the rights review use case.

REFERENCES

- [1] PREMIS Editorial Committee. PREMIS Data Dictionary for Preservation Metadata v3.0. June 2015. <http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf>
- [2] PREMIS Editorial Committee. PREMIS OWL Ontology Version 3.0.0. October 2018. <https://id.loc.gov/ontologies/premis-3-0-0.html>
- [3] PREMIS Editorial Committee. PREMIS Data Dictionary for Preservation Metadata v. 1.0, May 2005, p. 1-7. <https://www.loc.gov/standards/premis/v1/premis-dd 1.0 2005 May.pdf#page=17>.

"...PROVIDE A LASTING LEGACY FOR GLASGOW AND THE NATION"

Two years of transferring Scottish Cabinet records to National Records of Scotland

Garth Stewart

National Records of Scotland
UK
Garth.stewart@nrscotland.gov.uk

Abstract – This paper outlines the recent transfer of a series of high-level Scottish Government records to National Records of Scotland (NRS). The records needed to be made publicly available almost as soon as they arrived at our archive, and the transfers prompted NRS to consider how we could provide online access to born-digital records for two distinct use cases. This paper outlines; the background to this task, how NRS evaluated the solutions that could be used to deliver online access, and lessons learned from an enterprise carried out exclusively within the challenging context of the Covid-19 pandemic

Keywords – Transfer, Access, Re-use of technologies, collaboration, Public records

Conference Topics – Innovation; Resilience

Bid (NRS Reference SCR14/26/5) available at
<https://www.scotlandsppeople.gov.uk/>

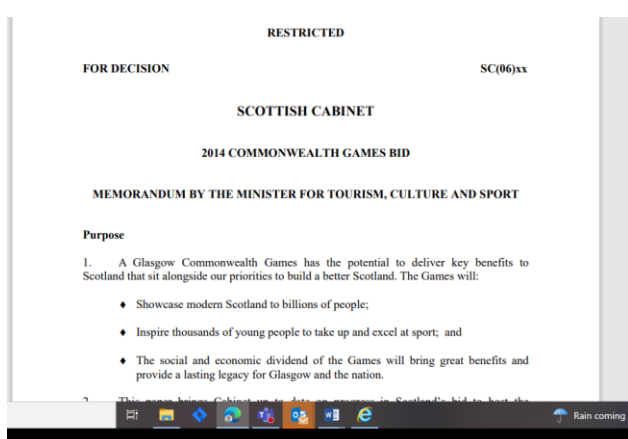
I. INTRODUCTION

This paper outlines the recent transfer of a series of high-level Scottish Government records to National Records of Scotland (NRS). The records needed to be made publicly available almost as soon as they arrived at our archive, and the transfers prompted NRS to consider how we could provide online access to born-digital records for two distinct use cases. This paper outlines; the background to this task, how NRS evaluated the solutions that could be used to deliver online access, and lessons learned from an enterprise carried out exclusively within the challenging context of the Covid-19 pandemic.

II. WHAT IS NRS?

NRS is a non-Ministerial Department of the Scottish Government (SG) - a devolved administration of the UK. NRS carries out a range of nationally-important statutory functions¹, including that it holds, preserves and makes available the national archive collection of Scotland.²

NRS has a wide and highly diverse range of depositors; from the SG and numerous public



Extract from records of the Scottish Cabinet, Minister for Tourism, Culture and Sport, Draft: 2014 Commonwealth Games

¹ For a full range see <https://www.nrscotland.gov.uk/about-us/what-we-do>

² For further information see: <https://www.nrscotland.gov.uk/research/research-guides/historical-records-an-overview>

authorities, to the Scottish Courts and a large array of private depositors.

NRS's archiving function extends to all record formats – from parchment to born-digital. NRS has been collecting born-digital records since 1998, and we have progressively developed policies and processes which combine the corporate tools and services at our disposal with the very latest in digital preservation good practice.

In the year that iPres comes to Scotland, the paper highlights the themes of community, innovation and resilience, which were paramount to delivering brand-new forms of access to one of Scotland's national collections.

III. DIGITAL ARCHIVING AT NRS

Our process for handing born-digital records is straightforward: our Archive Depositor Liaison (ADL) team work with the depositor ahead of transfer to agree precisely which records are to be deposited, and prepare these for transfer. Often colleagues from the Digital Records Unit (DRU) support this engagement and advise on format, metadata, export etc. Sensitivity review pre-transfer is undertaken by the depositor, who remains the owner and Data Controller for the records. NRS may ask for test data to be submitted to clarify technical matters, after which a 'transfer package' – records, manifest, and associated metadata - is transferred on encrypted USB Hard Drive. For further information see our [Depositor Guidance for the Transfer of Archival Born Digital Records](#).

Once at NRS, the transfer package is scanned for malware, and undergoes a series of ingest processes (fixity, characterisation, completeness checks)³, before being uploaded to the NRS Digital Repository. Until recently, all of these processes took place onsite.

Storage for the Digital Repository is built on existing NRS ICT infrastructure, ensuring we have a high level of support to maintain our security, maintenance, and storage capacity needs. Multiple

copies of records are kept on different media types and geographical locations.

Generally digital records are only catalogued to accession package level, however we are starting to catalogue more at a more granular level. Until recently, access to digital records could only be provided on-site on a standalone PC, or via provision of copies.

IV. CHALLENGING OUR PRACTICE: SCOTTISH CABINET RECORDS

Two recent transfers of significant government records challenged our access procedures.

The Scottish Cabinet (SCab) is the group of senior Ministers, including the First Minister of Scotland, which is responsible for SG policy. It came into existence in 1999, following the establishment of the Scottish Parliament and the devolved Scottish administration. The records document senior governmental decisions and policy matters in Scotland; from infrastructure and social policies, to global summits and health services.



Photograph of the 2005 Scottish Cabinet with the SG Permanent Secretary inside Bute House, Edinburgh

Crown copyright, National Records of Scotland, SCR14/6

To support government transparency and openness, SG records – including SCab - are transferred to NRS after 15 years⁴ – at which point they fall 'open' and are made available for public access by NRS, unless exempt under freedom of information legislation.⁵ Records of the SCab have been created and held entirely digitally from 2005, with records from that year scheduled for transfer to NRS in late 2020.

³ DROID, CSV Validator and Teracopy are now used for these processes

⁴ Reduced from 30 years in Scotland in 2009

⁵ The Freedom of Information (Scotland) Act 2002 (Historical Periods) Order 2013 (FOISA) confirmed that records become 'historical' after 15 years. This is slightly earlier than other UK administrations.

For at least 40 years, NRS has provided a 'media preview' to government records (including SCab), whereby journalists are invited to see selected records ahead of their public release date (normally on or around 1st January). Until December 2019, journalists attended our search rooms in-person to see/preview paper records, and the public could consult the paper files from the following January onwards.

In 2020 – and again in 2021 – this solution became untenable, due to Covid-19 restrictions. Nevertheless, the 2005 and 2006 records still had to be transferred and made publicly available – this time *online*. The show must go on, but how? The fact that the pandemic coincided with the first tranche of fully born-digital SCab files meant we *had* to undertake almost this entire operation remotely.

1. Part 1: Preservation

Transfer and ingest

The selection, transfer and ingest of the 2005 and 2006 SCab records went reasonably smoothly. SG uses Objective ECM as its main records management system, and all SCab records were drawn from this source. NRS also use this system for records management: this familiarity helped us understand the records we received.

SG carried out sensitivity review of records pre-transfer, leading to the generation of redacted sets for public access, and un-redacted sets (which remain 'Closed' until FOISA exemptions lapse). The files were exported from eRDM, and the export included manifest CSV files containing essential metadata such as original eRDM locations, Object IDs, file sizes and checksums.

Once transferred, ingest into the NRS Digital Repository was straightforward. Both record sets were successfully scanned for malware, profiled using DROID, and verified upon transfer using checksums contained in the manifest files.

'You may have forgotten to attach a file'

Email preservation presented a specific challenge. Upon comparing the 2006 record sets, we noticed that an identical number of records was contained within the 'Open' and 'Closed' sets. This was in contrast to 2005, where individual email attachments had been taken *out* of emails, redacted and saved as

separate objects. For 2006, it was evident that redacted attachments had been *added back into* the email files.

The Open record set contained 80 Outlook email files (x-fmt/430) which, in effect, were acting as container files for 265 attachment files (mainly Word document and pdf)! Future work will be undertaken to manage this preservation risk.

2. Part 2: Access

Unlocking access

Approaching transfer day in late 2020, NRS had *no* obvious means of providing online access to journalists or the wider public. In Autumn 2020 a project team was established to tackle this challenge. In both cases, we needed solutions which would be cheap (or free!) to procure, quick to install (given our deadlines), and easy for users to operate. In the case of the journalists, we also required a system which could be more strictly controlled, so that copies of records were not shared ahead of their release date.

NRS is a large government organisation with numerous statutory obligations and processes, and it has a large amount of digital assets, licences and services in operation at any one time. The project team took advantage of this by evaluating and creatively engineering the 're-use' of existing tools for born-digital access, rather than buying anything 'new'.

V. MEDIA PREVIEW

For journalists, a solution which would render a copy of the records in their original formats as far as possible was required. Objective Connect (OC) – a secure external file sharing application licensed for use by SG and NRS – was selected given its supporting features:

- a 'private' virtual reading-room could be created on OC for users, who would register/log-in to review content
- a folder structure could be created to replicate the original storage hierarchy of the records. Folders were labelled with NRS archive references and descriptions to facilitate browsing
- Each folder and document has a unique url, meaning that journalists could be provided with a hyperlinked index allowing them to

click to directly access particularly significant records identified by NRS

- Once within the reading room, documents could be accessed concurrently, and users would be unable to identify others 'in the room'. This supported the control and confidentiality of the process
- Access to every document could be tracked and audited by NRS
- Digital objects would be 'viewed' (referred to as 'Preview Mode') rather than downloaded by users: this mitigated the risk of copies being shared improperly before public release
- In practical terms, this was a familiar tool for NRS/SG colleagues, with no need for further investment, licensing or training

This solution worked very well – journalists received registration instructions in advance, as well as contextual information about the records, a detailed catalogue to enable browsing, and an extensive Index with hyperlinks to significant documents. An unexpected benefit of using OC was access audit trail functionality – this provided archive colleagues with data on what records were viewed the most and for how long – analytical data impractical to obtain in the paper world.

One issue did relate to our old friend, email files. OC's Preview Mode did not allow for embedded files (e.g. email attachments) to be opened. To get around this, NRS migrated copies of the email files from Outlook into eml format and used a simple Linux utility called '[munpack](#)' to extract attachments. This enabled us to create a folder structure for each email file so that the original email files could be reviewed alongside the attachments. This was a significant lesson learned, and for future accessions we intend to conduct this attachment extraction at point of receipt.



Munpack image taken from
<https://www.youtube.com/watch?v=SKkvb5jF0Qs>
Copyright Kris Occhipinti

VI. PUBLIC ACCESS

If delivering one born-digital access solution wasn't hard enough, how about designing a second for the wider public?!

OC would not suffice for this, given its limited scalability, but a parallel NRS project provided an option. The [ScotlandsPeople](#) website is NRS's primary platform for sharing digitised images of our collections. Used mostly for genealogical research, users can search record indexes for free (e.g. births marriages and deaths statutory registers, wills, valuation rolls etc.) and use a pay-per-view service to download images of relevant records.

During the early stages of the pandemic NRS was considering how to share more records online, via a new component on ScotlandsPeople called 'Virtual Volumes Online' (VVO). A large series of digitised church records were being prepared for release in early 2021, and this platform seemed suitable for SCab records too, given that it was scalable, secure, and soon to become part of the recognised ScotlandsPeople service offer. Again, no additional licensing would be required and this was selected for our use case.

As with any prototypical service, there were some snags! Unlike OC – which could render most formats - VVO had been set up to host image formats only: JPEG and TIFF. This configuration would present issues when converting formats such as email and Powerpoint into images for access. An alternative 'render' format for VVO was required. PDF was chosen, for a number of business reasons:

- Web-friendly: generally PDFs would be smaller files than the originals so would be

easier to download and make available on the Web

- Information security: As PDFs are a fairly simple file format, they had a lower risk of containing or hiding 'nasties' such as malware
- Information fixity: NRS wanted to 'fix' the content of records as much as possible. We didn't want to make content available that would be easy to alter and pass off as something that it wasn't. PDF offered this aspect
- Accessibility and obsolescence: As the original records would have been created in a 15+ year old version of MS Office, there could be greater format obsolescence risk. PDFs would be more backwards compatible for people to read.
- Efficiency and management: generating and tagging the metadata of JPEG images for each 'page' of transferred records would have been highly laborious. Using PDF simplified this process, whereby multiple documents could be appended together with the same metadata – ensuring provenance and original order was maintained - and presented as larger files e.g. one PDF per meeting of the Scottish Cabinet with all related documentation combined (be it Word, powerpoint, Excel etc.).

Adobe Acrobat was used to do the bulk of conversion (with licenses borrowed from another NRS team), and generally this pre-access processing approach worked. Such is life, there were still a few technical issues to troubleshoot.

VII. TROUBLESHOOTING

Email files would not convert to PDF via Adobe Acrobat. Instead our IT colleagues used, BitRecover PST to PDF Wizard, to successfully convert all Outlook files into PDF. Pleasingly, whilst IT colleagues pioneered the use of this tool for the 2005 records, they were able to pass on this learning to archive staff, who successfully converted 2006 emails using the same method.

It was difficult on some occasions to convert information stored in Excel files into PDF – if this presents an issue to users, we will provide a copy of the original Excel file via a suitable transfer mechanism.

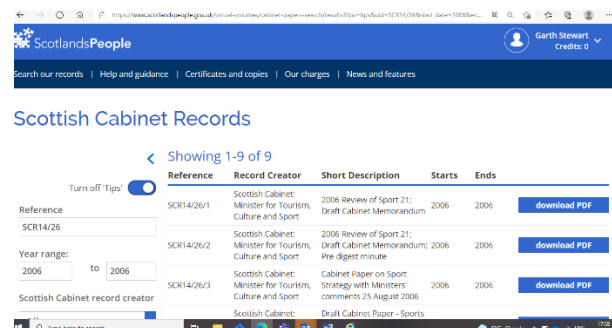
File names also proved quite painful: all files names needed to have an NRS reference added to them, in order to be uploaded to VVO. To do this at scale, the team used a Powershell command line script, which proved effective.

However, we did encounter issues likely familiar to digital preservation colleagues:

- File names sometimes exported in different orders depending on what tool was used for this (e.g. Windows file explorer, Powershell).
- Some file names were also uncomfortably long for our systems to cope with before they were converted to PDF, which obstructed file conversion.
- We also needed to ensure that the metadata populated for the access copy of the files corresponded with how they were catalogued within our central catalogue

Fixing these issues required a lot of data merging and time-consuming manual checks.

Lastly, in order to publish these records on ScotlandsPeople, new web pages, search functionality, and browsing features needed to be incorporated into the site by our third party provider. This meant that the 2005 records were not made fully available until October 2021, although the 2006 set were successfully published on 6 January 2022.



How the SCab papers are presented on ScotlandsPeople

VIII. CONCLUSIONS, LESSONS, AND THE FUTURE

Preparing these born-digital records for access was a fairly herculean effort. NRS's examples may not be suitable or feasible for other archive services who similarly need to develop online access to born-digital records from scratch.

What the examples do show however is what can be achieved in a short period of time, with the right people 'in the room', and with a defined common goal in mind. All collaboration needed for this enterprise – from initial conversations to project planning, to format conversion and record upload – took place entirely virtually – whether on MS Teams or in one of NRS's virtualised desktop environments (Adobe Acrobat and Powershell tasks, for example, were conducted in virtualised environments). The project was also delivered 'on a shoestring budget', and was supported by identifying alternative uses for corporate IT tools and services.

NRS has a long way to go in its digital archiving journey. We will need to take stock of the ScotlandsPeople platform (especially its 'normalisation' workflow) and receive user feedback on its strengths and limitations. The processing work required to process the records was also beset by particular inefficiencies, which we will work to overcome.

One thing is certain: NRS will continue to receive born-digital records, and be expected to provide public access to these – starting with the 2007 SCab records which will arrive later this year. With our Objective Connect and VVO prototypes successfully established, and with ideas on how to innovate our practice (e.g. use of further automation to handle ingest, use of natural language processing to query records for sensitivity and topic modelling), we have manufactured a much stronger foundation of digital archiving practice to confront our next set of challenges.

ACKNOWLEDGMENT

The author wishes to thank all colleagues at NRS who helped provide text, edit, or supply resources for this paper, which covers what was a cross-departmental, highly innovative success for our institution.

SEEKING SUSTAINABILITY

Developing a Modern Distributed Digital Preservation System

Nathan Tallman

*Penn State University
United States
ntt7@psu.edu
[0000-0002-5308-4100](tel:0000-0002-5308-4100)*

Hannah Wang

*Educopia Institute
United States
hannah.wang@educopia.org
[0000-0002-6676-1254](tel:0000-0002-6676-1254)*

Abstract – As modern commercial developments in storage infrastructure mature and become increasingly available through popular open-source projects, there are important opportunities for digital preservation communities to leverage the increased efficiency and flexibility that these technologies offer. Not only do these developments offer a way to “modernize” the digital preservation technology stack and make it more efficient, but they also may allow digital preservation communities to seek increased sustainability per the triple bottom line: reduce the costs of operations, reduce required labor to maintain, and reduce the environmental impact. The twin values of affordability and sustainability are core to the mission of digital preservation, and the MetaArchive Cooperative is pursuing the research and development of a modern distributed digital preservation system to better practice these values.

Since its inception in 2004, MetaArchive has used Stanford University's Lots of Copies Keep Stuff Safe (LOCKSS) software for its core infrastructure to achieve distributed digital preservation. While LOCKSS has been reliable software for many years, recent evaluations by MetaArchive and a desire to make its practices more sustainable have led to MetaArchive partnering with Keeper Technology to explore software-defined storage and serverless computing technologies for digital preservation. The results of this multi-phase project will be shared with the digital preservation community, with the hopes that it will encourage other digital preservation technological developments with a similar sustainability mindset.

Keywords – distributed digital preservation, sustainability, software defined storage, cloud infrastructure, inclusion.

Conference Topics – Innovation; Environment.

I. INTRODUCTION

The MetaArchive Cooperative is an international distributed digital preservation (DDP) network based

in the United States and hosted by the Educopia Institute [1]. MetaArchive started in 2004 as a project with funding from the National Digital Information Infrastructure and Preservation Program (NDIIP) initially involving six southern United States universities and the Library of Congress. After initial success, MetaArchive transitioned from a project to an independent network open to any cultural heritage organization in 2007 [2].

Since its inception, MetaArchive has relied upon Stanford University's Lots of Copies Keep Stuff Safe (LOCKSS) software [3] for its core infrastructure to achieve distributed digital preservation in a Private LOCKSS Network (PLN) [4]. While LOCKSS has been reliable software for many years, MetaArchive has experienced challenges in meeting network and member needs. In 2020, MetaArchive conducted an evaluation of LOCKSS to identify strengths, weakness and challenges, and opportunities [5]. Findings verified that LOCKSS (1.x) was a viable, demonstrated cost-effective solution for bit-level preservation that is still used within the community. However, LOCKSS' high maintenance costs (direct and indirect) in MetaArchive's implementation, issues with scaling, and difficulties in content management for members inhibit the growth of the MetaArchive Cooperative. Among other recommendations, two parallel paths were established: continue with LOCKSS & explore potential benefits of the LOCKSS Architected as Web Services project [6] and explore alternative options for DDP that leverage modern technological innovations.

In January 2022, MetaArchive began a multi-phase research and development project to create a modern DDP system. Along with leveraging technical infrastructure advancements in the commercial

sector and simplifying digital preservation systems, MetaArchive's goal is to make digital preservation more sustainable per the triple bottom line [7]: reduce the costs of operations, reduce required labor to maintain, and reduce the environmental impact.

II. TRIPLE BOTTOM LINE SUSTAINABILITY AND THE UN SUSTAINABLE DEVELOPMENT GOALS

As a member of the Digital Preservation Services Collaborative, MetaArchive's mission to preserve cultural heritage and research is rooted in a set of shared core values [8]. The research and development described in this paper is being undertaken in pursuit of greater adherence to these values, particularly technological diversity, inclusion, and, most immediately, affordability and sustainability. While sustainability may be a buzzword, it's a critical concern for digital preservation. If sustainability is ignored, only the most well-resourced organizations will be able to put forward the money, labor, and infrastructure to steward digital content into the future. Not only are individual resources at risk of being lost, but if only the collections of the largest organizations persist, the consequences to representation in the academic, scholarly, and cultural record are severe.

A. Environmental Sustainability

Ben Goldman points out that data centers have a significant environmental impact consuming large amounts of energy and water in "It's Not Easy Being Green(e)" [9]. The 2019 NDSA Storage Infrastructure Survey, with only 85 responses to the question, indicates that cultural heritage organizations are storing at least 51 petabytes and as much as 114 petabytes of unique digital content [10]; since there are hundreds or thousands of cultural heritage organizations worldwide this number is likely a substantial underestimate. As Keith Pendergrass and his coauthors point out in "Toward Environmentally Sustainable Digital Preservation," even conservative estimates indicate that the world's cultural heritage organizations consume over five exabytes of storage, without even accounting for the fact that most organizations replicate content multiple times [11]. Furthermore, if digital preservation practices and policies do not include reappraisal and deaccessioning of preserved content, digital storage consumption will only grow unchecked in perpetuity. Without even considering the practices that produce physical and virtual servers and other infrastructure

components, digital preservation itself has a significant environmental impact.

B. Economic Sustainability

Despite overall price trends in commodity storage, preservation storage is expensive. The Digital Preservation Storage Criteria project's 61 criteria are an indication of why this is the case [12]. Cultural heritage organizations need to be strategic about digital infrastructure and not operate in old paradigms. MetaArchive members are small and large, DDP technology needs to be as low-cost as possible to be accessible to all.

C. Labor Sustainability

Many digital preservation systems continue to perform preservation actions in the application level of the technology stack. The application layer is often the most time-consuming and expensive level to maintain. Moving preservation actions into lower levels of the infrastructure, such as the underlying storage system, can reduce the amount of labor. Additionally, by using commercial-sector infrastructure, it's easier to recruit skilled professionals; conversely, it can be challenging to recruit developers to maintain bespoke cultural heritage applications.

D. UN Sustainable Development Goals

The UN Sustainable Development Goals are a framework for improving life for all living creatures. Goal 9 specifically calls for building resilient infrastructure, promoting inclusive and sustainable industrialization, and fostering innovation [13]. While most efforts in Goal 9 are focused on manufacturing and industry, digital infrastructure in cultural heritage organizations needs to embrace these principles if our collections are to persist for future users. Innovating our technology to adopt strategies used by the world's leading corporations will make our infrastructure less expensive, more resilient, and available to more organizations.

III. A FRAMEWORK FOR MODERN DISTRIBUTED DIGITAL PRESERVATION SYSTEMS

MetaArchive anticipates implementing technologies created in the past two decades when building our next-generation DDP system and re-thinking how digital preservation occurs from the ground up. This paper will focus on only two of those technologies: software-defined storage and serverless computing. For a more detailed look at

modern technology for digital preservation, see Nathan Tallman's 2021 article in *Information Technology and Libraries* [14].

A. *Software-Defined Storage*

Software-defined storage is an application that operates in-between physical disks and the operating system. It replaces traditional storage management with a dynamic, flexible, and resilient system that can help manage basic preservation activities. Instead of using a file system managed by an operating system or a storage appliance that presents a file system to the operating system, software-defined storage lets you build your own storage appliance or storage network that supports multiple protocols for access including file, object, and block storage [15]. Popular open-source software-defined storage solutions include Ceph [16] and Gluster [17] and there are several commercial offerings as well. A software-defined storage network can be configured within a single data center or multiple data centers anywhere in the world.

Software-defined storage has several features that support digital preservation. First, because you can build your own storage appliance/network, you can achieve hardware diversity to mitigate risk related to single points of failure. Combining different hard drive manufacturers, batches, and hard drive technology, as well as the server components themselves provides a broad spectrum of hardware diversity. Second, like RAID, software-defined storage builds in erasure encoding for parity. This helps to protect against loss when a hard drive or even an entire node or cluster in a storage network fails. The level of protection can be configured as needed. Third, software-defined storage offers many options for achieving geographic-redundant replications. Software-defined object stores can be set up in different availability zones with bucket-level policies to ensure replication to as many locations as needed and supported by the storage network. If you have ever used AWS, Azure, or Google cloud storage, you have almost certainly been using software-defined storage.

When using traditional file storage (SMB/CIFS) in a software-defined storage network, it's also possible to leverage modern file systems such as OpenZFS [18] and BTRFS [19] that have built-in features to ensure integrity. This is achieved by tracking

checksums for the underlying blocks in file storage and leveraging parity to repair blocks found to have loss. Managing fixity in the storage-level simplifies higher-level applications and significantly reduces the environmental impact when compared to ongoing file-level fixity checks. For a fuller explanation of ZFS and digital preservation, see Alex Garnett, Mike Winter, and Justin Simpson's 2018 iPRES paper [20].

B. *Serverless Computing*

A Serverless computing, sometimes also called function-as-a-service, allows you to offload certain tasks. Instead of running commands on the same server as the repository, you call a function-as-a-service that executes the action on a different server. It's the ultimate microservice that can be finely tuned to consume only the required resources; instead of having a monolithic stack that is highly resourced for peak activity, the repository stack can be optimized for repository management while the serverless computing platform is optimized for high performance and throughput, often managed by another entity, and paid for based on usage. Serverless computing can be invoked by an application to perform tasks like file format characterization, format migrations, replications, and more. Depending on the platform and your needs, a serverless function may use its own container or simply run on its host platform. Using a serverless computing for basic preservation activities simplifies the digital preservation stack that needs to be maintained, functions can be called as needed. If using a commercial FaaS platform, this reduces the need to maximally resource servers and maintain more infrastructure, which in turn simplifies labor requirements.

IV. RESEARCH AND DEVELOPMENT

MetaArchive is partnering with Keeper Technology (KeeperTech), a Virginia-based storage and data solutions company with deep expertise in software-defined storage [21], in carrying out this work. The work is structured in three phases, with off ramps at the end of each stage for both parties. Phase 1 is to collaboratively develop a high-level design document and feature requirements for a modern DDP system. Before Phase 1 commencement in January 2022, MetaArchive members used the Digital Preservation Storage Criteria to prioritize, justify, and refine initial requirements [12]. Additionally, OSSArcFlow

diagrams were developed for select basic network tasks [22]. KeeperTech reviewed these inputs and developed several questions to explore with MetaArchive members. KeeperTech distributed a questionnaire to all MetaArchive members and a more detailed set of questions that were explored in focus groups. Based on responses to these questions, the provided inputs, and previous conversations between MetaArchive, KeeperTech will craft a white paper or high-level design document. This document will be shared with the digital preservation community.

At the conclusion of Phase 1, if mutually agreeable to both parties, MetaArchive and KeeperTech will begin to implement the high-level design in a working prototype in Phase 2. The working prototype will include the development of architectures, infrastructure, and applications that will be capable of demonstrating proof-of-concept. Phase 2 will also include creating testing requirements to ensure success. Again, outputs from this phase will be shared with the digital preservation community as open source.

Phase 3 will explore a fully operational DDP network. Like Phase 2, it will only be pursued if both parties agree, though either may choose to work independently. Phase 3 tasks will include blueprints for central and network operations, deployment guides, documentation, software packaging, and other items necessary to move into production. Although not yet determined, MetaArchive may choose to contract with KeeperTech for ongoing support or even to have them host centralized components of the system.

V. CONCLUSION

While the concept of DDP emerged in the 1990s, digital preservation communities don't have to continue to rely on 1990s technology, nor should they. As Trevor Owens, quoting Martha Anderson in *The Theory and Craft and Digital Preservation*, says, digital preservation is a relay race [23]. It is a chain of hand-offs between mediums, systems, and stewards. Preservationists should expect to update their technology as time goes on, the same way they forward-migrate digital content itself. Doing so ensures that communities are using the most sustainable, efficient, and affordable means to achieve DDP goals. As the MetaArchive Cooperative aims to put digital preservation in reach for organizations of any size, it is vital to ensure it is

fiducially responsible in meeting that goal. Ultimately, MetaArchive hopes that the results of this project will not only improve services for its own members, but also will encourage other digital preservation communities to adopt and pursue technological developments with a similar sustainability mindset.

REFERENCES

- [1] Educopia Institute, "Educopia Institute | Empowering Collaborative Communities," 2022. [Online]. Available: <https://educopia.org>. [Accessed 21 January 2022].
- [2] K. Skinner and M. Halbert, "The MetaArchive Cooperative: A Collaborative Approach to Distributed Digital Preservation," *Library Trends*, vol. 57, no. 3, pp. 371-392, 2009. [Online] Available: <https://doi.org/10.1353/lib.0.0042>. [Accessed 28 January 2022].
- [3] Stanford University, "LOCKSS," 2021. [Online]. Available: <https://www.lockss.org>. [Accessed 21 January 2022].
- [4] V. Reich and D. Rosenthal, "Distributed Digital Preservation: Private LOCKSS Networks as Business, Social, and Technical Frameworks," *Library Trends*, vol. 57, no. 3, pp. 461-475, 2009. [Online] Available: <https://doi.org/10.1353/lib.0.0047>. [Accessed 28 January 2022].
- [5] N. Tallman, Z. Vowell, B. Robbins and M. Schultz, "MetaArchive LOCKSS Evaluation," 2020. [Online] Available: <https://doi.org/10.26207/3acj-bs29>. [Accessed 4 February 2022].
- [6] Stanford University, "Frequently Asked Questions | LOCKSS," 2021. [Online]. Available: <https://www.lockss.org/about/frequently-asked-questions#laaws>. [Accessed 21 January 2022].
- [7] K. Miller, "The Triple Bottom Line: What It Is & Why It's Important," 08 December 2020. [Online]. Available: <https://online.hbs.edu/blog/post/what-is-the-triple-bottom-line>. [Accessed 28 January 2022].
- [8] Digital Preservation Services Collaborative, "Digital Preservation Declaration of Shared Values," 12 April 2018. [Online]. Available: https://dpscollaborative.org/shared-values_en.html. [Accessed 18 February 2022].
- [9] B. Goldman, "It's Not Easy Being Green(e): Digital Preservation in the Age of Climate Change," in *Archival Values: Essays in Honor of Mark A. Greene*, Chicago, American Library Association, 2018, pp. 274-295. [Online]. Available: <https://scholarsphere.psu.edu/resources/381e68bf-c199-4786-ae61-671aede4e041>. [Accessed 28 January 2022].
- [10] 2019 NDSA Storage Infrastructure Survey Working Group, "2019 Storage Infrastructure Survey Report," 24 February 2020. [Online]. Available: <https://doi.org/10.17605/OSF.IO/UWSG7>. [Accessed 28 January 2022].
- [11] K. Pendergrass, W. Sampson, T. Walsh, L. Alagna, "Toward Environmentally Sustainable Digital Preservation," *The American Archivist*, vol. 82, no. 1, 2019. [Online] Available: <https://doi.org/10.17723/0360-9081-82.1.165>. [Accessed 18 August 2022].
- [12] S. Schaefer, N. McGovern, A. Goethals, E. Zierau and G. Truman, "Digital Preservation Storage Criteria," 10 December 2018. [Online]. Available:

<https://doi.org/10.17605/OSF.IO/SIC6U>. [Accessed 28 January 2022].

- [13] United Nations, Department of Economic and Social Affairs, "Goal 9 | Department of Economic and Social Affairs," 2021. [Online]. Available: <https://sdgs.un.org/goals/goal9>. [Accessed 28 January 2022].
- [14] N. Tallman, "A 21st Century Technical Infrastructure for Digital Preservation," *Information Technology and Libraries*, vol. 40, no. 4, 2021. [Online] Available: <https://doi.org/10.6017/ital.v40i4.13355>. [Accessed 4 February 2022].
- [15] M. Carlson, A. Yoder, L. Schoeb, D. Deel, C. Pratt, C. Lionetti and D. Voigt, "Software Defined Storage," January 2015. [Online]. Available: https://www.snia.org/sites/default/files/SNIA_Software_Defined_Storage_%20White_Paper_v1.pdf. [Accessed 4 February 2022].
- [16] "Ceph.io," 2022. [Online]. Available: <https://ceph.io/>. [Accessed 2022 February 2022].
- [17] Red Hat, Inc., "Gluster," 2019. [Online]. Available: <https://www.gluster.org/>. [Accessed 2022 February 2022].
- [18] "OpenZFS," 2021. [Online]. Available: https://openzfs.org/wiki/Main_Page. [Accessed 4 February 2022].
- [19] "btrfs Wiki," 4 February 2022. [Online]. Available: https://btrfs.wiki.kernel.org/index.php/Main_Page. [Accessed 4 February 2022].
- [20] A. Garnett, M. Winter and J. Simpson, "Checksums on Modern Filesystems, or: On the Virtuous Consumption of CPU Cycles," in *iPres 2018 Conference [Proceedings]*, Boston, 2018. [Online] Available: <https://doi.org/10.17605/OSF.IO/Y4Z3E>. [Accessed 4 February 2022].
- [21] Keeper Technology, LLC., "KeeperTech," 2022. [Online]. Available: <https://www.keeper.tech/>. [Accessed 4 February 2022].
- [22] K. Skinner, S. Meister and C. Lee, "OSSArcFlow," 2020. [Online]. Available: <https://educopia.org/ossarcflow/>. [Accessed 4 February 2022].
- [23] T. Owens, *The Theory and Craft of Digital Preservation*, Baltimore, MD: Johns Hopkins University Press, 2018.

"A TARTAN RATHER THAN A PLAIN CLOTH"

Building a Shared Workflow to Preserve the Regional Ethnology of Scotland Project Archive

Sara Day Thomson

University of Edinburgh
UK

Sara.Thomson@ed.ac.uk

[0000-0002-3896-3414](tel:0000-0002-3896-3414)

Abstract – This paper provides a case study of a shared workflow to preserve the oral history recordings created through the Regional Ethnology of Scotland Project. This workflow has been developed as the first production run of a semi-automated integration of Archivematica at the Centre for Research Collections (CRC) at the University of Edinburgh. This experience demonstrates that digital preservation, above all, is about people and to succeed requires the input of a range of perspectives and skillsets.

Keywords – oral history, workflow documentation, Archivematica, automation

Conference Topics – Community; Resilience

I. BACKGROUND

In an interview from 20th August 2012 [1], Robert McQuistan from Carsluith tells his interviewer, Mark Mulhern, a lead researcher with the Regional Ethnology of Scotland Project (RESP) [2], what he has learned as a volunteer fieldworker collecting the stories of his neighbors:

"Oh, well, ah've learnt how intensely people feel about their own history, they, once they get into it and, one person says 'Oh ye know, it's taken me back and it's made me rethink and relive my past and stuff that ah'd forgotten about has come back.'

"So, in that personal sense, it's quite powerful for them but from my point of view just the flow of a person's life, just how it developed and evolved over the years. And the changes, just the remarkable changes from five to ten years, to twenty years, just how it all piles up.

"All of that," he says, 'it's like a rich tapestry, it's like a tartan rather than a plain cloth.'

The way McQuistan describes the accounts of individual lives in his village, the "remarkable changes", also reflects the challenges faced by digital preservation. While the relentless evolution of technology poses a risk to continued access, the inevitable transformation of staff, organizations, and end users also poses a risk of digital resources dissipating in the mists of time. The fieldworkers, volunteers, researchers, ethnographers, curators, and archivists involved in projects like the RESP Archive move on to other things. The institution looking after the resource - the files, documentation and all their idiosyncrasies - undergoes restructures, staff come and go, and priorities evolve. Digital preservation aims to anticipate these nebulous and unpredictable changes so that unique resources such as the RESP Archive have the best possible chance of persisting into the future.

This paper provides a case study of a shared digital preservation workflow built to withstand those changes. It reflects the steps taken to preserve the oral history recordings created through RESP, a project managed by the European Ethnological Research Centre [3]. This workflow has been developed as the first production run of the bespoke implementation of Archivematica [4] at the Centre for Research Collections (CRC) at the University of Edinburgh [5], the custodians of the RESP Archive. This experience demonstrates that digital preservation, above all, is about people and to succeed requires the input of a range of perspectives and skillsets. A digital archivist can help bridge the gap between a systems developer and project archivist. The nuanced understanding researchers

and project archivists possess about the digital resources has a direct impact on the effectiveness and robustness of a digital preservation strategy. Long term access to these life stories from across Scotland is best assured by different types of practitioners working together, implementing digital preservation measures as early as possible.

II. AUTOMATING DIGITAL PRESERVATION WITH ARCHIVEMATICA AT THE CENTRE FOR RESEARCH COLLECTIONS

The digital preservation system at the CRC - encompassing archives, rare books, art collections, museums, and reader services - uses Archivematica to process digital content and create AIPs and (in some cases) DIPs. This system, built by a dedicated developer (Hrafn Malmquist) in 2018-19, automates the transfer of AIPs and DIPs to DSpace [6] and ArchivesSpace [7] using the integration DSpace via REST API [8]. The system also integrates Archivematica with Tivoli Storage Manager (TSM), a proprietary tape storage system by IBM [9], used as primary preservation storage.

These Archivematica integrations fulfill the need to automate the transfer of digital archival materials directly to preservation storage and to the archives discovery system ArchivesSpace. The AIPs created by Archivematica are pushed to DSpace Collections and to TSM so that the AIP is duplicated. DSpace, running on disk storage, provides easy access to AIPs for investigating issues and evaluating processes. The DIPs created by Archivematica are pushed to DSpace Collections and then to the catalog record where they appear as Digital Objects. In sum, 1 Archivematica SIP = 1 Archivematica AIP/DIP = 1 DSpace Item = 1 ArchivesSpace Digital Object Record. The objective of this implementation was to create a seamless workflow from ingest to preservation to access.

These Archivematica integrations were built based around University Court Senate Records. These archival records mainly comprise meeting minutes in MS Word and Adobe PDF formats. This corpus of data posed relatively few technical complications for the envisioned workflow. The processing of the RESP Archive provided the first corpus of data to test the implementation with content that had not informed development.

III. A PROGRAMMATIC WORKFLOW FOR THE RESP ARCHIVE

The RESP Archive includes oral history recordings (audio and moving image files), transcriptions (PDF), and photographs (JPG) created through RESP, a project that enables communities across Scotland to work together to record information about their local life and society. This work is carried out on a regional basis with volunteer fieldworkers carrying out fieldwork interviews. To maximize the usability of the collections for researchers and others, detailed summaries for each item are provided in the ArchivesSpace catalog and all interviews have been transcribed in full. RESP considers the collection to be the creation of those who have made the recordings. As such, it is a central aim of the project that the recordings are made freely available in an easily accessible way, presented under Creative Commons. The RESP Archive, as a result, contains a collection of fieldwork interviews rich in detail about all aspects of life, place, and memory from different regions of Scotland. The project began in 2013 and to date over 1,500 recordings have been added to the collection and fully preserved. The RESP is funded until June 2023 with a hope that additional funding will be secure such that the project will continue until June 2028.

While the RESP interviews follow a relatively uniform model, the process also involves a good deal of organic growth and deviation, leading to a relatively complex archival structure. The RESP Archive is arranged by Region, then by Fieldworker, each Fieldworker comprising a series of interviews, some including only a single interview and others comprising closer to ten. Some Fieldworkers, like McQuistan, have also been interviewed to gather reflections on the methodology and experience. Some Fieldworkers are individuals and some are entire groups, like the Campie Primary School P5 pupils [10]. The hierarchy (and deviations across the collections) created a challenge for developing a workflow for a semi-automated Archivematica implementation developed for a relatively flat record series (the University Court meeting minutes). While automation has the power to exponentially speed up processing, it also requires materials and metadata to be structured in a particular way.

The structure of the RESP Archive wasn't the only challenge for the workflow. The ArchivesSpace interface did not support the accessibility and usability required by the target end users of the collection, who are not expected to be familiar with archival research. In ArchivesSpace - built as an

archives catalog not a digital repository system - discovering content and browsing the fully digital collection was cumbersome and opaque compared with the web-based discovery most people expect. The RESP Archive team, with the developer who built the Archivematica implementation, opted to build a website that would automatically pull metadata from ArchivesSpace and corresponding files from DSpace.

More fundamentally, the workflow suffered from a heavy dependency on developer support, as the workflow was initially developed with no digital archivist in post. The original developer, who had detailed knowledge of the systems involved, had moved to a completely different project. The project archivists had no training in Archivematica and digital preservation processing was outside their remit. These circumstances led to long delays in establishing the workflow, deciding requirements for preservation metadata, and processing content.

IV. ADAPTING A PROGRAMMATIC WORKFLOW TO A MANUAL WORKFLOW

To reduce dependency on developer support, the workflow needed to be re-designed so that it could be implemented by the digital archivist with support from the RESP Archive team. First, the programmatic workflow had to be documented (to a basic level) and broken down. The main tasks in the workflow which had been carried out programmatically included:

- Creation of a metadata file encoded in json to instruct Archivematica where to send the AIP and DIP
- Creation of directories in the required structure
- Transfer of files from network storage (Data Store) to the staging area for ingest into Archivematica
- Execution of the Archivematica process (referred to as a 'transfer')

Some parts of the workflow were already manual (or only partly automated):

- Creation of Collections in the dedicated DSpace repository
- Suppression of unredacted files in DSpace
- Deletion of unredacted files in ArchivesSpace
- Quality assurance of the website

Performing any of these previously automated tasks manually is more labor-intensive, but not

prohibitive. The more serious problem is that manual processing creates a greater risk of human error. However, because the RESP Archive team no longer had to wait for availability from a developer, the content could be transferred in smaller batches as it became available. Furthermore, members of the team are able to check each other's work throughout the workflow.

Due to the delays created by the developer dependency in the initial workflow, the RESP Archive team was keen to take on more parts of the digital preservation workflow. During the process of transforming the workflow, it was well-documented. As a result, it was relatively straightforward to identify those parts of the digital preservation workflow that could be handed over to project archivists upstream. The main barrier to handing these tasks over was knowledge of the systems and technologies involved. The digital archivist provided some basic training and step-by-step documentation for DSpace, alongside a general overview of Archivematica. While the metadata file encoded in json could be produced by hand in Notepad or a similar program, the team decided this approach entailed too high a risk of human error.

Narrowing down the remaining barriers to the workflow enabled the team to provide tightly scoped requirements to the development team. Fortunately, the head of the development team was able to include the creation of a lightweight, web-based tool as part of a larger internally-funded project. The result is JSON Convert, a simple form with metadata fields that transforms input text into the json structure required for the automation of the Archivematica workflow. Using JSON Convert, the RESP Archive team has taken ownership of a time-consuming task, creating the metadata file for each SIP alongside the creation of the ArchivesSpace catalog. This gives the project archivists, who work more closely with the fieldworkers and researchers, control over the basic metadata included with the AIPs.

Distributing parts of the digital preservation workflow has raised awareness of digital preservation and provided an opportunity for project archivists to upskill in digital methods and systems. Furthermore, the previously specialized, programmatic tasks in the workflow have now been transformed to manual tasks that a wider set of practitioners have the skills to perform. As the workflow has been devised and tested and revised,

the RESP Archive team has documented the digital preservation tasks so they can be implemented by practitioners with little or no experience of Archivematica or encoding information in JSON.

V. LESSONS LEARNED AND NEXT STEPS

While it may seem like a backwards step to transform a programmatic workflow into a manual one, the process has allowed the practitioners involved to evaluate the process in detail. Before expounding on future plans, it must first be acknowledged that the resource to process and catalog this collection so quickly, to such a high quality, is a real privilege made possible by external project funding. The majority of CRC collections do not enjoy this level of staff resource. As a result, the model created by the RESP Archive will not be easily applied to other archive collections (though it does provide valuable precedent).

The new, frontend workflow for the digital preservation of the RESP Archive has become well-established to work effectively. However, there are a number of improvements that have been identified. A new platform or digital repository purpose-built for discovering and engaging with digital archives would circumvent the need for creating project websites to improve accessibility and usability, removing a project-specific system from the workflow. Furthermore, the CRC remit includes large digitized collections, digital artworks, and datasets. ArchivesSpace will not support access for these non-archival formats.

Before leaving, the implementation developer submitted a request to Artefactual (the developers of Archivematica) for functionality to automatically suppress unredacted files (i.e. allow passing an authorization policy when depositing to DSpace). This automation would prevent the need to manually suppress or remove these files after the completion of an automated workflow [R]. The digital archivist aims to re-introduce automation more generally, for example to create structured directories. These refinements to requirements for automation have been possible to identify through implementation of the manual workflow.

Most significantly, the aim is to further develop JSON Convert so that it can be used for the submission of metadata and files directly to digital preservation. With time, this transfer method could be rolled out not only across the CRC, but across the

library and the wider university. Future development might be modeled on the AVP Exactly tool [11] which currently uses FTP transfer or the web form created for the CRC's Collecting Covid-19 Initiative which transfers files over HTTPS [12].

This workflow, developed for a remarkable collection of recorded memories and life stories, has provided a basis for the expansion of digital preservation at the CRC. The progress made on the RESP Archive shows great promise for what's to come, in the words of Robert McQuistan, "the changes, just the remarkable changes from five to ten years, to twenty years, just how it all piles up".

REFERENCES

- [1] Interviews of Robert McQuistan, EERC/DG/DG14/3. https://collections.ed.ac.uk/eerc/record/165128/archival_object.
- [2] RESP Project Website. <https://collections.ed.ac.uk/eerc/>.
- [3] EERC at the University of Edinburgh. <https://www.ed.ac.uk/literatures-languages-cultures/celtic-scottish-studies/research/eerc>.
- [4] H. Malmquist, 'Automating OAIS compliant digital preservation using Archivematica and DSpace' (26 November 2019). DOI: <https://doi.org/10.5281/zenodo.3554060>.
- [5] Centre for Research Collections, University of Edinburgh. <https://www.ed.ac.uk/information-services/library-museum-gallery/cultural-heritage-collections/crc>.
- [6] DSpace Wiki. <https://wiki.lyrasis.org/display/DSPACE/Home>.
- [7] ArchivesSpace. <https://archivesspace.org/>.
- [8] Artefactual, Archivematica Storage Service 0.13.0, Documentation, Administering the Storage Service. <https://www.archivematica.org/en/docs/storage-service-0.13/administrators/#dspace-via-sword2-api-or-dspace-via-rest-api>.
- [9] IBM, IBM Tivoli Storage Manager Documentation. <https://www.ibm.com/docs/ko/tsm>.
- [10] Interviews of Campie Primary School P5 pupils, EERC/EL/EL21. https://collections.ed.ac.uk/eerc/record/190167/archival_object.
- [11] Github, archivematica/issues, 'Feature: Allow passing an authorisation policy when depositing to DSpace #1316'. <https://github.com/archivematica/Issues/issues/1316>.
- [12] AVP, Exactly. <https://www.weareavp.com/products/exactly/>.
- [13] Centre for Research Collections, University of Edinburgh, 'Collecting Covid-19 Initiative'. <https://www.ed.ac.uk/information-services/library-museum-gallery/cultural-heritage-collections/crc/collecting-covid-19-initiative> (live URL) or <https://web.archive.org/web/20220305173433/https://www.ed.ac.uk/information-services/library-museum-gallery/cultural-heritage-collections/crc/collecting-covid-19-initiative> (archived URL, 2022-03-05).

VANISHED:

Preserving the Carmichael Watson Project Website Offline Using Webrecorder

Sara Day Thomson

*University of Edinburgh
UK*

*Sara.Thomson@ed.ac.uk
[0000-0002-3896-3414](tel:0000-0002-3896-3414)*

Anisa Hawes

*Freelance Researcher & Web Archivist
UK*

anisa.hawes@icloud.com

Abstract – In 2021, the Carmichael Watson project website — a highly valued resource of Gaelic culture, culminating in an investment of over £750,000 — faced imminent termination. This case study details how this project website, only online from 2013 until 2018, came to imminent risk of permanent loss. It then presents the strategy undertaken to transform it into a more sustainable format through web archiving and to revive its public accessibility.

Keywords – Web Archiving, Webrecorder, Technology Obsolescence, Disaster Recovery, Heritage Stewardship

Conference Topics – Community; Resilience.

I. INTRODUCTION

Due to security issues with underlying infrastructure, in 2021, the Carmichael Watson project website [1] — a highly valued resource of Gaelic culture, culminating in an investment of over £750,000 in external grant funding — faced imminent termination. The Digital Archivist at the University of Edinburgh, in collaboration with the Digital Library Development team and freelance web archivist Anisa Hawes, intervened to capture and preserve this resource using Webrecorder's ArchiveWeb.page [2]. Ultimately, the University of Edinburgh aims to embed good practice in digital preservation at the outset of creating digital resources, but like every steward of digital data or heritage resources, the university is subject to practical constraints. This case study details how this project website, only online from 2013 until 2018, came to imminent risk of permanent loss. It then presents the strategy undertaken to transform it into a more sustainable format through web archiving and to revive its public accessibility.

II. BACKGROUND

The Carmichael Watson Project, based on archives held by the Centre for Research Collections (CRC) at the University of Edinburgh [3], revolves around the papers of the pioneering folklorist Alexander Carmichael (1832-1912) and brings to life customs, stories, songs, and beliefs from the Gaelic-speaking areas of Scotland. It offers fundamental insights into the creation of Carmichael's greatest work *Carmina Gadelica* [4] but also supports interdisciplinary cooperation between local and scholarly communities for collaborative research in history, theology, literary criticism, philology, placenames, archaeology, botany, and environmental studies [5]. Through cataloging, indexing, transcribing, translating, digitizing, and conserving, this project opened up and made accessible this important collection to the academic and broader community. In particular, this web resource was developed to make erratic and multilingual notebook entries more readable and accessible, with transcriptions and detailed notebook entry-level descriptions. The project team also created EAC (Encoded Archival Context) records for many of the individuals from whom Carmichael collected material, giving prominence to people regarded as ordinary folk and marginalized groups. Linking to APIs to give geographical context to each individual item within a notebook provided increased usability and engagement with the resource, an essential part of key research interests.

III. A SERIES OF UNFORTUNATE EVENTS

The project website has been unavailable to the public since 2018 when the original stewards turned it off due to security concerns. The original URL, <http://www.carmichaelwatson.lib.ed.ac.uk>, now points to a holding page but the actual resource is not accessible to the public. However, the Library Development team maintained access for internal users via Virtual Machine (VM) with restricted access. Unfortunately, this VM for internal access was on RedHat 5, a technology that had reached its end of life by 2021. As a result, IT Infrastructure (ITI) notified the Development team to terminate all services running on it. In addition to the security issue posed by running end-of-life technology, ITI had been paying an additional license fee for the VMs, so maintaining the resource had been incurring an ongoing cost, despite being unavailable to the public.

The archives of the Carmichael Watson collections remain available through other dispersed access systems. When the website was taken offline in 2018, the project team ensured access to digitized versions of the notebooks and full archival descriptions through ArchivesSpace (the university's archives discovery system) [6]. However, the project team envisioned and invested in the website to make the materials more accessible and usable by researchers and members of the communities represented in the collections. The content and functionality underpinning that wider accessibility exist only in the web-based resource, including the TEI [7], the geotools, and the handwriting guide.

The underlying code for the project website was saved, but it would require substantial funding to build a new website from scratch or integrate this resource into another, newer platform. Resource for development at this scale — which would essentially repeat a project already substantially funded in the recent past — was simply not available. Furthermore, re-developing the resource on a new platform would inevitably lead to the same situation: the need to redevelop the resource when its underlying infrastructure inevitably becomes obsolete.

IV. WEB ARCHIVING ... OFFLINE

The first step to developing a strategy for archiving this project website was to check for a copy in a national web archive. The UK Web Archive [8], as a national program with a remit to capture UK websites, provides a first port of call when searching for legacy web resources from the University of

Edinburgh. The university itself does not have a web archiving program, but since 2020 has been collaborating with the UK Web Archive [9] to better look after its valuable (and vast) web estate. The UK Web Archive had, in fact, crawled the original URL on 17 occasions since 2013, noting on the backend that the Live Site Status was 'Vanished'. Unfortunately, under closer examination, none of the successful captures contained the original formatting and many pages were missing images. The 'QA Status' in the UK Web Archive's system W3ACT [10] was listed as 'none', which perhaps provides some explanation for why the crawls, even later ones, lacked the underlying style sheets and images (a problem affecting many of the university's web pages in the UKWA). Neither the Wayback Machine run by the Internet Archive [11] nor any other copies discovered through querying Memento Time Travel [12] held more than the top-level pages.

In the absence of a complete archived copy, the next option was to test if it could be archived offline — from the internal VM — using manual web archiving tools. Conifer [13] was able to capture a high-quality copy of a selection of pages, including authentic formatting and images. The Webrecorder tool used by the Conifer service [14] was built to preserve websites, even complex and interactive websites, to a high degree of fidelity (or accuracy). However, a manual tool like Webrecorder requires the user to click every link and activate every function in order to capture content. It would take a full-time archivist several weeks, if not months, to archive the Carmichael Watson project website in its entirety using this approach.

There are, however, important benefits to using a Webrecorder-driven approach to preserving this resource, namely the longevity of the output format and resulting persistent access. Using Webrecorder tools, the resource can be captured and transformed into a warc or wacz file, ingested into a digital preservation system, and accessed with appropriate user guidance through local systems. Perhaps most importantly, the tool is incredibly accessible and enabled the team, with no institutional web archiving infrastructure, to start capturing the resource immediately.

V. RESURRECTING CARMICHAEL WATSON WITH ARCHIVEWEB.PAGE

In collaboration with Anisa Hawes [15], the University of Edinburgh opted to pursue the

preservation of the Carmichael Watson project website using up-to-date Webrecorder tool ArchiveWeb.page, released in January 2021. ArchiveWeb.page allows users to systematically capture a web resource, mainly through a Chrome extension which enables capture through normal interactions with the browser.

The active capture of the website was only one part of the rescue workflow. To ensure no pages were overlooked, the website was first mapped and scoped to understand the boundaries and extent of the resulting archival resource. Due to the restricted access to the resource (via VPN), Anisa Hawes mapped the resource manually, documenting the URLs of each of the many navigation pathways in spreadsheets. Once capture commenced, the index of captured URLs in Archiveweb.page had to be cross-referenced to check all relevant pages had been captured and the quality of each capture (text, formatting, and functionality) assured. In order to support access, the archived resource has been annotated and documented. Annotation is particularly important to explain to target end users where the archived website 'ends' and when errors or missing content derive from the original and when from failures of the capture tool.

Capture commenced in summer 2021, though from the outset a number of challenges posed serious impediments to this planned workflow. The first and perhaps most prohibitive challenge was the design of the website itself. Rather than individual pages resolving to a persistent URL, individual pages had been duplicated in multiple user navigation pathways, including multiple access points and browsable indexes. Therefore, the same identical page exists at six, seven, eight, or more unique URLs culminating in nearly 20,000 URLs. This challenge highlighted the importance of resource mapping to establish a realistic scope and timescale. Ultimately, the team decided it was simply not feasible, in the available time, to capture the entire website manually. This decision drastically increased the importance of annotation and documentation to clearly communicate which pathways and access points are archived and which are excluded.

In addition to the time-consuming task of manually capturing so many pages, technology constraints also created barriers. As mentioned, the obsolete infrastructure underpinning the VM created security concerns, requiring VPN restrictions. While providing a layer of security, the VPN also slowed down the power and speed of ArchiveWeb.page and

the machine used for capture. The VPN also created a hindrance to experimenting with Browsertrix [16], a Webrecorder tool with automation to support capture. Though in theory Browsertrix could have made it feasible to capture the website in full, the VPN caused the tool to time out before crawls could be completed.

The functionality of ArchiveWeb.page, only released months before capture commenced, presented challenges as well. As the number of captured URLs accumulated, the ArchiveWeb.page index stopped displaying them all, making it difficult to check that all pages from a capture session had been successfully archived. Though the deprecated predecessor of ArchiveWeb.page (Webrecorder desktop) had functionality for adding metadata and annotation directly into the archived resource, this early version of Archiveweb.page does not, though a request for improved curatorial functionality has been submitted to Webrecorder. Therefore, annotation and descriptive metadata has been created separately, which will need to be maintained over time through the CRC's digital preservation program. A fuller presentation of the methodology and challenges of capturing an offline resource this way can be found on the Digital Preservation Coalition event page for the December 2021 Web Archiving & Preservation Working Group [17].

Though these technology challenges created hurdles for archiving the Carmichael Watson project website, the strategy for access looks much more promising. Based on recommendations from Anisa Hawes, the team will be hosting the archived resource on a local server and providing a link from the CRC's discovery record in ArchivesSpace. This approach, developed by Stanford University Libraries [18], will allow users to click on a link that takes them to the interface for the Webrecorder companion tool for replay — ReplayWeb.page [19] — to view and interact with the archived website through their browser.

The capture phase of the project has been completed and the VM turned off. The archive of the Carmichael Watson project website created by Anisa Hawes in Archiveweb.page is now the only, golden copy of the resource. Unfortunately, the large size of the warc file containing the archive exceeds the export limit of Archiveweb.page and, at the time of writing, an alternative export method is being explored with the developers. Arguably, the most difficult and important task remains: to clearly and effectively communicate this archived version of a

beloved resource to its target users, many of whom will never have heard of web archiving.

VI. LESSONS LEARNED

So far the team has learned some important lessons. The first lesson, of course, is to act sooner, acknowledging that the way resources are built or managed is not always in the control of the practitioner or team responsible for archiving. If web archiving had commenced before the resource came off the public web, automated tools like Browsertrix might have made the process much faster and cheaper. The CRC may have even been able to work with the UK Web Archive to improve automated capture, which would have secured a copy of the resource in one of the world's most well-supported and well-known web archives.

It has become clear that in order to be responsible custodians of its public record on the web and its web-based collections, the university needs to actively engage with the UK Web Archive and build local capacity for web archiving. Immediate investment is required in the staff resource to QA the University of Edinburgh Web Estate in the UK Web Archive and build relationships with content creators and stakeholders (a business case is currently pending). These relationships would enable, over time, the transformation of development practices for web-based resources. These relationships could also enable archiving from the point of creation in a way that supports long-term requirements, whether through the UK Web Archive or alternative approaches like Webrecorder.

In summary, the University of Edinburgh, as perhaps with other organizations in the HE sector and others, continues to struggle to embed digital preservation early in the lifecycle of digital materials. There is a lot of work to do to raise awareness of and build capacity for digital preservation across the university. In the meantime, it will be a rewarding victory to make the archived Carmichael Watson website available to the researchers, teachers, and members of the community who will now be able to use it for years to come.

REFERENCES

- [1] Original URL for the Carmichael Watson project website (2013-2018).
<http://www.carmichaelwatson.lib.ed.ac.uk/cwatson>
- [2] Webrecorder Tools, ArchiveWeb.page.
<https://webrecorder.net/tools#archivewebpage>
- [3] Centre for Research Collections, University of Edinburgh.
<https://www.ed.ac.uk/information-services/library-museum-gallery/cultural-heritage-collections/crc>
- [4] Carmichael, Alexander. *Carmina Gadelica* [or, *Ortha Nan Gàidheal*]: Hymns and Incantations with Illustrative Notes on Words, Rites, and Customs, Dying and Obsolete, Translated into English by Alexander Carmichael. Edinburgh: Floris Books, 2006.
- [5] List of publications referencing the Carmichael Watson archives. <https://www.ed.ac.uk/information-services/library-museum-gallery/crc/research-resources/gaelic/carmichael-watson/publications>
- [6] ArchivesSpace. <https://archivesspace.org/>
- [7] Text Encoding Initiative. <https://tei-c.org/>
- [8] UK Web Archive. <https://www.webarchive.org.uk/>
- [9] UK Web Archive blog. 'University of Edinburgh's Collecting Covid-19 Initiative — Collaborative Collection Building with the UK Web Archive' (12 March 2021), <https://blogs.bl.uk/webarchive/2021/03/university-of-edinburghs-collecting-covid-19-initiative-collaborative-collection-building-with-the-u.html>
- [10] UK Web Archive, W3ACT. <https://github.com/ukwa/w3act/wiki/W3ACT-User-Guide>
- [11] Internet Archive, Wayback Machine. <https://archive.org/>
- [12] Memento Time Travel. <http://timetravel.mementoweb.org/>
- [13] Conifer. <https://conifer.rhizome.org/faq>
- [14] Webrecorder Desktop. <https://github.com/webrecorder/webrecorder-desktop>
- [15] Anisa Hawes. <https://anisahawes.github.io/about/>
- [16] Browsertrix Crawler. <https://webrecorder.net/tools#browsertrix>
- [17] Web Archiving & Preservation Working Group, General Meeting December 2021. <https://www.dpconline.org/events/past-events/wapwg-meeting-dec2021>
- [18] Stanford Libraries, 'Archiving Instagram posts' (2021-09-09), <https://library.stanford.edu/blogs/stanford-libraries-blog/2021/09/archiving-instagram-posts>
- [19] Replayweb.page. <https://webrecorder.net/tools#replaywebpage>

MACINTOSH RESOURCE FORKS

Choosing File Formats for Preservation

Tyler Thorsted

Church of Jesus Christ of Latter-day Saints

Church History Library

USA

thorsted@churchofjesuschrist.org

[0000-0003-0292-0962](tel:0000-0003-0292-0962)

ABSTRACT - The preservation of files from early Macintosh Classic (OS <=9) may often require special handling in order to ensure long term preservation and rendering. The classic Macintosh operating system would use two “forks”, a data fork and a resource fork. Resource Forks may contain graphics, sounds, fonts, and additional code. In addition, the file system would store two 4 digit codes for each file, one identifying the creating software and another identifying the type of file as extensions were rarely used. Because of this unique information within the Macintosh file system, most modern preservation systems are only aware of the data fork and information can be lost. Round-tripping a file through a preservation system and back to the original OS can help identify potential loss.

Keywords – Macintosh, Resource Fork, HFS, Risk, Finder

Conference Topics – Exchange; Resilience

I. INTRODUCTION

All documents and files created on a Macintosh between the 1980-90's will have additional data which is important to proper functionality and rendering. Not all files used a resource fork to store some or all of its data, but many did. This paper will take an early project format from the popular audio recording application Pro Tools to illustrate different methods of preservation of files with these attributes.

DigiDesign Pro Tools [1] was an early digital audio workstation for recording audio. The software would create a folder structure for each session which included the session file, a folder of audio clips in the SoundDesigner 2 (SD2) file format [2], and additional folder of fades also in the SD2 format. The session file and the SD2 files each used a resource for part or all of their data.

A data set was created [3] on an original Macintosh running operating system 7.5. Pro Tools version 3.4 was used to create four different sessions, with and without linked SD2 audio files.

II. FILE FORMATS

A. Data Set

The two main file formats in this data set are the Pro Tools version 3 project file and native audio format used by Pro Tools, SoundDesigner 2. Pro Tools session files are saved preserving all the parameters of the recording project. What made the Pro Tools session format unique is that all the parameters were stored in the resource fork of the file with nothing in the data fork (see Figure 2). For the Sound Designer II files, the raw data was stored in the data fork, but all the information on sample rate, duration, channels were stored in the resource fork. Once development was made to support Mac OS X in version 6, all of this changed.

B. Preservation Formats

For Preservation purposes, the following formats were chosen to test preservation processes.

- Disk Image
- Original Logical Copy
- Stuffit
- MacBinary
- AppleSingle
- AppleDouble

III. INGEST

Each format was added to a folder and ingested in our preservation system. This ingest includes file

identification using DROID and scanned for viruses. Validation depends on tools available per format and identification, our system does not have anything in place for these formats.

A. Disk Image

The disk image was identified and ingested into our system. As seen in the chart below, the disk image was able to store all the original data, but only the disk image was visible to the preservation system, not the individual files on the disk.

B. Logical Copy

A direct copy from the disk image was made and added to a ZIP file for ingest. The ZIP container included the AppleDouble “.” resource forks and were extracted during ingest. AppleDouble files were identified, but Pro Tools project files failed as they contained zero bytes. With no extensions, the original files produced an error. Custom folder icons also caused a failure during ingest because of illegal characters in filename.

C. Stuffit Container

Adding the contents of the disk image to a Stuffit container made a single “.sit” file which was identified by DROID. Similar to the disk image, the individual files were not visible to the preservation system. One method is making individual files using Stuffit, adding info in the filename. A Stuffit file defaults to compression of the file, but this feature can be turned off, which may reduce risk.

MacBinary has been around a long time and is commonly used by Macintosh users. Encoding to MacBinary adds the resource fork, data fork and Finder information into a single file with a “.bin” extension. Many software titles exist to encode and decode into this format. This file format is not currently identified by DROID and “.bin” extension is common among other formats. Versions 1 & 2 of MacBinary don’t have static headers, making a PRONOM signature difficult. Each file in the data set can be encoded, even the custom icon for one of the folders. The format maintains original filename, even if encoded filename is changed, and also retains original timestamps.

E. Apple Single

Apple Single has all the same benefits of MacBinary, but was not as popular so software titles are more limited. The format is identified by DROID, making ingest easier.

F. Apple Double

Apple Double was ingested alongside the original logical copy having the prefix “.” for all files. Also identified by DROID, exporting and moving back to original system is cumbersome.

Format	Resource Fork	Type/Creator	Individual Files Visible	Identification	Validation	Compression	Common Software	Additional Steps
Disk Image	X	X		X			X	
Logical Copy			X					
Stuffit	X	X		X		X	X	X
MacBinary	X	X	X				X	X
AppleSingle	X	X	X	X				X
AppleDouble	X	X	X	X				X

Figure 1 Table showing each Preservation format method and related properties.

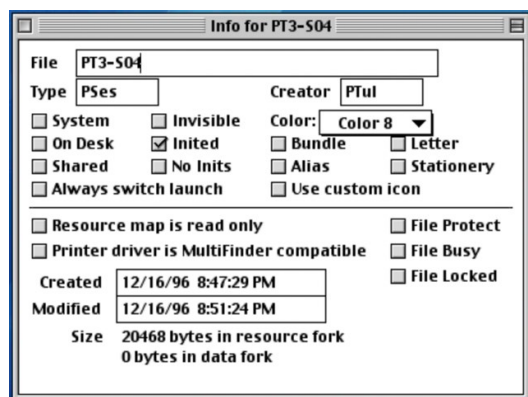


Figure 2 Information view of Pro Tools 3 session.

V. ROUND TRIP

Active digital preservation teaches us these ingest processes are “the actions required to maintain access to digital materials beyond the limits of media failure or technological change” [4]. Part of this process includes ensuring the files can be rendered correctly into the future.

The files ingested into the preservation system were then exported and copied back to a Macintosh for evaluation. The disk image and Stuffit file opened as expected and all content was retained. The MacBinary and AppleSingle files only required decoding using Mac OS X built-in tools [5] or a classic version of Stuffit Expander for MacBinary and a current version of Stuffit Expander for AppleSingle.

The original copied files had the most issues when trying to ingest into our preservation system, often failing entirely. A couple folders have a custom icon, which is a hidden resource fork only file with the name “Icon”, but because of an additional character at the end of the filename, errors occurred even before identification.

VI. CONCLUSION

A typical digital preservation workflow might include capture, extraction, organization, identification, validation, metadata extraction, and storage. Most preservation systems use file systems other than HFS, making this workflow and long-term preservation of this additional information challenging. Once the data from a HFS disk has been captured, determining the best processing and submission format is crucial to successful preservation.

There are many different digital preservation systems available and all may handle files differently. Plus, different preservation policies may influence decisions of preservation format. During our testing, our preservation system could not ingest files with

no data fork and no extension, but when tested with a different system, these files ingest with no issues. This is not necessarily good, as it may not alert us to potential issues.

Many may choose to ingest a disk image, as it requires the least amount of effort to ensure good preservation. If this method is chosen, it might be well to add DFXML or something similar to capture the contents of the disk image into the metadata [6].

AppleDouble is often created automatically and can be preserved fairly easily, and if handled correctly, the split data fork and resource fork can be recombined in the original system. They still may need additional steps to put things back together. One downside for file formats such as Pro Tools sessions, the Apple Double file may be preserved easily, but since the session has zero bytes in the data fork, many systems will not allow the files to be preserved. This is where AppleSingle may be a better choice.

MacBinary is very easy to work with in Classic and Modern Macintosh systems and would be my choice, but currently is not identifiable by DROID.

For our purposes, we feel the best solution is a combination of a disk image along with a MacBinary/AppleSingle.

REFERENCES

- [1] Wikipedia page about Pro Tools. <https://bit.ly/3MvclqT>
- [2] File Format Wiki page about Sound Designer II files. <https://bit.ly/35vkKM4>
- [3] Google Drive containing test files. <https://bit.ly/3hN55Ks>
- [4] DPC Digital Preservation Handbook. <https://bit.ly/3Cml0IX>
- [5] MacBinary and AppleSingle tool descriptions. <https://bit.ly/37a7QU6>
- [6] DANNING. <https://bit.ly/3tG1ZgZ>

A DECADE OF TRUSTWORTHY DIGITAL REPOSITORY CERTIFICATION: YET THERE WAS ONE

Jessica Tieman

*U.S. Government Publishing
Office
USA
jtieman@gpo.gov
[0000-0002-9547-0448](tel:0000-0002-9547-0448)*

Lisa LaPlant

*U.S. Government Publishing
Office
USA
llaplant@gpo.gov
[0000-0002-4924-1361](tel:0000-0002-4924-1361)*

David Walls

*U.S. Government Publishing
Office
USA
dwalls@gpo.gov
[0000-0001-5668-0402](tel:0000-0001-5668-0402)*

Abstract – As the only institution in the world with an ISO 16363:2012-certified repository, the U.S. Government Publishing Office (GPO) now sets forth to pursue assessment under CoreTrustSeal to supplement its internationally recognized certification. While ISO 16363:2012 is still recognized as the “gold standard” for digital preservation repository certification, GPO believes additional assessment will serve to mitigate risks related to the lag in digital preservation community adoption of formal certification and potential future unavailability of ISO 16919:2014 accredited certification bodies to administer audits. This short paper presents GPO’s work in progress to pursue this secondary form of assessment and GPO’s observations and lessons learned as an ISO 16363-certified repository thus far.

Keywords – trustworthy, digital repositories, audit, certification, standards

Conference Topics – community

I. BACKGROUND

Trustworthy Digital Repository Assessment has been a strategic priority at the United States Government Publishing Office (GPO) since the inception of its digital repository system in 2009. At the time, GPO’s GovInfo digital repository (formerly referred to as GPO’s Federal Digital System or FDsys) was designed from the ground up based upon ISO 14721: Reference Model for an Open Archival Information System (OAIS). GPO officially announced its commitment to pursue formal ISO 16363:2012 certification a year after the process for accrediting auditors to perform such audits and grant certifications was established under ISO 16919:2014.

In 2015, GPO initiated preparation for ISO 16363 certification while participating in the National Digital Stewardship Residency (NDSR) Program, hosted by the Library of Congress and the Institute of Museum and Library Services (IMLS). Through this program, GPO obtained a resident to perform the process of collecting and/or preparing all of the necessary documentation and internal-readiness assessments. This included looking at organizational infrastructure, digital object management, and security/risk infrastructure. Because the organizational infrastructure for GPO’s GovInfo digital repository spans multiple GPO organizational units, the resident primarily ensured that policies and procedures were explicit about activities across the units. The internal assessment included organizing narrative responses and collecting all relevant documents and evidence to support each of the 109 criteria of the ISO standard. In 2016, GPO developed and released a Request for Information (RFI). The purpose of the RFI was to elicit information and to better understand the auditing processes and certification opportunities for GovInfo under ISO 16363:2012 accredited certification organizations and to identify organizations that could perform the audit. Next, GPO released a Request for Proposal (RFP). The purpose of the RFP was to select an accredited certification organization to conduct the external audit of GPO. GPO then awarded the opportunity to perform the ISO 16363 audit to Primary Trustworthy Digital Repository Authorisation Body Ltd (PTAB). In December 2018, GPO made history by becoming the first organization

in the United States and second organization in the world to achieve certification as a trustworthy digital repository.

II. ISO 16363 POST-CERTIFICATION OBSERVATIONS

Since 2018, GPO has benefitted greatly as a result of attaining certification under ISO 16363:2012. Official, third-party recognition of GPO's digital repository as trustworthy and digital preservation community standards and best-practices compliant has bolstered trust in GPO's capability to leverage current technology, effectively mitigate long-term risks to digital objects, and operate a large-scale program that meets the needs and expectations of its Designated Community which generally includes, but is not limited to, Congress, Federal agencies and organizations, and the Federal Depository Library community. For these stakeholders, ISO 16363 certification is the only current method of repository certification that ensures a transparent process and removal of auditor bias. As a United States Federal Government institution, it is essential that the audit process is of the highest established credibility to maintain the integrity of the certification.

Despite the benefits GPO specifically receives as a Federal institution for maintaining this certification, no other institutions have publicly announced an intent to follow GPO's lead. A decade after ISO 16363:2012 was published, GPO remains the only repository to maintain its certification which requires renewal every 3 years and annual surveillance audits, and PTAB remains the only known accredited certification body. GPO has observed that many institutions feel underprepared to pursue the full process of repository certification, or they have concerns about gaining high-level administrative support to undergo such an extensive process. Many institutions may see the level of effort GPO expended to prepare for the audit and have hesitations about the human resources needed for organizing all of the required documentation. Repositories may also be unsure if the costs of ISO certification are fully understood or if funds are reliably available. It may also be challenging for an institution to agree upon a definition of its Designated Community. This is essential, as a repository's efficacy is effectively defined by its ability to meet the needs of its Designated Community; without thoroughly documenting those needs and expectations, it may be very difficult to provide evidence of trustworthiness under ISO 16363:2012. Regardless

the reasons many repositories have expressed for not pursuing full certification, GPO may be taking on some level of risk by prioritizing this form of certification if the rest of the digital preservation community continues to hesitate to adopt it in practice. As such, GPO has identified CoreTrustSeal as a secondary form of assessment worthwhile to supplement its existing certification, while staying actively involved with the international digital preservation community.

III. PURSUING CORETRUSTSEAL

The Core Trustworthy Data Repository Requirements [1] were developed by the DSA-WDS Partnership Working Group on Repository Audit and Certification, a Working Group (WG) of the Research Data Alliance [2]. According to DSA and ICSU-WDS, "The goal of the effort was to create a set of harmonized common requirements for certification of repositories at the core level, drawing from criteria already put in place by the Data Seal of Approval (DSA) and the ICSU World Data System (ICSU-WDS). An additional goal of the project was to develop common procedures to be implemented by both DSA and ICSU-WDS. Ultimately, the DSA and ICSU-WDS plan to collaborate on a global framework for repository certification that moves from the core to the extended (nestor-Seal DIN 31644), to the formal (ISO 16363) level." As such, it may seem duplicative or redundant for an institution like GPO to see value in pursuing this "core" level of assessment when GPO is already maintaining the "extended" level of assessment. However, GPO sees multiple benefits to achieving a dual model of assessment:

- Ensures the maintenance of at least one certification at any given time in the even that ISO 16363 accredited bodies are unavailable or other unforeseeable factors pose availability concerns for the performance of ISO 16363 audits
- Increases GPO's involvement in a professional community of over 100 international repositories pursuing "core" assessment but have not yet committed to "formal" assessment, including over 10 Federally operated digital repositories in the United States
- Provides potential opportunities for GPO to serve on CTS peer review boards and present or publish on the experience of attaining both forms of certification
- May allow for GPO to more directly interact with other digital repositories and encourage the

broader professional community to pursue formal certification beyond “core: assessment with GPO as a model of feasibility and success

IV. OPPORTUNITIES FOR GPO AND THE DIGITAL PRESERVATION COMMUNITY

One unique difference between CoreTrustSeal and ISO 16363:2012 certification that GPO appreciates is the requirement of CoreTrustSeal-approved applications to be made publicly available following certification. This requirement to make materials publicly available is one opportunity for GPO to more publicly share documentation and procedures with the digital preservation community in hopes of creating more transparency about the level of effort required to operate an ISO standards compliant repository. This may encourage more repositories to more willingly pursue ISO certification. Likewise, it may encourage other Federal institutions to adopt a more comprehensive view of what “Government Information” is and how text-based, or largely PDF-based repository collections, are still data collections. GPO additionally saw value in CoreTrustSeal over other alternatives, such as DIN31644⁴, as 10 United States Federal Government institutions have already pursued CoreTrustSeal, placing GPO within an existing national community of Federal information stewards.

Additionally, prior to beginning its ISO 16363:2012 audit, GPO participated in a high-level training course offered by PTAB (<http://www.iso16363.org/courses/>) in order to gain a better understanding of its own preparedness for an audit. Resources like this may be a great option for repository managers to better understand the level of effort required for repository certification against the ISO standard. Potentially, an effort like GPO’s to maintain dual certifications can encourage more training opportunities to become available to repositories that are interested in moving from “core” to “extended” models of certification.

By participating in the CoreTrustSeal process, GPO may also be better positioned to engage with other Federal institutions that have data and information repositories who may have difficulty navigating the ISO 16363:2012 standard and turning its requirements into actionable procedures. For instance, GPO’s participation in both models of assessment might help make existing community resources, such as the Digital Preservation Storage Criteria [3] more easily actionable in the context of

self-assessments for preservation infrastructure by being able to more closely examine GPO’s infrastructure practices as made publicly available through the CoreTrustSeal process.

REFERENCES

- [1] Intro to Core Trustworthy Data Repositories Requirements, CoreTrustSeal.
https://www.coretrustseal.org/wp-content/uploads/2017/01/Intro_To_Core_Trustworthy_Data_Repositories_Requirements_2016-11.pdf
- [2] Case Statement for the RDA Working Group and the full set of Working Group members, Members of the Working Group that created this document included the following individuals representing the Data Seal of Approval and ICSU World Data System: Michael Diepenbroek, Ingrid Dillo, Rorie Edmunds, Francoise Genova, Li Guoqing, Wim Hugo, Hervé L’Hours, Jean-Bernard Minster, Mustapha Mokrane, Lesley Rickards (Co-Chair), Paul Trilsbeek, Mary Vardigan (Co-Chair).
<https://www.rd-alliance.org/groups/repository-audit-and-certification-dsa%E2%80%93partnership-wg.html>
- [3] Digital Preservation Storage Criteria, Sibyl Schaefer Nancy McGovern, Andrea Goethals, Eld Zierau, Gail Truman. DOI 10.17605/OSF.IO/SJC6U. Last Updated: 2021-09-08.
<https://osf.io/sjc6u/>
- [4] DIN 31644, 2012 Edition, April 2012 - Information and documentation - Criteria for trustworthy digital archives.
https://global.ihs.com/doc_detail.cfm?document_name=DIN%2031644&item_s_key=00585595

ACT NOW, LATE OR NEVER:

Make Digital Objects (more) archivable early in their life cycle?

Katharina Markus

ZB MED – Information
Centre for Life Sciences
Germany
Markus@zbmed.de

Yvonne Tunnat

ZBW – Leibniz Information
Centre for Economics
Germany
y.tunnat@zbw.eu

Abstract – Newly acquired or published objects might be corrupt or might not conform to the archive's best practices. In some cases the library or archive even has to ask the data provider for replacements. The advantage of a pre-archival workflow to detect or prevent problems early in institutional data processing is depicted in this paper.

Keywords – Archivability, Digital Preservation, Validity, PDF format

Conference Topics – Exchange; Innovation

I. INTRODUCTION

When archives obtain objects and prepare them for ingest into the archival system, they follow digital preservation best practices. The archive department usually is responsible for the object preparation step which is sometimes conducted at a time significantly after the institution obtained the objects. But to consider best practices and to conduct this step earlier, e. g. directly after acquisition, might save time and curation effort later on. It may also allow preservation of information that is lost otherwise. This paper will introduce two use cases at the institutions ZBW – Leibniz Information Centre for Economics and ZB MED – Information Centre for Life Sciences. The institutions obtain control over objects relevant for this paper at two processing stages – publication (ZB MED) and acquisition (ZBW). This paper analyses in a qualitative manner the benefit of introducing preservation best practices into the early processing steps of publication and acquisition. The analysis is based on an implemented new workflow (ZBW) and implementation planning (ZB MED). It can serve as a basis for other institutions in similar situations where the archives deal with high

amounts of objects that also require relatively high amounts of curation.

II. BACKGROUND AND RELATED WORK

The preservation community defined general best practices for preservation [1]–[4]. One example for best practices used as quality criteria for objects is shown as follows: The German National Library (DNB) defines five different ingest levels, which increase the quality of files regarding preservation with each level from data integrity through identification of file formats, unrestricted access to files for DNB, available technical metadata and, finally, to valid files according to format validation. The DNB conducts quality checks during ingest and rejects files if integrity is not provided and formats are not identifiable [5].

Curating objects according to preservation best practices during transfer to archives may result in huge curation efforts for archives and archive departments, stalling objects in the pre-ingest or ingest step [6]. Efforts might be due to obtaining necessary rights [7], [8], dealing with non-standard and inconsistent infrastructure and data structure as well as missing files [6], [8], [9] or simply defective data [10]. Personal communication of the author Yvonne Tunnat with various members of the digital preservation community shortly after publishing a blog post regarding curation efforts and relevant tools used during acquisition shows interest in this topic as well.

Increasing conformance with these best practices was termed for this paper as increased *archivability* [11], which was defined by Banos and Manolopoulos

[11] as “whether [a website] has the potential to be archived with completeness and accuracy” but is used in this context for objects in a broader sense. The processing of objects to make them more archivable, like detecting and repairing defective files, were defined as *actions increasing archivability*. Institutions can conduct these actions during any step of the object processing workflow (from object creation until archiving, see fig. 1). For this paper, the authors divided processing steps and actions into those which are part of a *pre-archive* workflow (WF) and those that are part of the archive WF, where specifically pre-ingest and ingest steps are located. In the pre-archive WF, departments other than the archive are responsible for the object processing and actions are localized earlier in the processing (see fig. 1).

As far as the authors were able to determine, literature rarely analyses which archivability increasing actions are best conducted in early processing steps by departments other than the archive. Preservation best practices are implicitly targeted at archive departments and the archives are recommended to take “an active role in [digital

information’s] maintenance early in its life cycle” [12]. Still, the authors found the prospect of early curation actions mentioned in the context of digitization projects [13]–[15], research data [12] and web archiving [11]. Skinner and Schultz [13] address digitized objects but also consider born-digital material. They devote a chapter to preservation best practices for digitized objects as part of creation and acquisition, in which they recommend the set-up of an inventory, the definition and documentation of recommended file formats, metadata and data structures, the generation of checksums and establishment of explicit permission to preserve the objects.

Selected best practices for ZB MED and ZBW with relevance for this paper are similar: consistent data structure, recommended file formats, among them PDF/A with embedded open fonts, valid objects and metadata standards established in the research community. The actions that increase archivability are related to these best practices. The authors assume that introducing these actions in early object processing steps results in a reduction of total curation effort for the institution (see fig. 1).

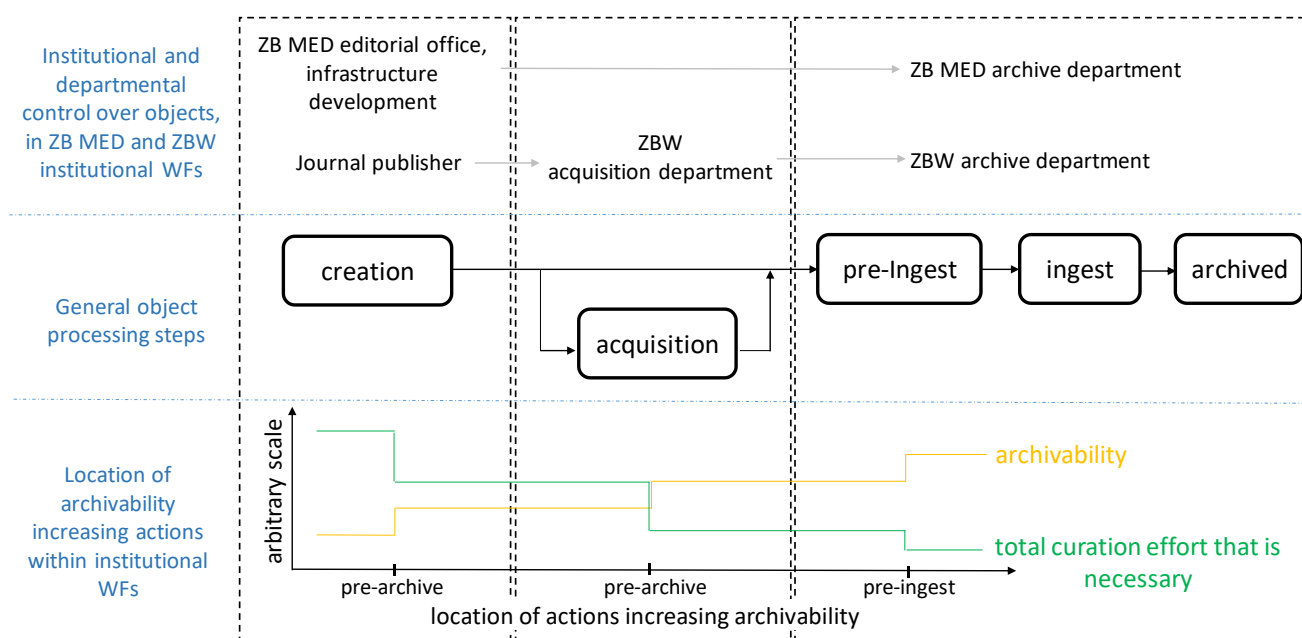


Figure 1 Depicted are object processing steps (from object creation to archived), the respective steps where ZB MED and ZBW obtain control of the objects, as well as three possible locations of archivability increasing actions (pre-archive during publishing, pre-archive during acquisition or pre-ingest within the archive department). The diagram shows the result of conducted archivability increasing actions as an increase in archivability and a decrease in total curation effort. Increase and decrease have been determined as a result from qualitative analysis instead of quantitative measurements and therefore the scale is arbitrary.

III. METHOD

To answer the research question "Is generating archivable digital objects early in their life cycle worth the (staff and process development) effort?", this short paper uses the method "actual practice" (as opposed to best practice, as examinations about this sub-topic in Digital Preservation seem to be rarely addressed in literature), analyzing available literature (chapter II) and two use cases: description and evaluation of ZB MED's pre-archival WF (chapter IV. A) and ZBW's pre-archival object processing (chapter IV. B). In the referenced blog posts the used tools are described and more practical points of the workflow are examined. This paper uses a qualitative analysis regarding benefits of early archivability actions and assesses the impact on the objects which have to be archived.

IV. USE CASES

A. ZB MED Use Case

ZB MED provides several publication services to the life science community, among them the PUBLISSO Gold publication portal [16] and, in collaboration with the Association of the Scientific Medical Societies in Germany, the German Medical Science (GMS) publication portal [17]. Publications on these platforms are intended to be archived in ZB MED's own archive, which is a separate system. Since a high number of data sets are archived retroactively, at a significantly later time than their publication date, various challenges became apparent when the archive collected data sets from the GMS portal. Still, ZB MED has control over formal quality assurance (QA) of publications on the platforms. Accordingly, it can incorporate various suggestions from its own archive department into publication processes and infrastructure in order to increase archivability. A summary of possible actions improving archivability during publication steps follows. These are in various states of implementation.

A well-known challenge for archiving is clearing necessary rights [7], [8]. In addition to rights of objects, fonts can also be copyrighted which hinders embedding during PDF/A migration. Using open fonts during publication allows later embedding in PDF/A without reviewing and verifying the usage rights of the used font. In case the font used during publication does not allow embedding by the archive, the institution can resort to another font

that does. But this change in fonts may lead to changes in content display, which requires additional QA, and therefore effort from the archive. Optimally, PDF/A with already embedded fonts is used for publication.

A significant challenge during ingest relates to the publication's data structure. If the structure is inconsistent, the institution preparing the data packages for the archive (data provider or the archive itself) cannot rely on an entirely automated workflow. Instead, it needs to identify exceptions and map the different data structures of provided objects to the archive's data structure. In the established workflow at ZB MED, about 0.1% publications contained exceptions that resulted in additional handling. While changes to data structures over time are probably unavoidable, the publisher may help with later data transfer by documenting a data structure schema as well as exceptions and new versions alongside respective objects. This may be useful not only for a transfer into archives but also for exit scenarios.

At the object level, recommendations of file formats that are more or less suited for digital preservation are well known in the digital preservation community [18]. Close collaboration with editorial offices helps with communicating these as best practices to authors. Additionally, the introduction of validation and a documentation of publication versions might also offer opportunities: The benefit of pre-archive validation is detailed in the ZBW Use Case (see chapter IV. B). Versioning of identifiers in metadata when new versions of a publication are generated allows for automated or partly automated update workflows connecting platform and archive. This, in turn, should reduce efforts of communicating updates between staff of different departments while also decreasing risks of human error.

Going beyond the purely technical level, markup languages can also serve as metadata standards as part of the object itself. As text publications are not necessarily restricted to the PDF format but become increasingly reusable for machines when published as XML, selecting subject-specific markup languages according to preservation best practices becomes relevant as well. Examples of well-known subject-specific markup languages are bioschemas [19] or MathML [20]. Recommending these standards for publications with preservation best practices in mind while also consulting the scientific community can evolve into a new task for subject-specific archives. In case of machine readability of molecular

structures of chemical compounds, ZB MED researched open, well-used and maintained markup languages. They consulted FAIRsharing [21], taking into account referenced maintainers, number of databases that use the standard and whether the standard is open. They investigated usage of the standard in popular software used in the research community, for which they referred to people with a background in chemistry and related fields. A preliminary selection resulted in openSMILES [22] as the preferred standard.

In general, integrating the above-mentioned recommendations into object creation processes is expected to reduce total curation efforts as an automated object transfer is enhanced and likelihood for later handling is decreased. For further work, ZB MED attempts the implementation of the mentioned suggestions as far as technically possible.

B. ZBW Use Case

The ZBW – Leibniz Information Centre for Economics provides digital documents like articles and research papers on its many presentation platforms like EconStor [23] or other instances which are all available via EconBiz [24].

The ZBW established Digital Preservation in 2015 to ensure long-term availability for their hosted content. The Digital Preservation Archive is a dark archive, based on the System Rosetta developed by Ex Libris [25]. All the content is presented to the users by other representation platforms, mostly based on DSpace [26].

However, Digital Preservation is the last step in the object processing pipeline, just as it is at ZB MED (see chapter IV. A). For most workflows that presents no problem, as the material on the DSpace platform is published immediately after acquisition and the ingest is done the night after.

For objects acquired under National and Alliance Licences, though, the hosting on ZBW servers and therefore the ingest to the Rosetta archive is done months or even years after the acquisition of the material.

After such a long time, the data providers (usually publishers like Emerald, De Gruyter and Elsevier) have long since moved on to other projects, so that it is time-consuming and sometimes impossible to get a replacement if parts of the data are missing or corrupted.

Therefore, the ZBW staff responsible for the acquisition has established an automated preliminary data check workflow pre-archive. The

publishers deliver the data, in most cases a large amount of PDF files, in Zip folders.

During the past years, the pre-archive workflow included:

- Unpacking the zipped files
- Integrity check (via checksums)
- Completeness check

Newly implemented into the pre-archive workflow are:

- Checking for password protection (which would impede data migration)
- Running the PDF files through tools to check for errors

The tools used are Grep, PDFinfo and, mainly: ExifTool [27]. The workflow in detail, the implementation of the workflow, the staff time used for daily work and the handling of different ExifTools error messages are described in detail in an OPF blog post published in February 2022 [28].

Tests have shown that certain error messages hint at the PDFs not being archivable, sometimes not even accessible for the users. For those, the ZBW acquisition department can ask for a replacement directly after acquisition. As many PDF files are password-protected, the ZBW rights department and the data providers have agreed to delete the password-protection. To accomplish this, the ZBW acquisition department has set up another automated workflow.

As only open source tools are used, the invested resources are calculated as curation effort, specifically as staff time of the involved departments acquisition and archive. This includes:

- copying the PDF files to the hotfolder where the tools conduct their actions
- preliminary judgment of the findings (especially if a new error occurs, which has not been evaluated yet)
- if a new error occurs, the ZBW archive department checks if the affected PDF files can be migrated to PDF/A-2b
- if a new error occurs, the acquisition department performs a manual check to see if the PDF is accessible. This is also done for some errors, such as "PDF header not at beginning of the file".

The curation effort for a bulk of 1,000 PDFs for the newly established workflow, in average, requires an

hour of staff time. This includes error-handling and asking for replacement when a PDF file is corrupted. This workflow now takes up more time during acquisition due to additional actions that aim for better archivability.

The ingest into the archive, in comparison, is now fully automated and usually does not need extra staff time. Only if errors occur does the ZBW archiving department have to work on these and spend staff time. Nevertheless, the new WF is worth the extra time during acquisition, as corrupted data is detected early and can be replaced, whereas it would be permanently lost to the archive or in general otherwise. No matter how good the digital curation workflows are: if the data is too corrupted to be repaired or even lacking contents to begin with, there is nothing to be done about it at a later stage. Either the contact to the publisher has gone cold, so that the ZBW acquisition team cannot get hold of the data provider and thus, the object anymore. Or the ZBW and the publisher negotiated that the contents can be hosted (and thus archived) when the data is no longer available from the publisher's websites. In this case, if defective data is discovered a significant time after archiving and the publisher does not provide it anymore either, the content is lost for good.

As a side effect: The data providers have so far been grateful for the information about corrupt files, as they also want to offer a high data quality on their platforms for their users.

The ZBW staff established these workflows quite recently. In the future the acquisition department will evaluate the workflows regularly and, if necessary, extend or alter them.

C. Tools

While ingesting the data into the ZBW Rosetta Archive, several tools are used: DROID, JHOVE, NLNZ Metadata Extractor, just to name the most important. These tools extract technical metadata like the file format including the format version, detect password-protected files and identify basic information about size, creation date and a lot of other information useful to ensure long term-availability.

As a side effect, the archive department usually detects files that are not accessible or otherwise corrupted.

During the pre-archive workflow after acquisition, the acquisition department uses Grep, PDFinfo and

ExifTool (see chapter IV. B). The usage of the tools is regularly evaluated, e. g. via tool benchmarking; comparing which tool is best suited for a certain task, mostly with regard to file validation. This has been done thoroughly for the file formats:

- TIFF [29]
- JPEG [30]
- GIF [31]
- PDF [32].

As tools and their usage frequently evolve, close preservation watch is essential. For instance, in December 2017, when the ZBW archive department examined the validation tools for PDF, ExifTool was not even considered, although it would have been of use back then. ZBW staff did not include it in the evaluation only because they did not yet know about the tool.

For some use cases, tools could also be inappropriate, as they take too long, give too many false alarms (false positives) or their validation is too thorough for pre-archive needs, like JHOVE for PDF [33].

The tools ZB MED uses for preparation of objects for archiving in its present workflow are a self-developed Submission Application (SubApp) as well as JHOVE and veraPDF in pre-ingest processing. The archive department is responsible for operating these. During the subsequent ingest the archive department uses further tools, the same as ZBW (see above) which are not detailed here. The SubApp generates data packages and detects exceptions in the data structure, whereas JHOVE so far detected invalid image files during pre-ingest processing. Exceptions and invalid files require individual processing by the archive and the editorial office, as part of the otherwise automated pre-ingest workflow. The archive department is in close contact with the publishing platforms regarding analysis and evaluation of tools and changes to objects and WFs.

V. FINDINGS AND SUMMARY

As shown in the use cases, ZBW and ZBW identified several actions which, when implemented in pre-archive workflows, may reduce curation efforts presently or in the future. As ZB MED detects various exceptions during pre-ingest with their present WF, they expect better automation if data structure is documented early in a stringent way. As additional opportunities, ZB MED identified the use of open

fonts for PDF publications, markup languages suited for preservation as well as documenting object versions in a standardized way. ZBW discovered corrupted data well after acquisition with their old WF. The new WF contains validation with ExifTool pre-archive, as part of acquisition. This allows early detection of invalid, password-protected and corrupted files and subsequent exchange of files when contact to the provider is still established. With these analyses, the authors expand on the recommendation by Skinner and Schultz [13] with specific tools (ExifTool) and proposed implementations (e. g. open fonts for PDF publications) based on actual practice.

Both institutions come to the conclusion that early incorporation of these best practices, tools and actions seems to prevent significantly higher efforts later. "Later" meaning here, if archivability increasing actions are conducted a significant time after publication or acquisition. The reasons are twofold: first, when the institution is still in contact with an external data provider, obtaining correct versions of files (corrupt, invalid) and clarifying rights (password protected) requires less effort than re-establishing contact months or years after data provision was concluded. Additionally, at this point in time the department sometimes can still obtain data that would be lost to the archive otherwise. Secondly, the departments involved in publishing already process objects at an individual level. Additional curation at that stage requires less effort than stopping automated archiving processes later on. Nonetheless, not every curation action can or should be implemented in pre-archival WFs. Therefore, the archive departments selected the above-mentioned actions and best practices in exchange with editorial offices and the acquisition department. They maintain contact with these pre-archive departments as well, re-evaluating workflows while also taking organizational and technical conditions into account. Still, this evaluation of prospective and actual implementations described here might help with the scaling of archiving workflows, not just for the institutions mentioned in this paper but for others as well, because all are faced with increasing amounts of all kinds of materials that need to be archived.

REFERENCES

- [1] J. Mitcham and P. Wheatley, "Digital Preservation Coalition Rapid Assessment Model (DPC RAM) (Version 2.0)." [Online]. Available: <http://doi.org/10.7207/dpcram21-02> (accessed Feb. 14, 2022)
- [2] A. Beking *et al.*, "Digital Curation Decision Guide," *NDSA Publ.*, Dec. 2020, doi: 10.17605/OSF.IO/Q8C47. (accessed Feb. 14, 2022)
- [3] E. Faulder *et al.*, "Digital Processing Framework," report, Aug. 2018. Accessed: Jan. 18, 2022. [Online]. Available: <https://ecommons.cornell.edu/handle/1813/57659>
- [4] CoreTrustSeal Standards and Certification Board, "CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022," Nov. 2019, doi: 10.5281/zenodo.3638211. (accessed Feb. 14, 2022)
- [5] Deutsche Nationalbibliothek, "Spezifikation der Dateiformat-Policy für die Sammlung von Netzpublikationen der Deutschen Nationalbibliothek. Version 1.0." Oct. 24, 2012. [Online]. Available: <http://nbn-resolving.de/urn:nbn:de:101-2012102408> (accessed Feb. 14, 2022)
- [6] S. Morrissey and A. Kirchhoff, "Managing Preservation Costs with managed Ingest: The Portico Straight-to-Ingest Project," in *Proceedings of 16th International Conference on Digital Preservation*, Amsterdam, Sep. 2019, pp. 317–322. doi: 10.17605/OSF.IO/VW7RJ. (accessed Feb. 14, 2022)
- [7] Rosenthal, David, "Why Preserve E-Journals? To Preserve The Record," Jun. 10, 2007. <https://blog.dshr.org/2007/06/why-preserve-e-journals-to-preserve.html> (accessed Dec. 26, 2021).
- [8] B. Sprout and M. Jordan, "Distributed digital preservation: preserving open journal systems content in the PKP PN," *Digit. Libr. Perspect.*, vol. 34, no. 4, pp. 246–261, Jan. 2018, doi: 10.1108/DLP-11-2017-0043. (accessed Feb. 14, 2022)
- [9] R. Arora, M. Esteva, and J. Trelogan, "Leveraging High Performance Computing for Managing Large and Evolving Data Collections," *Int. J. Digit. Curation*, vol. 9, no. 2, Art. pp. 17–27, Oct. 2014, doi: 10.2218/ijdc.v9i2.331. (accessed Feb. 14, 2022)
- [10] B. Sprout and S. Romkey, "A Persistent Digital Collections Strategy for UBC Library," in *Proceedings of the Memory of the World in the Digital Age: Digitization and Preservation Conference*, Vancouver, British Columbia, Canada, 2013, pp. 256–268. [Online]. Available: http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/VC_Sprout_Romkey_26_B_1540.pdf (accessed Feb. 14, 2022)
- [11] V. Banos and Y. Manolopoulos, "A quantitative approach to evaluate Website Archivability using the CLEAR+ method," *Int. J. Digit. Libr.*, vol. 17, no. 2, pp. 119–141, Jun. 2016, doi: 10.1007/s00799-015-0144-4. (accessed Feb. 14, 2022)
- [12] D. R. Bleakly, "Long-term spatial data preservation and archiving: What are the issues," *Sand Rep. SAND 2002*, vol. 107, 2002, doi: 10.2172/793225. (accessed Feb. 14, 2022)
- [13] K. Skinner and M. Schultz, *Guidelines for Digital Newspaper Preservation Readiness*. Atlanta, GA: University of North Texas Libraries, UNT Digital Library, <https://digital.library.unt.edu>, 2014. Accessed: Jan. 07, 2022. [Online]. Available: <https://digital.library.unt.edu/ark:/67531/metadac282586/>
- [14] K. Dohe and R. Pike, "Integration of Project Management Techniques in Digital Projects," in *Project Management in the Library Workplace*, vol. 38, Emerald Publishing Limited, 2018, pp. 151–166. doi: 10.1108/S0732-06712018000038013. (accessed Feb. 14, 2022)
- [15] Deutsche Forschungsgesellschaft, "12.151 DFG-Praxisregeln 'Digitalisierung' [12/16]." Accessed: Feb. 10, 2022. [Online].

Available:
http://www.dfg.de/formulare/12_151/12_151_de.pdf
(accessed Feb. 14, 2022)

and the Ugly," in *iPRES Japan*, Japan, Kyoto, Sep. 2017, p. 11.
Accessed: Feb. 21, 2022. [Online]. Available:
<https://files.dnb.de/nestor/weitere/ipres2017.pdf>

- [16] "PUBLISSO Gold Open Access publication portal." <https://www.publisso.de/en/publishing/> (accessed Feb. 21, 2022).
- [17] "German Medical Science publication portal." <https://www.egms.de/dynamic/en/index.htm> (accessed Feb. 21, 2022).
- [18] C. Arms and C. Fleischhauer, "Digital Formats: Factors for Sustainability, Functionality, and Quality," *Arch. Conf.*, vol. 2005, no. 1, pp. 222–227, Jan. 2005.
- [19] A. J. Gray, C. A. Goble, and R. Jimenez, "Bioschemas: from potato salad to protein annotation," presented at the International Semantic Web Conference (Posters, Demos & Industry Tracks), 2017.
- [20] R. Miner, "The importance of MathML to mathematics communication," *Not. AMS*, vol. 52, no. 5, pp. 532–538, 2005.
- [21] S.-A. Sansone *et al.*, "FAIRsharing as a community approach to standards, repositories and policies," *Nat. Biotechnol.*, vol. 37, no. 4, Art. pp. 358–367, Apr. 2019, doi: 10.1038/s41587-019-0080-8. (accessed Feb. 14, 2022)
- [22] "Simplified Molecular Input Line Entry Specification Format (SMILES)." <https://doi.org/10.25504/FAIRsharing.qv4b3c> (accessed Feb. 14, 2022)
- [23] ZBW, "EconStor." <https://www.econstor.eu/> (accessed Feb. 14, 2022).
- [24] ZBW, "EconBiz." <https://www.econbiz.de/> (accessed Feb. 14, 2022).
- [25] Ex Libris, "Rosetta." <https://exlibrisgroup.com/products/rosetta-digital-asset-management-and-preservation/> (accessed Feb. 14, 2022).
- [26] dSPACE, "DSpace." <https://www.dspace.com/en/pub/home.cfm> (accessed Feb. 14, 2022).
- [27] Harvey, Phil, "ExifTool." <https://exiftool.org/> (accessed Feb. 14, 2022).
- [28] Tunnat, Yvonne, "Validation with ExifTool: Quick and not so dirty," *Open Preservation Foundation Blogs*, Feb. 04, 2011. <https://openpreservation.org/blogs/pdf-validation-with-exiftool-quick-and-not-so-dirty/?q=1> (accessed Feb. 14, 2022).
- [29] Tunnat, Yvonne, "TIFF format validation: easy-peasy?," *Open Preservation Foundation Blogs*, Jan. 17, 2017. <https://openpreservation.org/blogs/tiff-format-validation-easy-peasy/> (accessed Feb. 14, 2022).
- [30] Tunnat, Yvonne, "Error detection of JPEG files with JHOVE and Bad Peggy – so who's the real Sherlock Holmes here?," *Open Preservation Foundation Blogs*, Nov. 29, 2016. <https://openpreservation.org/blogs/jpegvalidation/> (accessed Feb. 14, 2022).
- [31] Tunnat, Yvonne, "Good GIF hunting: JHOVE's GIF validation skills," *Open Preservation Blogs*, Dec. 05, 2017. <https://openpreservation.org/blogs/good-gif-hunting/> (accessed Feb. 14, 2022).
- [32] Tunnat, Yvonne, "JHOVE – the one and only PDF validator," *Open Preservation Foundation Blogs*, Dec. 19, 2017. <https://openpreservation.org/blogs/jhove-the-one-and-only-pdf-validator/> (accessed Feb. 14, 2022).
- [33] M. Lindlar, C. Wilson, and Tunnat, Yvonne, "A PDF Test-Set for Well-Formedness Validation in JHOVE - The Good, the Bad

THESE CRAWLS CAN TALK

Context Information for Web Collections

Susanne van den Eijkel

National Library of the Netherlands
Susanne.vandenEijkel@KB.nl
[0000-0002-8221-869X](tel:0000-0002-8221-869X)

Daniel Steinmeier

National Library of the Netherlands
Daniel.Steinmeier@KB.nl

Abstract – The National Library of the Netherlands has been harvesting the web since 2007. As is well known, an archived website is fundamentally different from a website on the live web. As a digital repository we create archival objects in the process of archiving, especially in the case of web harvesting since a website does not have clear boundaries. This means we define the limits ourselves. So how do we provide researchers with information on the choices we made during the process of selection, harvesting and ingest? How can we prove the integrity and authenticity of our web collections to our designated community? Based on the findings of Maemura et al. in 2018, three categories of information should be available for researchers of web collections: scope elements, process elements and context elements. This will help researchers understand what is present in a collection, what curatorial decisions have been made in the process and the reason behind the creation of the collection. In this article we will describe three documentation initiatives related to our web collections that we think could be seen as implementations of these three types of documentation elements.

Keywords – Preservation, Metadata, Context Information, Web Collections

Conference Topics – Exchange; Community

I. INTRODUCTION

As National Library of the Netherlands (KBNL) we are determined to guarantee long-term accessibility to our cultural heritage. The first thing people think of when they hear 'library' is books and indeed we have a lot of those, but for fifteen years now it has also been part of our mission to preserve the web. KBNL intends to collect a copy of every publication that was made in the Netherlands or is about the Netherlands, as defined in our content strategy [1]. In 2019, a 'digital first'-principle was added. This means that we only save a physical copy of the

content if there is no digital one available. Websites are considered publications just as much as books or articles. Therefore, selecting and archiving websites has the same priority as archiving books, newspapers and journals from our regular publishers.

In the field of web archiving, the focus is often on tools for harvesting certain types of content, or innovative technical solutions for providing access to web collections. However, access also means keeping material understandable for our users [2]. An important target group within our user community consists of researchers. How do we provide for this group when we want to keep material understandable? A lot of research about the web is based on statistical information extracted from data present in web archives. However, for these statistics to be reliable researchers should be able to understand the choices made by archivists during the selection and creation of the web collection. This is information that cannot simply be extracted or generated by tools but should preferably take the form of documentation written by the people making the day-to-day business decisions necessary in the process of archiving the web.

In our preservation plan we outlined three key themes important for long-term preservation: integrity, authenticity and long-term accessibility [3]. For us, integrity revolves around all measures necessary for ensuring completeness of the objects and the collection itself. The concept of authenticity requires information about provenance, the producer and the intention of the archival object. Both integrity and authenticity are important for considering archival material long-term accessible. In digital archiving a fair amount of trust is placed in

technical measures that are supposed to prove authenticity. However often non-technical factors like policy documents play an important role in proving authenticity in actual practice [4]. It is important to consider that many guidelines in digital preservation can be implemented using non-technical solutions. An example of this would be documentation meant to enhance understandability of the objects and the collection, described in the Open Archival Information System (OAIS) and the Trustworthy Digital Repository (ISO 16363) standards as representation information. After all, business decisions made by operators during archiving are best explained in written documentation and cannot be simply generated.

So, what would be important documentation to capture in the case of web collections? For this paper, we based our examples on the three element types described by Maemura et al. in the article 'If these crawls could talk', namely: scope elements, process elements and context elements [5]. All information on what is and is not included in the web collection is part of the scope elements. Scope is also part of the configuration of the web harvesting software we use. Harvesting is the process of collecting or crawling websites. Without scope settings the crawler would go on crawling the web indefinitely. The scope of the harvester defines boundaries, but this also has impact on the integrity of the resulting collection, so the choices made in the scope need to be explained and documented in detail so researchers will understand to what extent the collection can be considered 'complete'. Not only scope is important in this regard however, also the process of archiving needs to be mapped. This entails the process of harvesting, and the process elements present in the configuration of the harvester, but also the curatorial decisions made after harvesting. For instance, what actions are taken within the scope of doing quality analysis or in case of errors that occur in the process. Finally, an important piece of documentation concerns the context of the collection. Why is a collection harvested and how are targets for harvesting selected? So yes, these crawls can talk, if we make this context information available for our users. Taken together these three categories of documentation will provide researchers with information on the social factors impacting the integrity and authenticity of archived websites in heritage collections.

II. PROCESS ELEMENTS: QUALITY ASSURANCE ON SELECTIVE HARVESTS

In 2006, KBNL started a small collection of websites based on manual selection. This part of the collection is what we call the selective harvest, since it is a curated collection of websites that is deemed important from a cultural perspective, as defined in our collection policy. Every website in this collection has been selected because it represents in some way a part of Dutch language, culture and history. This selective approach is in line with the remit of KBNL, the resources available for web harvesting and legal considerations [6]. KBNL uses the Web Curator Tool (WCT) [7] to collect the selective harvests. This is an open-source tool for managing selective web harvesting for non-technical users. It provides a graphical interface for changing settings and adding basic descriptive metadata, but the underlying crawler software is Heritrix [8].

The web collection of KBNL currently consists of 20,000 websites. Among this number are the special collections. Since 2013, our curators started assembling special collections for the web archive in a similar fashion as the UK web archive. A few examples of these special collections are websites about the commemoration of 200 years Kingdom of the Netherlands, the First World War and the Covid-19 pandemic. This collection continues to grow, as our curators are still actively searching for popular websites or websites that reflect topics currently relevant in Dutch society [9]. What distinguishes a selective harvest from other forms of harvesting, like a domain crawl, is that a selective harvest consists of a relatively small number of websites. This means we are better able to analyse the results of the harvest thoroughly to see whether content is missing. Based on our findings we can alter the settings, in order to harvest a complete collection as possible.

We have documented in an internal manual how quality assurance (QA) is done and how the results of this process are added to the metadata as annotations via the WCT interface. By adding the outcome of the QA process in these notes, it will stay available as context information for internal users and researchers. Every two weeks employees check the content that was harvested in that period. In order to execute QA as effectively as possible, the new harvests are divided into three groups: websites smaller than 1MB, websites bigger than 1GB or with

a runtime longer than twenty-four hours and websites that have a divergent schedule or ran into a limit that we did not expect. Each target should have a schedule that indicates how often a website will be harvested, for example every half year. If a scheduled website contains less content than expected, the website will appear in the selection for QA. We try to find out if and why there is content missing and how to handle this. In the notes in WCT we document the findings and whether the problem was solved.

Our documentation also provides rules for URLs that must be included or excluded, defined as regular expressions. These rules are often linked to problems that were found during QA. Shopping carts, for example, are filtered out, because the crawler gets stuck on add-to-shopping-cart-buttons, resulting in unwanted content and a long runtime. Sometimes the homepages of websites refer to different URLs, even though they are part of the same website. Depending on how deep we wanted to harvest a specific website, we would add the other URL for the same website as include or as a secondary seed and documented it in the notes. In this way it is also clear to future users why earlier harvests were incomplete and differ from later harvests.

These annotations about quality assurance give users insight into the status of websites in our repository. Part of the harvest history as well as the decisions made during QA can be reconstructed with the help of these notes. For example, 'QA 2019 PC OK' means that during a QA-analysis in 2019 a website was picked out for error handling, that the problem has been fixed and that the website can be harvested again without errors. Often a specification follows the annotation in the notes, describing the problem and the solution briefly. It can also happen that a website no longer exists. In that case the annotations will mention 'QA 2021 PC cancelled', followed by an explanation in Dutch and the exact date that we found out that the website was no longer accessible. So, with the help of the documentation on how we handle quality assurance and notes that were added to specific targets, we can better understand the choices that were made in the harvest process such as why settings have been changed or why a website is no longer harvested.

III. SCOPE ELEMENTS: SETTINGS OF CRAWLER SOFTWARE FOR WEB ARCHIVING

The second documentation initiative we want to highlight in this paper is connected to a collection started as preparation for a national domain crawl. The Dutch national domain that we intend to harvest consists of 6.5 million websites, as was identified by KBNLs curator of digital collections. This is a vastly different scale from our selective harvest, which means different tooling is needed and manual quality analysis is ruled out. To prepare for this undertaking, we have been testing crawler tools and appropriate settings on a smaller domain: Frisia. This is a province in the Northern Netherlands, with its own history, culture and language. This province got its own top-level domain, namely .frl, in 2014. This domain is much smaller than the .nl-domain which makes it more suitable for testing crawler tools. On the Frisian websites we see the Frisian identity of the 21st century very clearly, for example in GIFs with Frisian puns. KBNL has selected approximately 10,000 Frisian websites and websites about Frisia. For example, Frisian Wikipedia (.org) and a Frisian news site on the .nl domain, are also included. In order to define the Frisian domain, researchers were asked which websites should be part of this collection, resulting in the current selection. It was a collaboration with researchers of digital humanities and the biggest cultural and heritage institution of Frisia: Tresoar. In a way this collection is still a selective crawl, but the websites will not undergo rigorous quality control afterwards because of the high number of websites harvested. Lessons learned from this regional domain crawl will be used for the Dutch national domain crawl.

For the first tests we decided to use NetarchiveSuite (NaS) [10]. This is an open-source tool, developed by the National Library of Denmark to harvest the Danish web. One of the reasons we decided to use NaS is because this also has a web interface on top of a Heritrix crawler, just like WCT. Testing a new tool and defining a new process of harvesting gave us the chance of rethinking the rationale behind our current harvesting settings and testing out variables to better understand how these have an impact on what is harvested. This resulted in better knowledge of the different settings available in Heritrix. We decided that this information would need to be stored for future reference, because it helps to better understand the collection.

During the harvest, multiple websites can be collected, by providing the crawler with more than one seed as starting point. The settings of the crawler

define the limits for the harvest. They determine, for example, how deep the website is crawled in terms of slashes in the URL (path-depth) and how far the crawler moves (hops) from the starting point (seed) when counting the number of links followed. The whole path of link-hops from the seed up till the current URL is called the discovery path and can be found in the log file after the harvest is done. By looking at the log file and finding patterns between unwanted URLs and types of hop in the discovery path, conclusions can be drawn on how many of the different types of hops should be allowed in the scope [2].

Our test strategy was to try and harvest websites as complete as possible at least insofar as they were still identifiably related to the seed. By using the default settings, we allowed five trans-hops (hops based on things like embed-links) and one speculative hop (hops based on links extracted from Javascript). This resulted in too much unwanted content, like login pages for social media in various languages. We concluded that we could filter out these pages by allowing less hops. We tested this theory by allowing only two trans-hops and no speculative hops. This limit was too strict. Relevant PDF-files and stylesheets for example, on the same domain as the original seed, were missing. To mitigate this, we ran a third test with adjusted settings. Once again, we permitted two trans-hops and a speculative hop, but only if the URL was from the same domain as the seed. This time we saw that the relevant PDF-files were indeed harvested, but with minimal unwanted content.

Now why is it important to document these choices and provide them to users as context information? Determining the correct settings is an intensive process and documenting the decision process means manual work. So why invest in this? It is important for archivists to understand the material in the archives, in order to preserve digital objects and define appropriate preservation strategies. On the other hand, context information is also vital for researchers. With access to context information, they can reconstruct the life cycle of a digital object. Insight into the settings will help them understand the choices that were made in the process. This includes understanding how the original website has been transformed into an archival object. In this way, researchers will be able to determine completeness and authenticity in much the same way as they would do with physical historical sources.

IV. CONTEXT ELEMENTS: COLLECTION DESCRIPTION XS4ALL

Context information, however, can be more than only technical details about the settings of the crawler software. Collection descriptions can also give us important historical context information, such as how the collection was created and why curators thought this collection was important to acquire [11]. A special subcollection within our selective harvest collection are the XS4ALL homepages. XS4ALL was one of the first Internet providers in the Netherlands to provide services that allowed individuals to create their own homepages from 1994 onwards. In 2019, it was announced that the brand name XS4ALL would no longer be used. This meant that original homepages would not be available anymore on the original URLs with the risk that they might disappear from the web during migration, and with that a great online source of early historical web content. This was reason enough for KBNL to take immediate action. Our curators compiled a list of the most important webpages to preserve, based on criteria like historical value, authenticity, rarity and technical and copyright considerations. A total of 19,000 websites was identified as XS4ALL homepages from the period between 1994 and 2001 [12]. The selection was harvested using WCT between 2019 and 2021. Effort not only went into harvesting the material, but also into describing the creation of this collection. This description contains information on the origins, the content and the sources that were used for discovering potential XS4ALL homepages. The collection of archived websites represents a cross section of the homepages that were hosted by XS4ALL and contains a wide variety of homepages about hobbies, animals, sports, music and personal online diaries.

Since the homepages of users were hosted on subdomains of the XS4ALL domain name and there was not a complete inventory from XS4ALL itself, tracking down all existing homepages was not a straightforward task. The collection description describes exactly which sources curators have used for arriving at the current inventory of XS4ALL homepages. It turned out that there were still old collections of links available on the web. These sources formed the groundwork for our selection. Found links were crawled and availability status was documented since many links were not available anymore at the time of harvesting. In the end not all websites that were still online have been added to

the collection. A selection has been made based on criteria such as age or cultural importance. These selection criteria have been documented, as well as the methods used for dating the webpage. These range from content elements within the page to last-modified date of images. This information will help researchers understand the rationale behind the selection and the reason why websites in the collection are deemed to be representative of a certain period in the history of the web.

Without a list of all existing items that are supposed to be in a collection, it is difficult to determine collection completeness. However, for research purposes it is important to be able to gauge whether an archive has a representative collection. Say a researcher discovers an XS4ALL homepage that is not present in our collection. Using the collection description, source material and other documentation mentioned, it will be possible to determine why this homepage was not archived. For instance, because it was not present in one of the existing link sources, or because the homepage was not available anymore at the time of archiving. This information is important from the perspective of source criticism and functions as an implementation of the concept of context elements. It is also valuable as an independent mechanism for establishing how complete the collection is as required by the guidelines for trustworthy digital repositories (ISO-16363).

V. CONCLUSION

In our opinion, context information can provide insight into the process of creating an archival object from files on the live web. The technology used, the process of quality control as well as the curatorial context of a web collection all impact the integrity and authenticity of the archival objects and the collection itself. The result of harvesting is what is stored in the archive and is therefore also the source material for research on the history and topography of the web. As a trustworthy digital repository, we have the responsibility to provide our users with enough information to keep our collection items understandable. For web archives we think it is important to provide information on the settings of the harvester, on the process of harvesting and quality control and on the curatorial decisions taken in order to acquire the content. Together these three types of information provide ways for our crawls to 'talk'. They tell us more about how the digital objects

came into being and which steps have been taken to ensure quality. In this way, context information will provide researchers with the necessary means for evaluating integrity and authenticity of the web collections that are part of our digital heritage.

ACKNOWLEDGMENT

The authors would like to express their thanks to Iris Geldermans, Johan van der Knijff and Kees Teszelszky for reviewing the first draft of this paper.

REFERENCES

- [1] National Library of the Netherlands, "Content Strategy", 2019. https://www.kb.nl/sites/default/files/documents/content_strategy_eng%20%282%29.pdf
- [2] S.C. van den Eijkel and D. Steinmeier, "What is an archived website?", unpublished
- [3] National Library of the Netherlands, "Preservation Plan", 2019. <https://www.kb.nl/en/about-us/expertise/preservation-policy>
- [4] C. Rogers, Authenticity of Digital Records: "A Survey of Professional Practice", in *Canadian Journal of Information and Library Science*, vol. 39, no. 2, pp. 97-113, June 2015. <https://muse.jhu.edu/article/590936>
- [5] E. Maemura, N. Worby, I. Milligan and C. Becker, "If These Crawls Could Talk: Studying and Documenting Web Archives Provenance," in *Journal of the AAOCIATION FOR Information Science and Technology*, vol. 69, no. 10, pp. 1223-1233, October 2018. <https://doi.org/10.1002/asi.24048>
- [6] B. Sierman and K. Teszelszky, "How can we improve our web collection? An evaluation of webarchiving at the KB National Library of the Netherlands (2007-2017)," in *Alexandria: The Journal of National and International Library and Information Issues*, vol. 27, no. 2, pp. 94-107, August 2017. <https://doi.org/10.1177/0955749017725930>
- [7] Web Curator Tool, <https://webcuratortool.org/>
- [8] Heritrix, <https://github.com/internetarchive/heritrix3>
- [9] I. Geldermans, "Historical growth of the KB web archive," online available at KB Lab, The Hague, 2021. <https://lab.kb.nl/dataset/historical-growth-kb-web-archive>
- [10] NetarchiveSuite, <https://sbforge.org/display/NAS/NetarchiveSuite>
- [11] K. Teszelszky, Chapter: "The historic context of web archiving and the web archive", in: N. Brügger and D. Laursen (eds.) *The Historical Web and Digital Humanities. The Case of National Web Domains*, 2019. <https://www.taylorfrancis.com/chapters/edit/10.4324/9781315231662-2/historic-context-web-archiving-web-archive-kees-teszelszky>
- [12] P. de Bode and K. Teszelszky, Collection description XS4ALL homepages, 2021. <https://zenodo.org/record/5055571#.YVGivH1cKUK> (Full text only available in Dutch, summary provided in English).

THE CO₂ EMISSIONS OF STORAGE AND USE OF DIGITAL OBJECTS AND DATA

Exploring climate actions

Lotte Wijsman

*National Archives of the
Netherlands,
The Netherlands
lotte.wijsman@nationaalarchief.nl*

Arie Groen

*National Library of the Netherlands,
The Netherlands
arie.groen@kb.nl*

Tamara van Zwol

*Dutch Digital Heritage Network,
The Netherlands
tvzwol@beeldengeluid.nl*

Robert Gillesse

*International Institute of Social History,
The Netherlands
robert.gillesse@iisg.nl*

Abstract – The storage and use of digital heritage objects produce carbon dioxide (CO₂) emissions. Cultural heritage organizations can take several measures into consideration in order to diminish these CO₂ emissions. However, how much CO₂ do storage and use produce and what measures could have (the most) effect? We examined the CO₂ impact and possible measures on the basis of a case study. We have focused our investigation on the impact of servers, infrastructure, cloud storage and use. But the story does not end there. We look ahead, beyond the case study and beyond boundaries, introducing a research agenda within the Dutch Digital Heritage Network (DDHN).

Keywords – carbon footprint, sustainability, storage, users, carbon dioxide emissions
Conference Topics – Environment

I. INTRODUCTION

Preserving digital objects for the public contributes, like many human activities, to carbon dioxide (CO₂) emissions and consequently has an impact on the environment. The Dutch digital heritage community is (becoming) conscious of the subject and wishes to examine the facts. What is the environmental impact of the storage and use of collections? What measures can be taken to lessen

the CO₂ impact? And what other issues are still untouched and are waiting for further investigation?

This paper provides insight into certain measures that can be taken when it comes to storage and use, based on a CO₂ impact case study of the Delpher platform [1]. The published results and recommendations based on the case study [2] proved only to be the start of exploring climate actions for heritage organizations. To gain more insights, we asked the community to help us expand the research agenda to their specific needs and wishes. But first we will discuss the findings and recommendations based on the Delpher case study.

The Delpher case study was executed by the company PHI Factory and the Green IT expert group within the Dutch Digital Heritage Network. In Delpher you can search and find millions of digitized texts from Dutch newspapers, books, and other published works. These documents come from the collections of various Dutch scientific institutions, libraries, and heritage institutions. Delpher is developed and managed by the National Library of The Netherlands (KB). We have examined storage and data use in this case study, focusing on the CO₂ impact of servers, the server environment/infrastructure, cloud storage, and the

end use: searching through the files on the platform and downloading files. The study was based on the GreenHouse Gas Protocol [3]. PHI Factory used the guidelines from 'The GreenHouse Gas Protocol' to measure the CO₂ footprint.

This paper presents our findings in those four areas.

II. SERVERS

Servers provide the computing power and storage required to store and make digital collections available for users. These servers are the main cause of CO₂ emissions. This is due to both the electricity consumption and the indirect CO₂ emissions from the production of the servers.

Creating digital compartments in the servers, like the KB has done for the data on Delpher, ensures that the capacity of these servers can be used more efficiently. This can be done by means of virtual machines or containers. The KB's servers consume now 242 MWh (or: 242,000 kWh) annually, which is equivalent to the electricity consumption of 98 average Dutch households in a year.

III. SERVER ENVIRONMENT

The location/environment of the servers has a major influence on the total of CO₂ emissions. If data is stored locally, on the level of one institution, there is a good chance that actions facilitating the servers, such as cooling them, consumes as much or even more energy than the servers themselves. To reduce the CO₂ impact of the infrastructure around the servers you can think about sharing servers with multiple organizations to use them most effectively. By moving the servers from the KB local location to a more efficient, external colocation data center, as in the case of Delpher, considerable savings can be made on electricity costs: saving annually the amount of 151 MWh. Because many servers are located here, facility systems such as cooling can do their job much more effectively. Therefore, this method is not only more sustainable, but also more economic.

You can also opt for more green energy, like the KB has done. Green energy is any energy type that is generated from natural resources, such as sunlight, wind or water. Because green energy is generated from a renewable source, the CO₂ emissions are a whole lot lower than in the case of energy from fossil sources. The annual carbon footprint of Delpher's

servers is less than 4 tons of CO₂ equivalents per year, which equals 4 hot air balloons of 200 m² (the size of a soccer field) filled with CO₂.

IV. CLOUD STORAGE

With cloud storage, the data and computing power of many companies is divided over servers. This makes for very efficient use of (the capacity of) the servers since every available space is being occupied. The advantage of storage in a cloud environment is that the type of providers behind it (e.g. Microsoft and Amazon) are at the forefront of the development of facility systems and the use of containers to make the capacity of their servers as efficient as possible. Naturally, cultural heritage organizations have to consider if they are willing to store their data in a large datacenter under the control of such a provider in perhaps a different country, under different rules and regulations. Because Delpher concerns itself with national Dutch cultural heritage, it has been decided to store the data at a Dutch colocation and not via an international cloud provider.

V. DATA USE

Retrieving files from a digital collection, loading webpages and using the search index causes CO₂ emissions. In the case of Delpher a large part of the digital collection will not be downloaded by a user, but searched, which has only a limited impact. Still, there are ways to even diminish this impact. This could be done by e.g. offering lower resolution versions of the digital object files. In addition, to make it even more effective, you can also limit the user features on the website so that fewer files have to be searched in the data store. For example if you do not offer 'search all' as a standard option, but let users indicate which specific material (newspapers, books or magazines) should be searched.

VI. DISCUSSION

A. Case study as a starting point

The case study provided calculated information on storage and use issues and recommendations for organizations to consider. But of course, one case study alone means there are limitations to what you can investigate. Some research topics are still left on the shelf.

For example we have not calculated the CO₂ emissions of digital preservation workflows like pre-

ingest. We have not considered the carbon costs of the data center building (materials) or the specific carbon costs of multiple information objects storage. However the case study proved to be a starting point for further research and raising awareness.

B. Further research

In order to determine the topics for further investigation we involved the heritage community. During two sessions in November 2021 and June 2022 on the theme of Green IT, heritage professionals from the network were asked about their experiences and wishes for further research.

Important to note is that the environmental impact of digital heritage was not yet included in the policy of most organizations. At the top of the wish list stood and stands therefore climate awareness. However, the heritage community recognizes the complexity of this topic, the sense of urgency and the demand for more knowledge. And certainly, some measures are being considered or already executed, like a stricter selection and avoidance of duplicate storage of digital objects, cleaning up digital data, switching to green power or relocation to other data centers. In order to help organizations raise awareness, the main results and recommendations were visualized in an infographic [4].

During the sessions there were discussions. Does it help to centralize storage? For example, placing audiovisual material with an organization that has the specific expertise and services to do so, instead of trying to solve everything in the local situation.

We should also think about the accessibility requirements. Must everything always be immediately available? Organizations store more and more data everyday but should climate considerations be a reason to make stricter selection choices. Those are some of the questions raised. We cannot answer for all of them, but we can keep open the discussion as a community.

The topics that were mentioned for the research agenda were eventually based round three themes: organizational impact (e.g. cost savings, shared data storage), technology and suppliers (the impact of digitization of materials, sustainable hardware, supplier comparisons) and use (user behavior, the impact of using audiovisual materials online, access of DIP on demand)

The network group has now new members from the community involved and is planning for a research agenda prioritization. Like DEN in the Netherlands, who share knowledge by bringing together sources on this topic [5].

C. Global outreach

Some of the more technological outcomes and recommendations in the case study, especially cloud storage and a more efficient way of dealing with servers and server environment, could be applicable for cultural heritage organizations around the world. We would also highly recommend for heritage organizations of different shapes and sizes to execute their own research and calculate CO₂ emissions, sharing their results with the international community. The more information we can gather together, the better the result will be for the community.

VII. CONCLUSIONS

With the findings from the case study and the aforementioned recommendations, cultural heritage institutions can start to examine the CO₂ impact of their own digital collections and make choices for a climate-resilient future. Also, the case study stimulates further discussion about selection and deduplications of collections in and between cultural heritage institutes. The case study is only a starting point for further research. Hopefully in future we can join forces with other (international) initiatives.

REFERENCES

- [1] Delpher platform <https://www.delpher.nl/>
- [2] De CO₂-impact van opslag en gebruik van digitaal erfgoed, DDHN <https://doi.org/10.5281/zenodo.6341483>
- [3] The Greenhouse Gas Protocol <https://ghgprotocol.org/>
- [4] Praatplaat CO₂-impact opslag en gebruik van digitaal erfgoed, DDHN <https://doi.org/10.5281/zenodo.6411668>
- [5] Ecological sustainability in digital transformation, DEN <https://www.den.nl/actueel/ecologische-duurzaamheid>

IMPROVING THE ARCHIVING AND CONTEXTUALIZATION OF ELECTRONIC MESSAGING IN FRENCH

Bénédicte Grailles

*Université d'Angers
France
benedicte.grailles@univ-
angers.fr
0000-0003-2042-855X*

Touria Aït El Mekki

*Université d'Angers
France
touria.aitmekki@univ-
angers.fr*

Édouard Vasseur

*École nationale des chartes
France
edouard.vasseur@chartes.psl.
eu
0000-0003-1503-2075*

Abstract – The objective of the Pêle-mél program is to propose a prototype tool for exploring and visualizing electronic messages and to test different strategies for contextualizing and classifying electronic messages in French, using NLP techniques, artificial intelligence and learning. Beyond that, the aim is also to draw conclusions on archiving strategies and to theorize the scope of the external knowledge required to succeed in this type of project. This program, financed by the French Ministry of Culture, innovates on two points: the will to understand messaging systems in their reticular context and the deployment of techniques adapted to French. The first results confirm, thanks to a detailed analysis of the messages, the interest of large-scale archiving and classification. They also show the difficulties linked to the temporal context of the studied mailboxes and to the hybrid character of the information supports. The palliative strategies put in place are costly but possible.

Keywords – email, machine learning, classification, terminology

Conference Topics – innovation; resilience

I. INTRODUCTION

The objective of the Pêle-mél program is to propose a prototype tool for exploring and visualizing electronic messages and to test different strategies for contextualizing boxes, correspondent networks and the information content of messages using Automatic Language Processing (ALP) techniques adapted to French, in particular classification. It is also a question of developing criteria to evaluate the archival value of messaging systems and to help in the decision making process,

in order to contribute to a relevant and reasoned selection. The current developments are based on mailboxes already collected by the Records and Archives office of the Ministry of Solidarity and Health and on knowledge external to these messaging systems [1]. The question of the cost of acquiring these criteria, the cost to provide or the actions to prioritize is also on the agenda. This project is financed by the French Ministry of Culture in the framework of a call for projects "innovative digital cultural services" [2].

II. CONTEXT AND STATE OF THE ART

For more than 30 years now, electronic messaging has become an essential tool for the production and transmission of information. Like other countries, France is concerned by the phenomenon, even at the highest governmental levels. For members of the government and their direct collaborators, electronic messaging is an essential tool. In daily work, they have become the medium of strategic information and often the unique traces of decision-making processes [3].

Since the beginning of the 2010s, records managers and archivists of ministries systematically archive documents produced and transmitted by means of electronic messaging by ministers and their collaborators, on the occasion of the cessation of functions. In the social ministries (health, solidarity, labor), electronic messages constituted a significant part of the documents collected in 2020

and represented 45% of the size of electronic archives kept.

Archival appraisal and access to archived messages is also a growing challenge. This is a crucial issue, whether to respond to requests for access from administrations, judges and journalists, or to prepare for the transfer of documents to the National Archives by carrying out the necessary sorting operations. Facilitating the search for meaningful information in the midst of the mass of archival messages is becoming a strategic challenge.

The preservation of electronic messages has been widely studied for several years, both in France and abroad, with several recent publications such as those produced under the authority of Christopher J. Prom [4]. In France, the interministerial electronic archiving team Vitam has been particularly interested in the subject, both from a theoretical and practical point of view [5]. It has developed a java library, MailExtract, which allows the extraction of a tree structure of messages in .eml format from the raw exported files, taking into account the specificities of messages in French, notably accented characters.

On the other hand, the question of access is still little studied, except in the United States. Based on the results of the ePADD project led by Stanford University [6], the RATOM project led by the University of North Carolina has initiated the use of named entity recognition using NLP libraries, in order to facilitate the identification of messages, the publication and public access [7]. Furthermore, email processing is often used for information extraction: spam detection, email categorization, contact analysis, email network property analysis, and email visualization [8]. The value of NLP techniques combining rules and machine learning and using contextual information, has been shown [9]. The value of Topic Modeling to analyze big unclassified text is proven [10] and experiments with pre-annotation of named entities have been conducted [11].

However, standard clustering tools are not sufficient due to typography, the absence of a model or pre-trained corpus, the grammatical specificities to be taken into account in the models, or the semantic characteristics of terms when they contain several words or compound words, which require term extraction based on language.

Undertaking research to facilitate access to archived messages is therefore becoming a necessity in the French-speaking world, because of its linguistic specificities.

III. CORPUS AND PRE-PROCESSING

For this pilot project, we have at our disposal two electronic mailboxes from female advisors in the office of the Minister Roselyne Bachelot-Narquin, comprising 8,636 messages and their attachments, covering the period 2007-2011. It is these boxes that we seek to contextualize and interpret. The records and archives department also provided us with copies of paper directories and organization charts, some of which were native digital and others digitized. We also retrieved two thesauri used by the ministry's documentation center, one from 2014, the closest in date, the other from 2020 (7,000 descriptors) and a corpus of 810 speeches delivered by the minister between 2010 and 2012. These different sets have undergone a certain amount of pre-processing in order to be manipulated.

The messages were taken from Outlook via pst exports. They were processed by the MailExtract tool and are in the form of eml files documented by an xml file that complies with the data exchange standard for archiving [12]. Each message constitutes a separate silo.

The message metadata was retrieved from the xml files and separated from the attachments and the message body. Attachments, message objects, and message texts were morpho-syntactically tagged using the Natural Language Toolkit (NLTK) python software library [13]. The labels were also applied to the speech previously transferred from pdf to txt. Different term extractors, capable of categorizing the names of natural persons, legal entities, abbreviations, locations, etc. and supporting French, could be tested. Only those that could be installed locally were selected in order to preserve the confidentiality of the data. The corpora were also pre-processed to lemmatize and remove accented characters.

The organization charts and directories were entered manually into a spreadsheet. Indeed, these various documents did not present sufficient regularity to attempt an automatic recovery.

IV. FIRST RESULTS

One third of the messages are in a thread (in reply or re-posting). Morpho-syntactic tagging shows that these messages are correctly written, in a relatively elaborate language. The median number of sentences per message is 6 and the median number of words is 228. These sentences include nouns, adjectives and verbs. They are well structured. Attachments are present in 30% to 70% of messages. It is therefore imperative to include these documents in the classification. In total, a network of more than 2,700 correspondents is involved. 30% of the addresses correspond to mailing lists that had to be located in the corpus thanks to regularities (capital letters, punctuation) and then a light manual cleaning. The real people concerned by these lists are not known; the groups have not been archived and cannot be reconstructed with certainty. Nevertheless, the names of the groups are quite transparent and the cross-referencing with directories and organization charts can be exploited. The statistics on recipients show that the information is mainly circulated internally within the minister's office and more broadly within the department.

The exploration tool is based on a relational database that relies on the initial metadata, the attached files and their naming, and the message body. Within the body of the message, signatures were identified and extracted to enrich a directory of individuals, their affiliation and their function. This capitalization is not sufficient in itself. The extraction of the named entities makes it possible to identify natural persons whose function and affiliation must be identified. The directories and organization charts make it possible to inject external knowledge to contribute to the identification.

One of the challenges of contextualizing messages is the identification and resolution of acronyms. The part-of-speech tagger offer an "abbreviations" category. To establish the most complete list possible, we relied on the descriptors of the thesaurus that we projected onto the messages, their subject and attachments, on the implementation of rules and on a dictionary of acronyms and acronyms of the administration crossed with the identification of named entities. More than 400 acronyms have been identified.

A unsupervised clustering was implemented with K-Means and Iramuteq [14]. The result was not very convincing. The thesaurus descriptors were projected onto the messages, their subject and attachments. 60% of the 7000 descriptors are used

in the email corpus. Despite this figure, the thesaurus is not very helpful because it is not sufficient to discriminate between messages that turn out to be very similar.

It was decided to test word embedding models which represent words by vectors and document embedding model which, instead of vector representation of words in the documents present in the data, focuses on creating vector representations of documents regardless of its length: Fasttext [15], Word2Vec, Doc2Vec. Relationships are identified through generic and specific patterns. The results of this phase are currently being verified, but the results are correct.

V. CONCLUSION

The program is now in the middle of the road. The classification gives usable results and the ontology is under construction. The visualizations are very advanced. The question of named entities still needs to be explored. It is still difficult to draw a conclusion. Strategies exist but they are costly. Relevant data is lost: hyperlinks point to nothing, LDAP directory are not kept; mailing lists have lost their contacts. Acquiring external knowledge is complex and requires manual rework at this pivotal time when information is hybrid, paper and electronic. However, it is equally true that there are interesting strategies that could be called upon depending on the level of results expected.

REFERENCES

- [1] In addition to Touria Aït el Mekki, Bénédicte Grailles and Tsanta Randriatsitohaina (University of Angers), Anne Lambert and Chloé Moser (Ministry of Health and Solidarity) and Édouard Vasseur (École nationale des Chartes, Jean-Mabillon Center) are participating in this project.
- [2] List of winners of the call for innovative digital services 2020, Ministry of Culture. [\[https://www.culture.gouv.fr/Thematiques/Innovation-numerique/Appel-a-projets-Services-numeriques-innovants-SNI/Laureats-de-l-appel-a-projets-Services-numeriques-innovants-2020\]](https://www.culture.gouv.fr/Thematiques/Innovation-numerique/Appel-a-projets-Services-numeriques-innovants-SNI/Laureats-de-l-appel-a-projets-Services-numeriques-innovants-2020)
- [3] S. Bretesché., B. de Geffroy, F. de Corbière., *E-bureaucratie: le travail emmaillé des cadres*, Paris: Presses des Mines, 2018.
- [4] C. Prom *Preserving email.*, 2nd ed., London: DPC, 2019. <https://www.dpconline.org/docs/technology-watch-reports/2159-twr19-01/file>

- [5] Programme VITAM, *L'archivage des messageries électroniques. Preuve de concept VITAM*, Paris: Ministère des Affaires étrangères/Ministère de la Culture/Ministère de la Défense, 2013.
- [6] ePADD Project Website Homepage, University of Stanford. <https://library.stanford.edu/projects/epadd>
- [7] RATOM Project Website Homepage, University of North Carolina. <https://ratom.web.unc.edu/>
- [8] G. Tang., J. Pei, & W.S. Luk, "Email mining: tasks, common techniques, and tools", *Knowledge and Information Systems*, vol. 41, no 1, pp. 1-31, 2014
- [9] R. Alghamdi., K. Alfalqi, "A survey of topic modeling in text mining", *International Journal of Advanced Computer Science and Applications*, vol. 6.no 1, 2015.
- [10] D. Maynard, Y. Li, P. Wim, "NLP Techniques for Term Extraction and Ontology Population." In Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. IOS Press, NLD, pp. 107-127, 2008.
- [11] P. Suárez, Y. Dupont., B. Muller, L. Romary, B. Sagot, "Establishing a New State-of-the-Art for French Named Entity Recognition" in LREC 2020 - 12th Language Resources and Evaluation Conference, May 2020, Marseille, France. ([hal-02617950v2](https://hal.archives-ouvertes.fr/hal-02617950v2))
- [12] SEDA standard Homepage, Ministry of Culture. <https://www.francearchives.fr/seda/>.
- [13] Natural Language Toolkit Documentation. <https://www.nltk.org/>
- [14] Iramuteq Website. <http://iramuteq.org/>.
- [15] FastText Website. <https://fasttext.cc/>

FROM OUTPOST TO COMMUNITY:

Strengthening support for the Australasian digital preservation community through regional presence

Jaye Weatherburn

*The University of
Melbourne
Australia*
*jaye.weatherburn@unimelb
.edu.au*
[0000-0002-2325-0331](tel:0000-0002-2325-0331)

Alexis Tindall

*University of Adelaide
Library
Australia*
*alexis.tindall@adelaide.edu
.au*
[0000-0002-1888-6693](tel:0000-0002-1888-6693)

Michaela Hart

*Department of Health
Victoria
Australia*
*michaela.hart@health.vic.g
ov.au*
[0000-0001-9901-0476](tel:0000-0001-9901-0476)

Abstract – The digital preservation workforce is dispersed across organizations, roles, and the world. The value of grass-roots communities of practice and membership organizations to support digital preservation work are evident with the success of such initiatives as the Australasia Preserves regional community of practice supporting practitioners in the Australasian region, and the Digital Preservation Coalition (DPC), an international charity that works to secure our digital legacy. With DPC members dispersed across the globe – and across time zones – engagement can be challenging. In March 2020 the DPC and the University of Melbourne commenced a partnership to embark on a two-year investigative project (2020-2021), to establish a staffed office improving access to the DPC's program of activities for Australasian DPC members and more broadly, digital preservation practitioners in the region such as those participating in the Australasia Preserves community of practice. In early 2022 the DPC announced that this partnership would continue with expansion of the DPC Australasia remit and staffing presence following successful project outcomes. This continuation, the thriving DPC Australasia Stakeholder Group, and the ongoing development of a sustainable business model are undeniable indicators of the success of this project.

Keywords – Engagement, Community, Capacity, Partnerships, Collaboration

Conference Topics – Community

I. THE CHALLENGE AND THE OPPORTUNITY

Digital Preservation Coalition (DPC) membership continues to grow across the world and across sectors, as more and more organizations realize the

considerable and urgent challenge to be addressed. The membership includes a range of types of organizations, from galleries, libraries, archives, and museums, through to banks, government agencies, and commercial entities. Digital preservationists come from diverse professional backgrounds and are, in many cases, facing similar challenges in disparate and isolated environments. The digital record of all organizations grows day by day, and those that have joined the DPC realize the benefit of being part of a network of organizations facing shared challenges.

One of the key benefits of DPC membership is connection with a peer and professional community, whether that be through sharing successes at Connecting the Bits [1], sharing challenges and failures at Digital Preservationists Anonymous [2], or aggregating Rapid Assessment Model (RAM) [3] results to realistically benchmark organisations' digital preservation capability within a professional community. Maintaining that community across countries, and specifically time zones, can be difficult, especially given the need to build trusted communities of practice.

In January 2018 the DPC embarked on a new strategic plan to prepare the transition to a truly global foundation. That ambition was elaborated in June 2019 with the adoption of an appendix to the strategic plan which recognised that digital preservation is a global concern which needs to be addressed as such. This built upon an interest in

inviting international membership, declared from 2016. This commitment indicated that to foster the growth of the global digital preservation community in new markets and geographies, the DPC would be scaled to the extent of the challenge. The DPC's mission was therefore formally expanded to include the enhancement of members' experience and the capacity of the digital preservation community around the world through the provision of a stable and trusted platform for collaboration, owned and run for the benefit of the global digital preservation community, and accountable to them through membership.

This is the context and the reason why, from 1st January 2020 to 31st December 2021, the DPC established an exploratory project in partnership with the University of Melbourne (Australia) [4], whose digital preservation program had supported the founding and ongoing success of the Australasia Preserves digital preservation community of practice. This partnership saw the start of increased delivery of the DPC program in the region, while also exploring the requirements of sustainable ongoing operations to meet DPC members' needs in Australasia. The initiative was implemented through a secondment arrangement, whereby the University of Melbourne provided a part-time member of staff (0.5FTE) to work on 5 strategic goals:

- 1) Sustain and expand Australasia Preserves, the regional digital preservation community of practice instigated by the University of Melbourne's digital preservation program
- 2) Deliver a DPC program in Australasia and surrounding territories
- 3) Develop self-sustaining membership to support a permanent office
- 4) Amplify digital preservation activities in Australasia for the benefit of DPC members
- 5) Amplify the DPC's messages about the need for and benefits of digital preservation

II. ACHIEVEMENTS OF DPC AUSTRALASIA

Indications Of Success

At the commencement of the partnership in early 2020, plans were in place for events, engagement, and travel throughout the region to fully realize the goals of the project. In March 2020 the COVID-19 pandemic greatly disrupted all expected activities. Plans were rapidly redeveloped to take into account

the new uncertain reality. Multiple long state-wide lockdowns in Naarm (Melbourne), Victoria, Australia, particularly impacted plans as the DPC Australasia member of staff was required to work from home for the majority of the two-year project period. Closure of the Australian border to international visitors and between states and territories further impacted the ability to connect with DPC staff, DPC members, and the broader community throughout this time.

Despite the pandemic greatly impacting the program delivery and planning for sustainable operations (including inability to travel or hold face-to-face meetings and workshops during 2020-2021 for planning, development, and membership relationship building and expansion), DPC membership in the region grew from 3 to 15 members during the project period. Many achievements were made in each of the project's five strategic goals.

Support for the Australasia Preserves digital preservation community of practice continued, with community membership growing to over 400 members. Monthly meetups were organized and hosted for the community throughout 2020, moving to quarterly meetups throughout 2021, attracting attendance of 40 to 100 people (varying due to topic and timings), and many of these events are recorded and shared openly with the international digital preservation community on the Australasia Preserves YouTube Channel [5]. A volunteer organizers' group formed in April 2020 continues to support community activities, growth, and forum management [6]. "Digital Preservation Essentials" training modules were developed by facilitated community working groups [7] in 2020, with these resources made openly available to the broad digital preservation community [8].

The delivery of a DPC program of activities in the Australasian region began, guided by local members' input and needs. A local work plan was developed collaboratively with regional members in early 2020 to guide ongoing work for the project, and was reviewed and updated for work in the second year of the project (2021). #DPConnect [9] informal networking sessions were hosted in the Australasian time zone, with feedback overwhelmingly positive: "These short and sweet weekly meets have been a lovely way to share an inspiring number of lockdown projects, from getting on with addressing legacy to new work... it's been great to debrief with like minds whilst achieving some ISO [isolation] relief...very

informal, very supportive” (posted on the Australasia Preserves online community forum).

Various events were facilitated and hosted in the Australasian time zone, including a Rapid Assessment Model (RAM) workshop and webinar, web archiving training, briefing day watch parties showcasing recordings of events held in inconvenient time zones, including preservation planning and technology watch, EDRMS preservation, Digital Preservation Futures sessions, and advocacy training.

DPC membership in the region further strengthened a timely and important collaborative response to a regional priority in the higher education sector during the project period. Australian university members of the DPC (the University of Melbourne with input and endorsement from the University of Adelaide, the University of Sydney, and Monash University) wrote a joint digital preservation response statement to the Australian Research Data Commons (ARDC) “Institutional Underpinnings” process (submitted on 14th July 2021). The ARDC-led Institutional Underpinnings program [10] aimed to develop a framework for institutional research data management (RDM) across Australia’s universities, with 25 participating Universities jointly developing a framework. Digital preservation was missing from the draft framework, so the university-based DPC members convened and co-developed a statement as feedback to raise the profile of the digital preservation challenge. The collaborative statement on digital preservation was reviewed by the ARDC Institutional Underpinnings program and framework editorial committee and 25 participating universities, and was accepted as one of 16 essential elements of a developing national framework for institutional research data management. Without the regional activity that DPC Australasia has enabled, and active stakeholder group that it has stimulated, local digital preservation specialists may not have been alert to this opportunity to contribute to an influential initiative, and it would have been significantly harder to elicit such a quick and relevant contribution.

An extensive member needs analysis was also conducted between April and June 2021 to inform and develop the DPC Australasia Organization Development Plan for the DPC in Australasia 2022-2025, beyond the conclusion of the initial project period. This was developed with input from Australasian-based DPC members during 2021:

Australia’s Academic and Research Network (AARNet), The National Archives of Australia, The University of Sydney, The University of Adelaide Library, National and State Libraries Australasia, Monash University, Records and Information Management Professionals Australasia (RIMPA), Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS), Australian Society of Archivists, National Library Board Singapore, Public Records Office Victoria, National Film and Sound Archive, Queensland State Archives, the National Library of New Zealand, and the University of Melbourne. The needs analysis was undertaken collaboratively over three months (April-June 2021), with 13 hours of member consultations (via videoconference), and more than 340 comments captured to guide development of the plan.

The DPC Australasia Organization Development Plan aims to further progress the international strategic objective of the DPC’s strategic plan, and includes forecasts and recommendations based on the best knowledge available in December 2021 and on practical experience. It is therefore also a provisional outline with explicit and implicit assumptions that will be tested and are subject to emerging risks and opportunities.

Consultation with members throughout the development of the plan highlighted the value of a supported local network, with the forward plan highlighting the “keen need to foster connections between other members in the region, especially in order to build trust and enable robust open discussion of digital preservation capacity and approaches in a collegial environment” [11].

III. REGIONAL CASE STUDIES

Case Study: Sole Digital Archivist, Department of Health, Victoria

Working in large organizations in digital preservation can be similarly isolating to those in smaller ones. Transitioning traditionally paper based archival and recordkeeping functions into digital with robust and sustainable preservation workflows requires advocacy, education, and training. Access to the Australasia Preserves Community of Practice and DPC membership through the Australian Society of Archivists has provided this Digital Archivist access to training and resources that would otherwise be unavailable.

The familiar mantra that digitisation is not digital preservation expresses itself frequently when advocating for large collections of paper-based materials. It may not seem analogous however the reality is that many archivists work in hybrid collections and while sustainable digital preservation is a goal, practitioners must take a holistic view of an entire collection and prioritize and justify the areas of greatest need.

Australasia Preserves provides a space for testing ideas and sharing challenges and mistakes so that when practitioners are at the point of making a business case or operationalizing digital preservation workflows much of the macro thinking has already been progressed.

An example of this has been when working with colleagues in system design. Considerable advocacy and relationship building with these colleagues has shifted the role of Digital Archivist from the person to go to when a system is being retired, to a colleague to involve at the development stage. This has resulted in a true picture of the technology and costs required to maintain information for the whole of its life.

Case Study: University of Adelaide Library Strategy and Roadmap Development

In mid-2020 the University of Adelaide Library commenced work in earnest to improve digital preservation, as part of our strategic commitment to be leaders in information management [12], by joining the Digital Preservation Coalition and commencing development of a Digital Preservation Strategy and Roadmap. While a cross-organisational working group gathered considerable archival, library and other relevant expertise, organizational digital preservation experience was awareness-level.

Benefits of DPC membership, including access to DPC online resources, technical guidance, and tools such as the rapid assessment model and Novice to Know-how training helped focus the Strategy Working Group considerably and rapidly improved the organization's maturity as we tackled this challenge.

The Library's Digital Preservation Strategy and Roadmap was endorsed by the Library Leadership Team and published online in July 2021 [13]. The document was influenced by our participation in the DPC Australasia Stakeholder Group as a peer network. Formally, membership of the DPC provided

the Library with feedback on the document before it was finalized. In addition to this, library representatives to the Stakeholder Group had an opportunity to connect with similar sized organizations facing similar challenges in familiar professional and budgetary circumstances. While published case studies are useful research, a trusted peer network and direct connection with equivalent organizations helped us benchmark ourselves even more effectively.

These relationships with like organizations impacted the phasing and timeline of the Library's Roadmap, as well as influencing how we approach staff upskilling. In the final phase of the Roadmap, the Library will consider implementing a digital preservation system. At the time of writing, the Library has not completed the first phase of the roadmap, but the DPC Stakeholder Group helps us keenly observe the experience of similarly-sized colleague organizations using preservation systems already. The Library is especially interested in the progress of those that are in the early stages of implementing commonly used systems. Their experience will prove a rich supplement to a formal process of systems procurement when the time comes.

IV. THE FUTURE OF DPC AUSTRALASIA

Looking Forward

The work undertaken during this partnership project has produced one emerging approach for setting up a regional DPC presence – an approach which could potentially be used in other regions to help scale other communities to the global challenge of digital preservation. In early 2022, recruitment was underway for a full-time dedicated Head of Australasia and Asia-Pacific strategic leadership role, a role focussed on continuing to grow the DPC presence and membership sustainably beyond the project period, with human infrastructure support provided by the University of Melbourne for a three-year period, linked into the university's digital preservation program with its strategic focus of national and international collaboration.

Throughout 2021, DPC members in the Australasian region collaboratively developed a vision, mission, and values to guide ongoing development of the DPC in the region. At this early stage of development, the region of 'Australasia' has not been definitively determined, in part due to the

lack of standard worldwide consensus as to what is included (without an official definition, Australasian and Asia-Pacific (APAC) countries vary depending on context, with some lists including Russia, the US, Canada, Chile, India, Mongolia, etc). During the project period, it was recognised that wider growth in the region surrounding Australia will depend on local initiatives and resourcing to seed new chapters of the DPC, while many factors including cultural distinctions, language diversity, further time zone challenges, and differing vision, mission, and values may emerge and require addressing in local contexts. As the DPC in Australasia initiative further develops in 2022 and beyond, it is expected these subjects of name, context, and regional definition will further evolve as part of the DPC's ongoing international strategy development.

DPC members in the Australasian region have expressed a keen need to foster further connections with other members in the region, especially in order to expand trusted networks and enable robust open discussion of digital preservation capacity and approaches in a collegial environment. For this reason, additional staff duties have been identified in the DPC Australasia Organization Development Plan and prioritized to meet this fundamental need. In addition to a strategic leadership role responsible for coordination, planning, research and development, other recommended staff duties for prioritizing in the Australasian region include workforce development and skills/training expertise, and communication, events, and administration expertise.

In terms of facilitating broader, more systemic advocacy for digital preservation in the Australasian and Asia-Pacific regions, it has been recognised this endeavor will require larger institutions and organizations becoming involved who have comprehensive reach and understanding of the current environments. This goal could potentially also be achieved aided by more dedicated resourcing of communication and strategic guidance and expertise for DPC operations in Australasia as this initiative further grows and matures. In any case, digital preservation capacity and advocacy has been greatly enhanced through this partnership, through the project period (2020-2021), and is expected to further contribute to regional capability to secure our digital legacy in the coming years.

REFERENCES

- [1] Connecting the Bits Members' Unconference, DPC. <https://www.dpconline.org/events/past-events/ctb2021>
- [2] Digital Preservationists (DP) Anonymous, DPC. <https://www.dpconline.org/events/dp-anon-july2019>
- [3] Rapid Assessment Model (RAM), DPC. <https://www.dpconline.org/events/dp-anon-july2019>
- [4] DPC Launches new office in Melbourne, Australia, DPC. <https://www.dpconline.org/news/dpc-aus-office-launched>
- [5] Australasia Preserves Youtube Channel, Australasia Preserves. <https://www.youtube.com/channel/UCRO-yOP6cWYbLIsZdv9bncg>
- [6] About Australasia Preserves (Co-Organisers), Australasia Preserves. <https://www.australasiapreserves.org/p/about.html>
- [7] Australasia Preserves Working Groups, Australasia Preserves. <https://www.australasiapreserves.org/p/working-groups.html>
- [8] Digital Preservation Essentials, Australasia Preserves. <https://www.australasiapreserves.org/p/working-groups.html>
- [9] #DPCconnect informal networking sessions, DPC. <https://www.dpconline.org/events/dpconnect>
- [10] Institutional Underpinnings, ARDC. <https://ardc.edu.au/collaborations/strategic-activities/national-data-assets/institutional-underpinnings/>
- [11] Document DOCRC1221C *Report to Council and Forward Plan Dec 2021 - March 2022*
- [12] Beyond the Library of the Future: University Library Strategic Plan 2019-21, University of Adelaide Library. <https://www.adelaide.edu.au/library/about-the-library/reports-publications>
- [13] University Library: Digital Preservation Strategy and Roadmap, University of Adelaide Library. <https://www.adelaide.edu.au/library/about-the-library/reports-publications>

PRESERVATION WATCH

Working Towards A Supra-Organizational Preservation Watch Function Within The Dutch Digital Heritage Network*

Tamara van Zwol

*Dutch Digital Heritage Network
Netherlands
tvzwol@beeldengeluid.nl*

**Eva van den Hurk – van
't Klooster**

*Regional Historical Centre Vecht
en Venen
Netherlands
e.vandenhurk@rhcvectenvenen.nl*

Lotte Wijsman

*National Archives of the
Netherlands
Netherlands
lotte.wijsman@nationalearchief.nl*

Abstract – Preservation Watch is a vital function when it comes to monitoring internal and external developments that can benefit or risk digital objects. However, given the abundance of developments and risks, it is hard for organizations to keep up. As a solution, the Dutch Digital Heritage Network started work on a supra-organizational Preservation Watch function led by a group of experts from the field. This paper will expand on the scope of the group, our goals, and our stock agenda for this year.

**Keywords – Preservation Watch, Community, Supra-Organizational, Exchange
Conference Topics – Community; Exchange**

I. INTRODUCTION

Preservation of digital information objects can and will be influenced by all types of factors, such as advancing technologies, organizational policies, the changing needs of your designated communities¹, or even climate change. Some of these developments can pose a risk or, in the best of times, prove to be a benefit to the life cycle and sustainability of digital information objects. Therefore it is important to monitor internal and

external developments in order to take appropriate measures on time. This monitoring function is called Preservation Watch [1]. By implementing Preservation Watch into your organization and its preservation policy, you can monitor the potential risks and act accordingly.

But how do you keep track of the array of developments and possible risks? Especially organizations with limited resources can struggle to keep up. Being part of a network helps: a community of heritage organizations where you can find and share (practical) expertise, signal and observe developments, do research, and ask experts to address specific topics within the context of Preservation Watch.

For these reasons, the Dutch Digital Heritage Network started building a supra-organizational Preservation Watch function in March 2021 by forming a group of experts.² The Netherlands Institute of Sound and Vision is the coordinating party in this group, working with the experts from various Dutch heritage organizations and the field of digital humanities.

* A supraorganization is an organization whose members or stakeholders are organizations rather than individuals and which performs an overarching function. With supra-organizational in the context of this paper we refer to our team that consists of members who are representing their organizations and how we want to have an overarching Preservation Watch function to aid other heritage institutions.

¹ The Designated Community is an identified group of potential Consumers who should be able to understand a particular set of information. From: Reference Model for an Open Archival Information System, <https://public.ccsds.org/pubs/650x0m2.pdf>.

² Although we acknowledge the importance of the two other preservation functions: Preservation Planning and Preservation Action, this group will focus solely on Preservation Watch. This is due to the fact that the other two functions are primarily carried out by individual organizations, rather than lending itself to a supra-organizational approach.

The Dutch Digital Heritage Network is formed by organizations in the fields of culture, heritage, education, and research together. With suppliers of heritage software, provinces and municipalities we are working on the implementation of the National Strategy Digital Heritage, supported by the Ministry of Education, Culture and Science. Across the boundaries of our organizations and collections we can secure public access to heritage information for the future. This ambition binds together diverse organizations and partnerships in the Dutch Digital Heritage Network. The network, established in 2015, consists of organizations of different sizes and from different heritage sectors, yet the six main national institutions that contribute to and represent this network are the KNAW Humanities Cluster, the KB, National Library of the Netherlands, the National Archives of the Netherlands, The Netherlands Institute of Sound and Vision, Het Nieuwe Instituut, and the Cultural Heritage Agency.

In this short paper, we will expand on Preservation Watch, how we are building and organizing a facility that transcends the level of the individual organization, in what ways we want to involve and include the (international) community, and which topics are to be monitored on this supra-organizational level.

II. DEFINITION AND SCOPE PRESERVATION WATCH

Preservation Watch is:

- Monitoring internal and external developments (threats and opportunities) that may have an impact on the sustainable accessibility of digital information objects;
- Weighing the risks and opportunities that these developments entail;
- Testing (new) tools and services that may be helpful in ensuring the sustainable accessibility of digital information objects;
- Documenting and sharing the results of all these actions.

Preservation Watch serves digital preservation and it entails all activities and processes that are

necessary to keep digital information objects accessible to users for as long as necessary. The field of digital preservation covers the entire life cycle of digital information objects, i.e. all processes related to their creation, acquisition, selection, processing, storage, management, and the provision of access to them. Preservation Watch therefore has a broad scope and covers:

- Technological developments (e.g. software becoming end-of-life);
- Organizational developments (e.g. changes in budget);
- Political and social developments (e.g. the introduction of the General Data Protection Regulation)[2]³;
- Developments in use or in user groups, the so-called designated communities (e.g. younger generations not being familiar with WordPerfect).

It is hard to tackle this broad scope of developments as a single heritage organization. So how can different organizations work together on this? Ideally Preservation Watch has a cyclic effect: institutions include relevant developments in preservation planning, after which these developments are acted upon. The resulting experiences are processed into best practices that are shared with the (international) community, so others can benefit from the work that has already been done.

III. GOALS AND IMPLEMENTATION

A. Strategic Goal

The supra-organizational Preservation Watch function is set up with the following strategic goal in mind:

Heritage and knowledge institutions work together as network partners to gain experience with Preservation Watch at a network level and from there to come to (agreements about) an efficient and effective organization of the Preservation Watch function in the heritage sector and the digital

The idea is to start small in 2022 and to use the experience acquired to formulate a proposal for further development after the first year. This year,

³ The General Data Protection Regulation is a regulation in EU law on data protection and privacy in the European Union and the European Economic Area.

the main focus will be on getting a form of supra-organizational monitoring off the ground, gaining initial experience, and organizing and/or facilitating a form where knowledge can be shared and exchanged.

For the sake of feasibility,⁴ the scope in 2022 will be limited to the technological developments as specified in the definition given above.⁵ Within this scope, the emphasis will be on topics that are, judging from signals from the network, considered to be the most urgent:

- File formats and preservation tools;
- Metadata models/schemes/standards;
- Storage techniques.

Setting up a supra-organizational Preservation Watch presupposes that a form of Preservation Watch is also set up at the organizational level. The supra-organizational level is intended to complement, not replace. The supra-organizational Preservation Watch is primarily aimed at signaling and monitoring new developments, weighing the extent to which developments are promising or threatening for the heritage organizations and communicating this to the heritage organizations in question. The supra-organizational Preservation Watch does not have an advisory role for the network.

Taking action in response to identified developments is primarily the responsibility of individual organizations. However, it is conceivable that certain actions will be taken in the context of the network. For example: if the Preservation Watch expert group that also exists within the Dutch Digital Heritage Network indicates that a new file format is on the rise, the Preferred Formats expert group can pick up this signal and supplement the Preferred Formats Guide with information on this new file format [3].

B. Operational Objectives 2022

Several operational objectives have been set for 2022. As a start, a Preservation Watch expert group has been set up with experts from various heritage

organizations, who will experiment with the design of the supra-organizational Preservation Watch function. Furthermore, the expert group will identify, follow, and discuss developments in the topics of file formats and preservation tools, metadata models/schemes/standards and storage techniques.

The group will record their acquired knowledge, but will also create lists that present the sources of knowledge that are essential when monitoring the technological developments in the three topics mentioned earlier, and which resources and techniques can be used to perform the Preservation Watch function.

Another important objective concerns communication. As mentioned previously, we wish to share the acquired knowledge with the community. This requires us to find a method that will weigh the opportunities and risks of technological developments for the heritage sector. Furthermore, we wish to find a way to make the deliverables of the expert group available to all network partners via an existing knowledge platform and at least two sessions (in person or online).

As mentioned previously, 2022 is the start of this project and we will start small (by focusing on technological developments first). However, part of this first year is also to investigate the possibilities of expanding the supra-organizational Preservation Watch after 2022 (to include organizational, political, social, and use(r) developments, and also strengthening it by attracting other types of partners such as suppliers, people from the field of digital humanities, and Flemish partners).

With the supra-organizational Preservation Watch, we wish to cover the full scope of the Dutch Digital Heritage Network: Archival organizations, museums, libraries, institutes in the field of audio-visual heritage and media, and scientific institutes in the area of digital humanities.

C. Expert Group

In line with the starting point of the Dutch National Strategy Digital Heritage [4], the supra-organizational Preservation Watch starts from the cooperation within the Dutch Digital Heritage Network. The function takes shape and is implemented in the expert group, which is made up

⁴ Factors that include management and administration, e.g. budgets and collections, are specific to each organization.

Therefore, we have decided to concentrate on technological trends and developments.

⁵ See II. Definition And Scope Preservation Watch

of expert staff from various network partners. Nationally active heritage organizations play a key role, but other organizations are also involved. The members of the group will also act as linking pins to the heritage community. At the start of 2022, the expert group consisted of members from the Dutch Digital Heritage Network (project lead and coordination), the National Library of the Netherlands, the National Archives of the Netherlands, Data Archiving and Networked Services, Regional Historical Centre Vecht en Venen, Amsterdam City Archives, the Netherlands Institute for Sound and Vision, EYE (the Dutch Film Museum), and the Utrecht Archives. In the future the expert group will expand and add members from other heritage organizations from the Dutch Digital Heritage Network. Each member of the expert group will be part of a subgroup that focusses on one of the three main topics: file formats and preservation tools, metadata models/schemes/standards, and storage techniques.

2.

IV. STOCK AGENDA

For each topic, the expert group works with a stock agenda of developments that need to be kept up to date. The stock agenda can be adjusted and/or supplemented during the course of the year, with the needs of the designated community, individuals working at heritage institutions, as the guiding principle. On the one hand, each subject is about monitoring existing products, services, and developments that change, become outdated and/or obsolete and therefore pose possible risks. On the other hand, it is about identifying new products, services, and developments that may have potential for the heritage sector.

While we have already made our focus for 2022 smaller by focusing on technological developments first, the three topics are still quite extensive. Therefore, the expert group will be supported by external experts ('watchers'). These watchers can provide the necessary monitoring actions, to enable the expert group to focus on weighing risk factors and share the most important risks and developments for the heritage community with the network.

1. Stock Agenda Topic File Formats And Preservation Tools

For the topic of file formats and preservation tools we have already selected several subjects we will build up knowledge about. Concerning file formats, we wish to investigate the obsolescence, phasing out, and disuse of file formats [5], and the properties that accompany them. Concomitant to this, we will research the possible and/or necessary preservation actions needed with the file formats in question. We also wish to explore (the developments regarding) new file formats, new versions of existing formats, and new functions in existing file formats.

Regarding preservation tools, we plan to research tools for setting up a (preferably automated) process for monitoring the obsolescence of file formats or file properties. Additionally, we will monitor international developments regarding preservation tools, from all parts of the world (including Asia, South America, and Africa).

<i>Stock</i>	<i>Agenda</i>	<i>Topic</i>	<i>Metadata</i>
<i>Models/Schemes/Standards</i>			

For the topics of metadata models/schemes/standards we monitor several developments. The development of PREMIS [6], METS [7], MDTO [8], RIC/RIC-O [9], and other metadata models/schemes/standards will be monitored: who is working on them, in what direction are they developing, and what is the expected impact of this for the heritage organizations? Also monitored will be E-ARK projects concerning the development of design principles [10] for an information package, the impact of the Proof of Provenance project⁶, developments around the detection and repair of 'unconscious bias' in collection metadata [11], developments concerning the automatic allocation of metadata in the e-Depot or digital repository, and the possible impact of new or amended legislation and regulations (e.g. Copyright laws) on metadata models/schemes/standards.

Stock Agenda Topic Storage Techniques

Concerning the topic of storage techniques, we feel it is necessary to develop criteria for sustainable storage technologies. This allows us to compare the various techniques in the same manner.

⁶ Project of Sound and Vision on recording (metadata about) the origin of digital/digitised content, important in relation to Linked Open Data; project is still in application phase.

The topic storage techniques can be subsumed into three categories: techniques that are obsolete, phased out, or disused, techniques that are current (e.g. Object storage, and Cloud storage), and techniques that are still being developed and interesting for the future (e.g. DNA storage, and optical carriers such as glass).

V. PRESENT

The operational objectives for 2022 have been put into planning. Overarching the entire year will be the research on the three topics and the accompanying stock agenda's. At the time of writing this short paper we are still at the start of putting our plan into action. We will focus on several starting points such as choosing and arranging a knowledge platform to work from, establishing our stock agenda's on a more in-depth level, writing a communication plan, finding sources of knowledge, and finding several experts that will help the expert group in their monitoring. After these steps have been taken, we can shift our attention to exploring the possible expansion of the group with other partners, and investigating how to broaden the Preservation Watch function beyond the technological developments. Additionally, we will offer several sessions to the members of the Dutch Digital Heritage Network to get to know the supra-organizational Preservation Watch, and to share the knowledge that has been acquired up to that point.

ACKNOWLEDGMENTS

We would like to thank all experts involved in the Preservation Watch project for their past and present contributions: Marjolein Steeman, Sam Alloing, Valentijn Gilissen, Liesbeth Oskamp, Walter Swagemakers, Annelot Vijn, Ana van Meegen, Robert Gillesse, Hans Laagland, Barbara Sierman, Marcel Ras, and, for writing the original project plan 'Generieke Preservation Watch 2022' which forms the basis of this paper, Margreet Windhorst.

REFERENCES

- [1] https://wiki.dpconline.org/index.php?title=Preservation_Watch; http://www.planets-project.eu/docs/reports/Planets_PP7-D6_EvaluationOfPPWithinOAIS.pdf
- [2] <https://gdpr.eu/what-is-gdpr/>
- [3] <https://www.wegwijzervoorkeursformaten.nl/index.php/Hoofdpagina>
- [4] <https://www.rijksoverheid.nl/documenten/beleidsnotas/2021/03/15/nationale-strategie-digitaal-erfgoed-2021-2024>
- [5] <https://www.dpconline.org/digipres/champion-digital-preservation/bit-list>
- [6] <https://www.loc.gov/standards/premis/>
- [7] <https://www.loc.gov/standards/mets/>
- [8] <https://www.nationaalarchief.nl/archiveren/nieuws/van-tmlo-en-tp-rijk-naar-mdto>
- [9] <https://www.ica.org/standards/RiC/ontology.html>
- [10] <https://dila.eu/specifications/sip>
- [11] See the SABIO project: <https://netwerkdigitaalerfgoed.nl/nieuws/project-sabio-wil-bias-in-collectiebeschrijvingen-via-ai-opsporen/>

PEERING INTO THE JUNGLE

Challenges in determining preservation status of open access books

Mikael Laakso

Hanken School of Economics
Helsinki, Finland
mikael.laakso@hanken.fi
[0000-0003-3951-7990](tel:0000-0003-3951-7990)

Alicia Wise

CLOCKSS
California, USA
awise@clockss.org
[0000-0002-7898-3428](tel:0000-0002-7898-3428)

Ronald Snijder

OAPEN Foundation
The Hague, The Netherlands
r.snijder@oapen.org
[0000-0001-9260-4941](tel:0000-0001-9260-4941)

Abstract – This paper reports on some of the initial observations from an ongoing study focused on determining the preservation status of academic open access books. The central challenges discussed revolve around lack of common definitions, metadata, and established practices for openly recording preservation status for books.

Keywords – open access, books, monographs, preservation

Conference Topic – Resilience

I. INTRODUCTION

Making academic content openly available for everyone using the web has never been easier from a technical and financial cost standpoint, the maturity and widespread adoption of web and document standards take care of a lot of challenges that were creating friction in the past. Web services that facilitate content upload and open distribution of academic works like monographs, book chapters, individual article manuscripts, and entire journals are sprawling up at an unprecedented pace which has led to a rapidly increasing volume of academic content available out in the open. While the act of making something openly available provides open access (OA) to the content for the moment, the practices for ensuring preservation to such content for the long-term are still developing, and to a degree unknown. Based on evidence from recent interviews and workshops on OA book preservation with key stakeholders, many of the central questions related to best practices of preservation are still evolving and there is a need to gain more information about current practices and work towards robust preservation solutions[1,2].

A recent study gauged the degree to which content from OA journals has vanished from the web since the year 2000, finding that at least 174 OA journals had vanished from the active web and had lacking preservation coverage for their published materials [3]. Partly inspired by the findings of this study Project JASPER (JournAIS are Preserved forever) was initiated which is a collaboration between CLOCKSS, DOAJ, The Internet Archive, The Keepers Registry, and PKP [4]. There is currently no similar overview of materials lost or at risk of being lost due to lacking preservation coverage concerning OA books. As there is growing momentum for advancing OA to academic books through science policy it would be important to scope the landscape through a systemic study to map the current preservation status of published materials.

II. TOWARDS BETTER KNOWLEDGE ABOUT PRESERVATION COVERAGE

The aim of an ongoing study is to conduct a data-driven mapping of the current landscape of preservation within the content domain of OA books. The focus of the study is on academic monographs and edited books that are or have been available OA. Excluded are non-published theses and dissertations, and individual book chapters. The definition of at-risk materials is lack of preservation inclusion in a preservation service e.g. Portico, CLOCKSS, or other similar recognized infrastructure. This study is not focused on issues related to specific file formats of preservation, merely that an indication of some preservation exists for a specific title.

Already from the outset it was known that the data collection circumstances for vanished and currently online OA book content differs significantly from that of scholarly journals. Laakso, Matthias & Jahn (2021) utilized mainly past and present journal lists provided by journal indexing services to identify potentially vanished journals, and verified the preservation status through information from the Keepers Registry and Internet Archive snapshots of the last known URL. For OA books the situation is more fragmented due to the lack of comprehensive international services for content indexation, and for registering preservation inclusion across service providers.

An additional component in this ongoing study will be to figure out what domains host the OA book content, by checking which URLs their DOIs or full text links point to. This is not a way of verifying preservation, but such an exploration can shed light on what the long tail for content providers looks like and potentially what type of organizations are running them if it can be derived from the domain names.

III. CHALLENGES OBSERVED SO FAR

1. *Definitions: When is a book an academic book, and when is it open access?*

Not all books on the web are of key interest to this study, where focus is on non-fiction academic books. Most bibliometric databases provide filtering to either “Book” and/or “Monograph” with very few offering further ways to reliably narrow the scope down from there. There is no widely used tag for “peer reviewed” or similar that would make it possible to filter the large quantity of entries down, leaving it up to the inclusion criteria/data harvesting methods of each service provider to what is included and what is not. Further, as categories are so wide there is often a lot of thesis, reports, and individual book chapters sprinkled in among the search results which are hard to identify and separate in any automated way. This is not only a factor that concerns only metadata, but also overall transparency and knowledge available about what kind of editorial processes are behind published works.

Ambiguity is also introduced by the concept of OA, as some sources allow filtering to content available in full text for free (without any distinction between OA types), some do not have OA filtering at

all, and some have very granular metadata concerning OA metadata. The circumstances for preservation are different if the content is available in a document repository in manuscript form compared to the publisher’s website to which also the publications DOI also points to. Barnes, Bell & Cole et al [2] found out through their interviews with stakeholders in the landscape there are some publishers that upload their published content to local repositories, but if that archival counts as preservation depends on the policies and precautions of the institution running the service, which makes gauging the viability of such practices hard at scale. There has been a lot of progress in this area but there is still work to be done with it comes to reliably filtering OA content across key services.

2. *Data management: Physical extraction of metadata to represent the “global bookshelf” of academic OA books*

The amount and quality of freely available metadata describing publications has never been as good as it is now. However, the growing size of increasingly detailed and comprehensive metadata comes at the price of data size (and to some degree data precision, as some other mentioned challenges point out).

Slicing out book metadata from some of the widest openly available bibliometric datasets in the world (e.g. Crossref, OpenAlex, OpenAIRE) requires either downloading the entire datasets which are often in the 100s of gigabytes uncompressed, mapping the JSON files to a database, and designing queries to extract the wanted data concerning books contained in the data. Many bibliometric datasets are becoming challenging to process locally even on a modern desktop computer since they do not fit to be processed into available computer working memory. What is of interest for the purposes of book preservation information is relatively small, but extracting it often requires dealing with the entire dataset at the outset which limits accessibility.

The services mentioned above also offer API access which means that they can be queried programmatically for extraction of specific records. This requires some familiarity with programming or setting up scripts to send requests for multiple chained queries as only a limited number of records are given as response per request. Unfortunately, some API services like OpenAIRE do not allow queries to be filtered to only books, making that path

unviable for book-related queries. Crossref on the other hand has no reliable way to extract only items available OA.

3. *Unique identifiers: Taming the wilderness of identifier metadata describing OA books*

Though the volume and quality of openly available metadata concerning OA books is better than it has ever been and is constantly improving, there are some obstacles for straightforward duplication checking when data is aggregated from several complementary data sources. Matching by title or author is not reliable due to even small differences in spelling, format and punctuation leading to incorrect matches. There is varying use of unique identifiers for books, where ISBNs might be the only value available in one dataset (e.g. WorldCat) or but not available for any entries in another service (e.g. Lense) where DOIs are the key identifier used. Many services also have their own unique identifiers for entries but these are of little use when the data is to be aggregated with data from elsewhere. Table 1 provides a coarse overview of some key data sources, their estimated volume of OA books together with unique identifier availability in the metadata of the records.

Table 1

Overview of bibliometric sources containing records of OA books

Service	Scope of OA book content	Unique identifier availability in metadata
OpenAlex	4 545 046 indexed objects of type "Book", 203 857 objects of type "Monograph" + OA	ISBN = 0% DOI = Not yet measured, but high share
Crossref	Works of types book or monograph 328 098 that have license information and link to full-text (not necessarily OA)	ISBN = Not yet measured, but high share DOI = Not yet measured, but high share
WorldCat (OCLC)	4 597 non-fiction e-books tagged as OA	ISBN = 100% DOI = 0%
DOAB	55 723 academic peer-reviewed books, all OA	ISBN = 86% DOI = 83%
Scielo Books	1564 complete titles of which 963 are OA	ISBN = 100% DOI = 93%
Lense	348 267 records under "Open Access" and "Book" published between year 0 and 2050.	ISBN = 0% DOI = 99%
OpenAIRE	211 749 records under "Open Access" and "Books" after removal of individual chapters,	ISBN = 0% DOI = 99%

	thesis, reports, and preprints.	
--	---------------------------------	--

4. *Openly available preservation data: Preservation data is scarce for all but the largest service providers, and even their datasets could be improved*

The challenges mentioned so far have concerned creating a comprehensive dataset of OA books, but none of the data so far is capable of providing indication for which titles are reported to be preserved through some service. CLOCKSS [5], Portico [6], and Global LOCKSS Network [7] all provide open datasets that describe which books they have included in their coverage. None of these three provide DOI's for their records, only ISBNs which is not optimal as most of the major bibliometric service providers focusing on OA book content rely on DOIs.

National libraries have good data within them but programmatic access from outside is still limited. Barnes, Bell & Cole et al [2] found that some OA monograph publishers deposit copies into national library holdings, something which would be very interesting to obtain more information about on a larger scale. However, the holdings of libraries around the world are not easy to query programmatically from the outside.

5. *Building a path forward*

With all these intertwining challenges present, pinning down the status for preservation of OA books is not a straightforward process and will even under optimal circumstances be an estimate rather than absolute and comprehensive as the definitions and practices in the landscape are still emerging. Below are some observations that could help shape the path forward for a more transparent preservation landscape for OA books.

Data sources that include book materials should strive to include both ISBNs and DOIs in the metadata when they are available since that makes matching to preservation data much more reliable. Early experiments have shown promise in fuzzy matching of book titles to preservation records based on combinations of author information, book title, and publisher. However, the approach needs to be assessed more extensively but in cases where direct matching does not garner results such an approach might show utility as long as the number of false positive matches can be contained.

It could be argued that OA content would benefit from OA status information for preservation, i.e. that there would be practices and data in place that would make it easy to both deposit and verify where specific pieces of openly available content are properly preserved. Concerning preservation data national libraries could on their own or through collaboration make available open machine-readable data concerning which books are preserved in their digital holdings. A service similar to The Keepers Registry that the ISSN International Centre maintains for journals would be very much needed for books as well, where preservation service providers could automatically report which titles they include in their coverage.

REFERENCES

- [1] Bell, E. (2020). COPIM Archiving and Preservation Workshop, September 2020. COPIM. <https://doi.org/10.21428/785a6451.0e666456>
- [2] Barnes, M., Bell, E., Cole, G., Fry, J., Gatti, R., & Stone, G. (2022). WP7 Scoping Report on Archiving and Preserving OA Monographs (1.0). Zenodo. <https://doi.org/10.5281/zenodo.6725309>
- [3] Laakso, M., Matthias, L., Jahn, N. (2021). Open is not forever: A study of vanished open access journals. J Assoc Inf Sci Technol. 2021; 72: 1099– 1112. <https://doi.org/10.1002/asi.24460>
- [4] DOAJ.org (2021). Project JASPER - Open access journals must be preserved forever. <https://web.archive.org/web/20210916132815/https://doaj.org/preservation/>
- [5] CLOCKSS (2022) <https://reports.clockss.org/keepers/keepers-CLOCKSS-books-report.csv>
- [6] Portico (2022) <https://api.portico.org/holdings/ebooks/e-books-part1.xlsx> and <https://api.portico.org/holdings/ebooks/e-books-part2.xlsx>
- [7] Global LOCKSS Network (2022) <https://reports.lockss.org/keepers/keepers-LOCKSS-books-report.csv>

EVALUATING A TAXONOMY FOR VIDEO GAME DEVELOPMENT ARTIFACTS

Taxonomies for New and Innovative Domains

Marc Schmalz

University of Washington
USA

mschmalz@uw.edu
[0000-0003-1027-696X](https://orcid.org/0000-0003-1027-696X)

Kylie Snyder

University of Washington
USA

ksnyd@uw.edu
[0000-0002-3374-2364](https://orcid.org/0000-0002-3374-2364)

Lidia Morris

University of Washington
USA

ljmorris@uw.edu
[0000-0002-4702-6975](https://orcid.org/0000-0002-4702-6975)

Corey Cherrington

University of Washington
USA

cherri93@uw.edu
[0000-0001-6965-9987](https://orcid.org/0000-0001-6965-9987)

Tara Disher

University of Washington
USA

tdisher@uw.edu
[0000-0002-3285-1411](https://orcid.org/0000-0002-3285-1411)

Jin Ha Lee

University of Washington
USA

jinhalee@uw.edu
[0000-0002-9007-514X](https://orcid.org/0000-0002-9007-514X)

Abstract – Digital game development is innovative and intersectional, producing cultural texts in an emerging field across new technology, physical, and digital media. As such, it offers fertile ground for designing and evaluating structures to help creators, information professionals, and others organize and preserve new domains, and to expand the processes of knowledge organization. Participants classified digital game development artifacts from one online and two physical archives. Data were analyzed with mixed methods, generating recommendations for improving the taxonomy and insights on evaluation framework.

Keywords – Metadata, Games, Taxonomy Evaluation

Conference Topics – Innovation

I. INTRODUCTION

Institutions such as the Strong National Museum of Play, the Stanford Libraries, and the National Media Museum in the UK now seek to catalog, classify, and preserve digital games, but are primarily focused on preservation of the final product. Less consideration has been given to the artifacts associated with their development, materials that future researchers, historians, and professionals will rely upon. Development artifacts are vital for study of the medium, helping us understand game design,

intended audience, public reception, and impact on the parent organization. Today, many of these artifacts are born-digital, facing a new set of challenges for archiving. Without organized efforts to preserve such materials, they will be lost.

Researchers at the University of Washington Information School received a National Leadership Grant from the Institute of Museum and Library Services (IMLS) in 2018 to “create a conceptual data model and metadata schema for describing and representing artifacts related to the development of digital games” [1, p. 1]: the Taxonomy of Video Game Development Artifacts (TVGDA). Evaluation of the TVGDA, as a newly developed taxonomy, is ongoing. We contribute to its evaluation by having users apply it to real-life collections and evaluating the results.

There is limited prior research on evaluating methods for taxonomies in library and information science literature. Reference [2] does provide a set of qualities by which a taxonomy may be evaluated (concise, robust, comprehensive, extendible, and explanatory) and that serves as a frame for analysis.

Stated formally, our research questions are: 1) In what manner does the TVGDA exhibit the qualities of the criteria suggested in [2] for evaluating

taxonomies?; 2) What suggestions can be made to further improve the TVGDA?; 3) How can this evaluation of the TVGDA inform and improve upon taxonomy evaluation processes?

Analysis of the TVGDA based on user feedback informs innovative strategies for preservation of materials about digital interactive games, as the TVGDA provides a controlled vocabulary (CV) capable of describing this unique set of information objects known as video game development (VGD) artifacts.

II. LITERATURE REVIEW

Though suggested best practices for CVs and metadata exist, there is no common standard for taxonomy quality and few specifics for testing CV efficacy. “Most evaluation seeks to identify and improve metadata quality, but few attempt to define concretely what ‘quality’ entails” [3, p. 3]. Due to this lack of literature, many taxonomies remain unevaluated [4]. Reference [4] breaks down the evaluative practices discovered through their literature review into five categories: Logical Argument, Expert Evaluation, Action Research, Case Study, and Illustrative Scenario.

Reference [2] recommends a set of attributes by which taxonomies might be evaluated, proposing that good taxonomies are concise (limited enough in detail as to afford easy use), robust (complex enough to differentiate between objects), comprehensive (able to address any object in the domain and/or addressing all aspects of objects in the domain), extendible (able to include new dimensions), and explanatory (“provide useful explanations of the nature of the objects under study” [2, p. 342]). Specific tests for these attributes are not offered. Instead, the authors point out that most appropriate methods for evaluating a taxonomy are dependent on how the taxonomy will be used as implemented. While best practices are likely to emerge as taxonomies are improved and tested over time, there will likely remain no one-size-fits-all approach.

A. Considerations for Classifying VGD Artifacts

For audio and video, organizations such as the Association of Moving Image Archivists (AMIA) assist archivists in describing “Moving Image and Sound Collections” [5]. While these collections have received more attention in recent years (c.f. [6]), existing guidelines lack information on video games and their development artifacts. Reference [7] finds that the closest conceptual standard for the

description of VGD artifacts is in Describing Archives: A Content Standard (DACS) [8], noting there have not been enough accessions of VGD artifacts performed to properly judge its appropriateness. This assertion necessitates content standards specific to VGD artifacts. VGD projects, while creating digital products, began before the “digital revolution... when email blasts replaced circulated paper” [9, p. 85; 10], so their legacy artifacts are both physical and born-digital. The industry has obvious archiving needs for born-digital artifacts, which is an area of classification for which DACS may not yet be well equipped [7]. There continue to be several types of artifacts involved in VGD that are in physical format, too. Thus, tools designed for video game archivists and historians need to consider born-digital, physical, and digitized artifacts. The TVGDA framework is intended to remain stable as technology evolves while providing enough context in scope notes that new forms of VGD artifacts will be classifiable.

B. User-focused Research Principles

Preserving video game information through metadata is a massive challenge [11]. The more we understand video games, the more we recognize the difficulties of applying current standards and rules to describe them, and even more so for VGD artifacts.

There have been initiatives to improve the organization and description of games and VGD artifacts. For example, [12] established the Video Game Metadata Schema (VGMS) as a “list of elements which form a metadata schema for describing video games.” Reference [13] utilized user interviews to “derive and discuss key design implications for video game information systems [(IS)]...” [p. 833] to improve game-related IS. These user-focused research projects form the basis of similar metadata and taxonomical structures related to video game digital assets and ephemera.

The TVGDA is one such example, specifically targeted at organizing and describing VGD artifacts. The TVGDA was created for three classes of users: industry professionals, information professionals, and game researchers. It has a single dimension, used to describe a VGD artifact’s type. Taking its warrant from industry use, the TVGDA “is organized into three broad sections including (a) Development (with seven subsections), (b) Organization-Related Materials, and (c) Marketing (with four subsections), representing different aspects and timelines of game

development" [7, p. 548]. It includes 123 industry terms with scope notes and additional lead-in terms.

III. METHODS

Per [8], TVGDA evaluation began by applying the taxonomy to 1,000 VGD objects, supplemented by two expert evaluation interviews. Data gathering took place as a graduate-level class cataloging assignment based on an online archive of VGD artifacts and a pair of individual tests on institutional collections. These methods allowed us to test consistency of use as well as applicability across multiple collections. The expert evaluation interviews were conducted as follow-up interviews on the latter two collections.

IV. FINDINGS AND DISCUSSION

We used the criteria in [2] to frame our evaluation: We consider whether the TVGDA is comprehensive, concise, robust, explanatory, and extendible.

A. *Comprehensiveness*

Comprehensive is the quality of covering all objects in the intended domain, or of including enough dimensions to describe the domain [4].

Coverage. At 123 terms, the TVGDA is an extensive representation of VGD artifacts. Analysis shows far more difficulty deciding between terms than finding an applicable term. That said, results did include indications of possible missing terms.

There were more than 20 comments from graduate catalogers requesting or suggesting new terms, often narrower terms for artifacts with niche purposes. These suggestions usually came from working with difficult-to-describe materials, including game control or navigation graphics, tables of contents, barcodes, and physical comic books. Still, these suggested terms may be requesting a level of specificity that may not be necessary in the TVGDA.

Dimensionality. Comprehensive taxonomies should contain enough dimensions to describe the domain. Specifically, [2] says: "a useful taxonomy includes all dimensions of objects of interest" [p. 341]. This is differentiated from robustness, defined as, "enough dimensions and characteristics to clearly differentiate the objects of interest" [p. 341]. Comprehension and robustness must be balanced in each taxonomy. The TVGDA is a unidimensional

expression of an artifact's type in a highly specific domain, but institutional implementations would certainly include other dimensions (also present in the student cataloging exercise). We continue the analysis of dimensionality below, under Robustness.

B. *Conciseness*

Concision is the quality of parsimony and limited complexity and is at tension with the quality comprehensiveness [2]. Graduate catalogers were not asked to review this quality of the TVGDA but the research team notes that the TVGDA prefers comprehensiveness to concision.

The sheer number of terms may also be a factor in relatively low intercoder agreement (see Explanatory Power, below) as catalogers fall back on familiar terms where less-used terms may be more appropriate. The TVGDA's caretakers should consider user studies to reduce the number of broad terms, easing the conceptual load required for high-level classification of artifacts while allowing interested parties to use narrower terms for detail.

C. *Robustness*

Robust is the quality of allowing catalogers to differentiate between objects: "enough dimensions and characteristics to clearly differentiate the objects of interest" [2, p. 341].

Test catalogers had problems choosing terms in this unidimensional taxonomy. For example, participants vacillated between *screenshot* ("Image captured from a game during play" [14, p. 6] and *art asset* ("Any artwork used in a released version of the video game, such as 3D models or 2D artwork" [14 p. 5]). The intended purposes of an image may be required to determine whether any given image is a *screenshot* or one of the other graphical artifact types.

The all-digital online archive highlights another issue with images: Catalogers sometimes classified the artifact as an image and sometimes as the object depicted. Cataloging standards help information professionals understand how to classify these objects, but the TVGDA is intended to be used by creators as well.

To make the TVGDA more robust, the research team suggests that its caretakers consider separate terms for nature and function, add additional guidance on which aspect should take priority, consider guidance on the use of multiple terms for

single artifacts, and clarify the function of digital representations of physical artifacts for creators.

D. Explanatory Power

Explanatory is the quality of adequately describing the domain, providing “useful explanations of the nature of the objects under study” [4, p. 342]. We compared the behavior of multiple catalogers to see if they shared a common understanding of the items described. Quantitative analysis provides insights.

Since the artifacts were distributed to multiple groups who further distributed the work to members, assumptions for use of Cohen’s kappa (sets being compared being completed by one and only one coder [15]) were not met and could not be used. Additionally, many students selected no terms or multiple terms, so these pairings were omitted. Analysis used simple statistics regarding matches between valid pairs, but at multiple levels.

Many catalogers labeled items as “ambiguous” and required further analysis. Participants often noted that a lack of context meant they could not accurately assign any terms. Instead, they relied on several other indicators to attempt a classification, including inferences from file names and online searches for authoritative information. While failure to agree on understood objects shows room for improvement in the TVGDA, its caretakers have little ability to control the ambiguity of the nature of a given item, or its lack of context in an archive.

E. Extendibility

Extendable is the capacity for a taxonomy to be revised. The TVGDA is a single-dimension taxonomy, and there are no barriers to extending it in terms of adding another identified dimension or adding sections or terms to the existing dimension. In this regard, the TVGDA is extendable.

V. THE EVALUATION FRAMEWORK AND BROADER IMPLICATIONS

Comprehensiveness represents two qualities—coverage and dimensionality—which were treated separately. We found no benefit to considering them as a single quality. Coverage was the most intuitive sub-quality for the research team to grasp and evaluate: Have participants identified artifacts which cannot be satisfactorily classified with the current taxonomy? Dimensionality is difficult to address in a

unidimensional taxonomy without specific participant feedback, and we had none.

Conciseness is in tension with comprehensiveness and is tied to the cognitive load required to apply the taxonomy. Aside from the two interviews, we did not inquire directly about cognitive load, though we believe quantitative analysis of cataloger agreement offers us some insight in this work. We believe guidance for how and when to use broader or narrower terms should help situationally balance conciseness with comprehensiveness.

Robustness seems tied to multi-term classification with this unidimensional taxonomy. The volume of multi-term suggestions and lack of coder agreement seem like inverse measures for robustness, indicating room for improvement.

Explanatory power was evaluated here by quantitatively evaluating cataloger agreement, as a sufficiently explanatory taxonomy will provide a common understanding of the domain and make classification easier. This ties explanatory power to robustness: Lacking full explanatory power in taxonomy, our users often found that additional term selection (tied to robustness) helped them explain an artifact. This is complicated by the presence of ambiguous artifacts: those whose nature is explained through the taxonomy but not understood by evaluators due, perhaps, to lack of experience in the domain or with the specific collection.

Many concerns found in the research seemed to have multiple modes of evaluation. Troubles identified with *screenshot* could situationally apply to robustness or comprehensiveness, for example. Still, the team found these qualities to be useful evaluative concepts and intend to use them in the future to help researchers formulate better methods for their evaluations for any given specific taxonomy.

VI. CONCLUSION

Our approach to assessing this taxonomy combines the meaningful elements of established heuristics and user testing to create situated vocabularies and taxonomies to establish best practices for defining the breadth and depth of relevant artifacts across evolving domains. The TVGDA offered the opportunity to contribute through evaluation of both a new user-centered taxonomy and a framework of taxonomy qualities,

making recommendations for improvement of both. Focus on the other two user groups (creators and researchers) represent an additional opportunity for expanding tests of the TVGDA. VGD is constantly evolving and intersectional, producing cultural texts by developing new technology and spanning physical and digital media in a relatively unexplored domain that has become academically legitimized only relatively recently. As such, it offers fertile ground for designing and evaluating structures to help creators, information professionals, and other users organize and share the domain, and to expand our knowledge of knowledge organization as well.

ACKNOWLEDGMENT

This project was made possible in part by the Institute of Museum and Library Services.

REFERENCES

- [1] LG-86-18-0060-18, <http://www.ims.gov/grants/awarded/lg-86-18-0060-18-0>, last accessed 2021/09/13.
- [2] Nickerson, R.C., Varshney, U., Muntermann, J.: A method for taxonomy development and its application in information systems. *European Journal of Information Systems*. 22, 336–359 (2013).
- [3] Lee, J.H., Clarke, R.I., Perti, A.: Empirical evaluation of metadata for video games and interactive media. *Journal of the Association for Information Science and Technology*. 66, 2609–2625 (2015).
- [4] Szopinski, D., Schoormann, T., Kundisch, D.: Because Your Taxonomy Is Worth It: Towards a Framework for Taxonomy Evaluation. *Research Papers*. (2019).
- [5] Cocciolo, A.: Moving image and sound collections for archivists. (2018).
- [6] AIMS Work Group: AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship, https://dcs.library.virginia.edu/files/2013/02/AIMS_final_text.pdf.
- [7] McDonald, C., Schmalz, M., Monheim, A., Keating, S., Lewin, K., Cifaldi, F., Lee, J.H.: Describing, Organizing, and Maintaining Video Game Development Artifacts. *Journal of the Association for Information Science and Technology*. 1–14 (2020).
- [8] Society of American Archivists: Describing Archives: A Content Standard (DACS), <https://www2.archivists.org/groups/technical-subcommittee-on-describing-archives-a-content-standard-dacs/describing-archives-a-content-standard-dacs-second->.
- [9] Williams, J.A., Berilla, E.M.: Minutes, Migration, and Migraines: Establishing a Digital Archives at a Small Institution. *The American Archivist*. 78, 84–95 (2015).
- [10] Wolf, M.J.P. ed: *The video game explosion: A history from PONG to Playstation and beyond*. Greenwood Press, Westport, Conn (2008).
- [11] Winget, M.A.: Videogame preservation and massively multiplayer online role-playing games: A review of the literature. *Journal of the American Society for Information Science and Technology*. 62, 1869–1883 (2011).
- [12] Lee, J.H., Perti, A., Clarke, R.I., Windleharth, T.W., Schmalz, M.: UW/SIMM Video Game Metadata Schema Version 4.0. (2017).
- [13] Lee, J.H., Clarke, R.I., Rossi, S.: A qualitative investigation of users' discovery, access, and organization of video games as information objects. *Journal of Information Science*. 42, 833–850 (2016).
- [14] Lee, J.H., McDonald, C., Schmalz, M., Windleharth, T., Keating, S., Monheim, A., Cifaldi, F., Lewin, K.: Taxonomy of Video Game Development Artifacts. (2020).
- [15] Hoyt, W.T.: Interrater Reliability and Agreement. In: Hancock, G.R. and Mueller, R.O. (eds.) *The reviewer's guide to quantitative methods in the social sciences*. Routledge, New York (2010).

ARCHIVER

Sustainable Preservation of Scientific Data

Matthew Addis

Arkivum Ltd
UK
matthew.addis@arkivum.com
[0000-0002-3837-2526](tel:0000-0002-3837-2526)

Teo Redondo

LIBNOVA
Spain
teo.redondo@libnova.com
[0000-0001-6465-7771](tel:0000-0001-6465-7771)

João Fernandes

CERN
Switzerland
joao.fernandes@cern.ch
[0000-0002-0445-7038](tel:0000-0002-0445-7038)

Abstract – The ARCHIVER project (Archiving and Preservation for Research Environments) has spent over 3 years designing, prototyping and piloting innovative new services for the Long Term Digital Preservation (LTDP) of scientific datasets. During the project, multiple Data intensive organizations representing several research domains (CERN, DESY, PIC and EMBL-EBI) have worked closely and collaboratively with suppliers (Arkivum and LIBNOVA) on the research and development of new services and solutions for scientific data preservation relevant for the European Open Science Cloud (EOSC). This panel session will see the ARCHIVER project participants discuss and share the experience and lessons learned during the project. Topics will include: the benefits of a collaborative approach between end-users and commercial suppliers; the challenges that were addressed along the way and the solutions that were created; and what still needs to be done in order to realize the project vision of sustainable digital preservation services for the whole scientific community that address the needs of organizations both large and small. iPRES 2022 comes just two months after the end of the final Pilot phase of the ARCHIVER project which makes it an ideal time for the ARCHIVER participants to share their insights and experiences with the wider LTDP community.

Keywords – Digital Preservation, Scientific Data, Trusted Digital Repository, Sustainability, Scalability
Conference Topics - Environment; Innovation.

I. INTRODUCTION

With a procurement budget of 3.4 million euros, the ARCHIVER project [1] has used a Pre-Commercial Procurement (PCP) approach to competitively procure R&D services from a range of vendors in order to create new services and solutions for LTDP of scientific datasets. The three stages of the project

cover design, prototyping and pilots and have taken place from Jan 2019 - Jun 2022.

ARCHIVER is driven by the needs of a diverse range of stakeholders including CERN, who operate the Large Hadron Collider near Geneva, DESY (the Deutsches Elektronen-Synchrotron, based in Hamburg and Berlin), the EMBL-EBI (European Bioinformatics Institute, based in Cambridge), and PIC (Port d'Informació Científica, situated near Barcelona).

The importance and benefits of making scientific data open and reusable according to FAIR principles [2] (Findable Accessible Interoperable Reusable) has been clear for some time now. Practical advice and guidelines are now available from initiatives such as the FAIRsFAIR [3] project. However, there are still major gaps [5] when it comes to long-term accessibility and usability of research data, for example as discussed in the FAIR Forever report [4] from the Digital Preservation Coalition. These gaps put research data at long-term risk, they prevent the construction and operation of sustainable Trusted Digital Repositories [6], and they affect organizations both large and small who are tasked with being custodians of valuable research data resources. This is what the ARCHIVER project sets out to solve.

The aim of ARCHIVER is to achieve substantially improved archiving and digital preservation, not just for petabyte-scale data-intensive research, but also for the Long Tail of Science (LToS) [7]. To support the requirements of European scientists, ARCHIVER provided R&D funding to European SME sector specialist to stimulate new end-to-end archival and preservation services for the vast and ever-growing datasets generated by world-leading research

institutions. Reflecting the move toward large-scale collaborative research supported by cloud infrastructures, ARCHIVER embraces and tackles issues such as scalable and interoperable LTDP services in the cloud, new business and commercial models for archiving, accessing and reusing large datasets, and crucially how to do this in a way that is both economically and environmental sustainability.

This panel will discuss the ARCHIVER project, the results that have been achieved, the approaches that were taken, what worked and what didn't, what still needs to be done to make LTDP a reality for the scientific data community, and how the ARCHIVER experience and approach could be translated into other domains and markets. The aim of the panel is to follow the ARCHIVER spirit of being open, honest and transparent and to share our experience and thoughts with the community as a whole.

II. PANEL DISCUSSION TOPICS

The panel will discuss a range of topics and questions that include:

- What worked well in ARCHIVER (and what needs improvement) when the suppliers and end-users collaborated together using the project's pre-competitive R&D approach? Do you think this could be a good template for other sectors?
- What were the main challenges in ARCHIVER and how were they overcome? Were they organisational, technological, economic or a mix of all these things?
- What are the three most important features of each of the ARCHIVER resulting solutions? How did the end-users articulate their priorities, and how did the suppliers go about building their solutions?
- How can LTDP services support organisations who produce research data from the large scale through to the long tail, for example in the context of initiatives such as the European Open Science Cloud? Are there economies of scale, is there a one size fits all, and how can the LToS benefit from the services developed in ARCHIVER that are primarily for large organisations?
- Is it possible to preserve and provide access to huge volumes of data in a way that is environmentally sustainable? Does the cloud help, or does it make it harder? What does the carbon footprint of LTDP look like in practice?
- How does digital preservation fit into making data FAIR Forever? How do digital preservation

standards and good practices help organizations build trusted repositories?

- The value of scientific data is often in its reuse, for example re-running computations and applications against archived data. How do the ARCHIVER resulting services address the need to both preserve data and at the same time support active access and reuse?
- ARCHIVER has done a lot of work on technological solutions, but what about the economics and business models? What commercialisation approaches do the ARCHIVER team foresee for the resulting services developed in the context of the project?
- Research data lives for longer than any vendor, system or technology. How do the ARCHIVER resulting LTDP services prevent vendor lock-in and encourage portability and interoperability, yet at the same time make it attractive for new commercial services to enter the market? Are these in conflict with each other?
- Do you think that the lessons learnt and the solutions developed in ARCHIVER are transferable from ARCHIVER to other disciplines and domains? What would be your number one recommendation?

REFERENCES

- [1] www.archiver-project.eu
- [2] Findable Accessible Interoperable Reusable (FAIR). <https://www.go-fair.org/fair-principles/>
- [3] <https://www.fairsfair.eu/>
- [4] A. Currie, W. Killbride. "FAIR Forever? Long Term Data Preservation Roles and Responsibilities, Final Report". Feb 2021. <https://zenodo.org/record/4574234>
- [5] J. Fernandes et al. "ARCHIVER D2.1- State of the Art, Community Requirements and OMC Results ". Jan 2020. <https://zenodo.org/record/3618215>
- [6] D Lin et al. "The TRUST Principles for digital repositories". Scientific Data. 7, Article 144. May 2020.
- [7] M. Devouassoux, B. Jones, J. Fernandes. "Long-Tail-of-Science's Requirements for Commodity Cloud Services in Europe" Oct 2019. <https://zenodo.org/record/3564668>

HOW CAN BRINGING TOGETHER THE WORKFLOWS OF PUBLISHING & PRESERVATION LEAD TO BETTER, LONGER-TERM SOLUTIONS THAT BENEFIT BOTH?:

A Panel with COPIM Work Package 7, the Embedding Preservability in New Forms of Scholarship Project (NYU), and Project JASPER

Dr Miranda Barnes

Loughborough University/
COPIM
United Kingdom
m.l.barnes@lboro.ac.uk

Karen Hanson

Embedding Preservability
USA
karen.hanson@ithaka.org

Alicia Wise

Project JASPER
United Kingdom
awise@clockss.org

Abstract – Rather than preservation and archiving being an afterthought for digitally published works, research is being done to explore how the concepts, processes, and requirements of preservation can be embedded into publishing, especially OA publishing. How might this be integrated further, and what benefits might extend to academics and researchers themselves? Often the difficulties or challenges of preservation result from scholarly research being generated and published without a preservation policy in mind, which can result in the knowledge becoming lost. This has a particularly emphasised effect upon smaller and scholar-led presses, who often do not have the inbuilt resilience typically provided by either a large business model or a memory institution, which can allow for archiving and preservation to occur procedurally. Our panel will consider the workflows involved, potential solutions, and what additional engagement may be necessary to increase awareness among publishers and researchers. COPIM's Work Package 7 engages with complex digital OA monographs and the scholar-led publishing community. The Mellon-funded Embedding Preservability in New Forms of Scholarship project (NYU) embeds digital preservation experts with publishers from the beginning of the publishing process to help them to make choices that result in publications, including very complex ones, that can be preserved at scale. And Project JASPER works with small, independent OA journals to facilitate preservation.

Keywords – open access, digital monographs, digital preservation, archiving, scaling small

Conference Topics – Resilience; Community

I. INTRODUCTION: OA AND MIA?

More and more monographs and articles are published in both a born-digital fashion, as well as open access, so the issue of archiving and preservation becomes more pressing. As has been regularly noted, particularly in the work of Project JASPER and others, high-value resources and important scholarly knowledge can easily disappear from the internet. If a journal or OA monograph press folds and they had no active preservation workflow, the content is likely gone forever.

Laakso's article 'Open is not forever: A study of vanished open access journals'¹ (2021) discusses the shift in responsibility, and the uncertainty surrounding it, for the preservation of born-digital books and articles, which has in part led to the loss of a multitude of articles and monographs. Laakso's study found that in terms of academic journals, 174 OA journals had disappeared from the internet since 2000. Though to date no similar study has been performed for OA monographs, the need for infrastructure is clear: UUK's Open Access Monograph's Group 2019 paper, while only briefly touching upon archiving and preservation, states that "a robust infrastructure needs to be in place to ensure digital outputs are preserved."² The 2017 Knowledge Exchange Landscape study on open access and monographs found that "82% of the interviewed libraries to 'strongly agree/agree' with

the development of a central OA monograph repository for their respective country.”³

II. SOLUTIONS, CHALLENGES & THE “AFTERMATH EFFECT”

Though solutions exist for archiving and preservation, such as Portico, CLOCKSS, and similar services, smaller publishers of both journals and monographs often lack the awareness or resources to participate in some form of long-term preservation. COPIM’s Work Package 7 found levels of inconsistency within the preservation practices of the publishers surveyed in their workshops and interviews, which has a direct correlation to the uncertainty around who is responsible in the digital publishing landscape for preserving the scholarly record. DOAJ, the central hub for Project JASPER, found 7500 journals in their platform with no preservation – evidence that the “long tail” of smaller publishers is more at risk of disappearing. DOAJ is one of the five organisations working together under Project JASPER towards a solution, alongside CLOCKSS, the Keepers Registry, PKP (Public Knowledge Project), and the Internet Archive, to formulate and deploy three preservation options for OA journals with no current preservation in place.

Additional challenges result from what we will call the “aftermath effect” in preserving digital monographs. Most preservation activities, by necessity and overarching practice, occur in the aftermath of publication, which means that it is a retrospective action that must respond to the content already created. With this comes issues with a variety of file formats, software, and the general multitude of content types and methods that have become possible via the lightspeed advance of digital publishing technology over the last twenty to thirty years. These issues become even more evident when examining experimental or complex digital monographs, which may contain embedded audio-visual content, geospatial data, or be created within a specific software or platform, moving beyond the traditional boundaries of a ‘book’ and the more traditional digital format of a PDF.

III. RESPONSIVE RESEARCH: ENGAGEMENT

Responding to these quandaries means engaging with publishers, academics, and publishing-software developers at an earlier stage, to increase awareness of what preservation means and why it is important, and to introduce the

implications of preservation’s limitations and requirements. As well as preservation being a flexible and responsive process, this can allow for publishing and content creation to incorporate and engage with preservation as an essential part of their processes, and allow for preservation to be an extended form of knowledge dissemination and reusability. The Embedding Preservability project, which follows on from the Preserving New Forms of Scholarship project at NYU, will embed four preservation experts with various publishers and their publishing-software developers. This team will directly observe and participate in the technology-decision process and learn the editorial and production workflow at each publisher in order “to identify opportunities for implementing changes that favor preservation during the creation process.”⁴ (Hanson, 2021)

While the work of Project JASPER engages the small OA journal publisher, and Embedding Preservability primarily involves OA publications at small- and mid-size university presses, the smaller or scholar-led open access monograph publisher is where COPIM is positioned. COPIM’s focus is on scaling small and providing further support and guidance for small and scholar-led publishers in order to assure equity in the publishing and preservation landscape. The resource challenges for these small publishers, in terms of finance and staff, but also technology, remain and WP7 is working to develop guidance and solutions to assist these publishers with the archiving and preservation process, as well as advocate for longer term, more centralised infrastructure. Also, the role of the academic researcher is one we hope to more actively involve. While at present the majority of engagement is directed at publishers in order to increase their understanding of the importance in having an archiving and preservation policy, there is definite scope for engaging researchers and academics. What steps could be taken to engage researchers in the processes of preservation and increase awareness? What level of responsibility do researchers have in understanding the future preservation of their work? What might be the best avenues to reach researchers to convey these concepts and engage their participation?

REFERENCES

- [1] Laakso, M., Matthias, L. and Jahn, N., 2021. Open is not forever: A study of vanished open access journals. *Journal of*

the Association for Information Science and Technology, 72(9), pp.1099-1112.

- [2] 2019. *Open Access and Monographs*. [Online] London: Universities UK Open Access and Monographs Group, p.8. Available: <https://www.universitiesuk.ac.uk/sites/default/files/field/downloads/2021-07/open-access-and-monographs.pdf>
- [3] Ferwerda, E., Pinter, F. and Stern, N., 2017. *A landscape study on open access and monographs: Policies, funding and publishing in eight European countries*. [Online] Bristol: Knowledge Exchange, p.126. Available at: <https://doi.org/10.5281/zenodo.815932>
- [4] K. Hanson, "Partnering with publishers to break down barriers to preserving new forms of scholarship," Digital Preservation Coalition, blog, November 3, 2021 [Online]. Available: <https://www.dpconline.org/blog/wdpd/khanson-wdpd21>

A LABOR OF LANGUAGE

Building The Global Preservation Community Through Funded Translation Projects

Rebecca Fraimow

GBH Archives
United States
rebecca_fraimow@wgbh.org
[0000-0003-4025-9503](tel:0000-0003-4025-9503)

Juana Suárez

New York University
United States
juanar@nyu.edu
[0000-0002-4574-4738](tel:0000-0002-4574-4738)

Pamela Vízner

AVP
United States
pamela@weareavp.com

Lorena Ramírez-

López

Webrecorder
United States / Chile
DaleLoreNY@gmail.com
[0000-0003-1297-1990](tel:0000-0003-1297-1990)

Abstract – Translation of digital preservation documentation into other languages is an invaluable tool for global knowledge exchange in the field. In this panel, professionals working on translation projects in the preservation field will discuss the value of translation, what best practices around these projects look like, and methods for sustainably supporting translation efforts without reliance on volunteer labor.

Keywords – documentation, translation, metadata, labor

Conference Topics – community, exchange

I. INTRODUCTION

In the constantly evolving field of digital preservation, documentation around best practices, and global standards provides a crucial tool for institutions as they develop their preservation workflows; it's also vital for the cross-institutional exchanges, partnerships and projects that support and strengthen the field. However, language barriers severely limit the accessibility of this documentation. Additionally, when translation projects do occur, they often rely on volunteer labor, placing additional burdens on bilingual preservation professionals who hope to make documentation accessible beyond the hegemony of English as a global language.

This panel will present a discussion with translators and partners on translation projects in the preservation field. Speakers will each briefly present on recent projects translating preservation documentation, with an emphasis on preliminary planning, workflows, and presentation/information exchange, before entering into a discussion of best practices and sustainability to support translation projects and more broadly within the field.

Rebecca Fraimow and Lorena Ramírez-López will discuss the NEH-funded translation of documentation around the PBCore cataloging standard into Spanish. Topics will include the process of writing translation funding into the grant, decision-making around direct translation versus re-creation of example documentation in different contexts, and the process of outreach and website development to ensure that Spanish-language users are able to access the documentation.

Pamela and Juana will discuss some relevant issues related to the process of translating technical documents such as the suitability and credentials of the translator vs. the misconception that any native speaker can do the translation; the difference between translating different types of documents, in particular, those related to digital technologies; the

importance of good writing skills in the target language and ability to adapt writing to the different variations and nuances of the target language (Spanish in this case). All facts considered, the discussion makes emphasis on the costs of a translation project when utmost professional standards are expected, and also reminds the importance to include texts from other languages to English as part of diversifying efforts in the profession. Finally, we will discuss the importance of prioritization and selection of content to be translated for major impact and usability, as documentation can have embedded practices and workflows, whose translation can imply an assumption that these practices can be implemented in any context, regardless of language and preservation methods used.

By entering into an open discussion of the labor involved in translation projects and best practices for ensuring useful outcomes, the panelists hope to provide valuable examples for ensuring the sustainable continuation of this work in the future and further open up possibilities for global exchange within the field.

RIGHT CLICK TO PRESERVE

Preservation, NFTs, and Distributed Ledgers

John Bell

*Dartmouth College
United States*

*john.p.bell@dartmouth.edu
[0000-0003-0514-1585](tel:0000-0003-0514-1585)*

Regina Harsanyi

*Independent Conservator
United States*

regina.harsanyi@gmail.com

Jon Ippolito

*University of Maine
United States*

jippolito@maine.edu

Abstract: Artists have experimented with cryptocurrency-incentivized distributed ledgers such as blockchains since the advent of Bitcoin. In parallel, crypto advocates have claimed that distributed ledger protocols will ensure an accessible and immutable record of anything registered to it, including artwork. This panel examines this idea with nuance, neither buying into the mass deception behind NFT marketing nor rejecting the reality that a subset of artists are creating significant, challenging works that inherently utilize these technologies. Each of the three panelists considers a specific approach to how preservation professionals can keep such works alive. We then jointly compare their merits in a variety of preservation contexts.

Keywords: NFT, Blockchain, Emulation, Variable Media

Conference Topics – Innovation.

I. DISTRIBUTED LEDGERS AND PRESERVATION

A claim frequently repeated in the last year by advocates of blockchains and NFTs is that distributed cryptographic ledgers will ensure an accessible and immutable record of born-digital and digitized art for posterity. The reality, borne out by both analysis of cutting-edge experiments as well as historical precedent, is less rosy. While the field is still in its infancy, the panelists will forecast the viability of the most promising proposals for preserving, and being preserved by, blockchains.

Media coverage and competing narratives around NFTs have clouded collective understanding of these concepts, making it important to clarify what is meant by these terms:

A **distributed ledger** is a database that is consensually shared and synchronized across multiple nodes. Any changes or additions made to the ledger are reflected and copied to all participants in a matter of seconds or minutes.

A **blockchain** is the most popular example of a distributed ledger, using a block data structure with the primary objective to record verifiable transactions.

An **NFT** (Non-Fungible Token) is a set of electronic instructions called a “smart contract,” with a unique cryptographic hash published to a distributed ledger that can reference virtually any object, tangible or digital. Misperceptions notwithstanding, most NFTs do not contain or convey rights to media files, but merely point to them.

A **distributed storage protocol** is a peer-to-peer network for storing and sharing data, distributed on multiple file servers or multiple locations. It allows users to access or store isolated files redundantly and make them available programmatically.

II. APPROACHES TO PRESERVING NFTs AND THEIR ASSETS

A. Archival Packages

The first and only cross-chain models in practice for preserving non-fungible token data and associated assets were created and put into practice by Protean, a variable media art conservation initiative for educating, practicing, and publishing novel standards for emerging technologies in both the public and private sector. The Protean method dodges platform impediments by preparing archival packages off-chain first and working with artists to generate URIs for long-term maintenance plans, which are then referenced in the smart contract metadata. They often hold multiple files, including uncompressed, lossless archival copies, README.txt files, detailed manuals, technical artist questionnaires that emphasize media-specific assessment, and supplemental documents.

The history of net-based art conservation has exposed many tales of link rot. Protean takes this into account by educating artists and art custodians on generic digital strategies such as migration, but also newer tactics such as becoming Filecoin and IPFS node operators themselves.

This process also takes into account the reality that immutability is antithetical to media art preservation. Unlike the singular asset associated with typical NFTs, a Protean link might point to a range of releases that are added to an artist's server or an IPFS directory as the work is updated over time.

B. Emulation

Though it may seem a counterintuitive approach to preserving a distributed system, it is possible to emulate an entire set of blockchain nodes to preserve a distributed ledger on a centralized system. Emulation may be appropriate once the original chain or supporting dependencies become unavailable or somehow in dispute: a particular ledger system may lose popularity and not have enough participating nodes to continue, the ledger itself may fork for technical or social reasons, or external links stored in the ledger may go stale.

Running several containerized nodes together on a single machine is simply a matter of launching multiple iterations of the same container, each of which can then be virtually networked with the others.[1]

Technical convenience does not imply emulation is always the best solution, however. An emulated private node network only preserves data in the ledger itself; since ledger data is often a pointer to outside media, code, or assets, those outside resources may also need to be emulated. A work may depend on the activity of public users or new data being added to a ledger that is now isolated from the outside world. In cases where such considerations are not critical to the work, though, emulation offers the ability to simply reproduce a complex technical ecosystem.

C. Media-independent assessment

Another integration of crypto art into current conservation practices would be to take the approach of the Variable Media Questionnaire,[2] which records opinions about how to preserve creative works when their current medium becomes obsolete.

The current version of the Questionnaire looks at artworks as ensembles of Parts, though its purpose is less to track sundry gadgets like cables or disk drives than to understand the key elements of a work that are critical to its function, such as source code or media display. Structuring the Questionnaire in this way makes it easier to compare different artworks created with similar parts.

The Variable Media Questionnaire could be leveraged to help preserve blockchain art by creating a dedicated Package. A variable media Package is an ensemble of Parts that might be used in common creative formats such as video installations and websites. Some of the questions in a blockchain Package would overlap with other formats, such as the resolution or color depth of a digital image, or what happens to user contributions when a work is loaned. Others might be specific to the blockchain, such as what to do when a chain forks and how critical a given cryptocurrency is to the work's function.

III. CONCLUSION

The vulnerabilities of distributed ledgers suggests that—outside of works created on the blockchain itself—such systems are unlikely to suffice as a preservation solution for art in the future, and in fact may make a preservationist's job more difficult. Nevertheless, the panelists believe that a significant slice of artworks using these technologies are worth saving for the future. Rather than a one-size-fits-all preservation solution for these works, we recommend a case-by-case analysis of the best way to translate the artistic qualities of the work into future scenarios in which the original chains and associated services are defunct.

REFERENCES

- [1] The go-ethereum Authors. 2021. "Private Networks." April 06. Accessed January 29, 2022. <https://geth.ethereum.org/docs/interface/private-network>.
- [2] Forging the Future. n.d. The Variable Media Questionnaire. Accessed January 29, 2022. <http://variablemediaquestionnaire.net>

CORETRUSTSEAL v3.0

IN A PRESERVATION AND COMMUNITY CONTEXT

Jonathan Crabtree

*Jonathan Crabtree
Odum Institute Data Archive
USA
jonathan_crabtree@unc.edu
[0000-0002-0139-7025](tel:0000-0002-0139-7025)*

Ingrid Dillo

*DANS Data Archiving &
Networked Services
Netherlands
ingrid.dillo@dans.knaw.nl
[0000-0001-5654-2392](tel:0000-0001-5654-2392)*

Hervé L'Hours

*UK Data Service,
Consortium of Social Science
Data Archives (CESSDA)
United Kingdom
herve@essex.ac.uk
[0000-0001-5137-3032](tel:0000-0001-5137-3032)*

Abstract – The authors will present the CoreTrustSeal Requirements including changes in version 3.0 in the wider context of an ongoing task force to promote digital preservation and an emerging community of trustworthy digital repositories. The input from the iPres audience during the panel discussion will be collated, aligned and fed back into these active areas of assessment, community building and policy development.

Keywords – CoreTrustSeal, Trustworthy Digital Repository, Preservation, Community

Conference Topics – Community; Innovation.

auspices of the Certification of Repositories Interest Group [7]. Version 2.0 of the CoreTrustSeal Requirements sought to support the transition from prior certifications while integrating the experience of initial applicants, reviewers and Board members. Version 3.0 of the Requirements for 2023-2025 emerges at a critical point for repositories, the preservation community and global scientific infrastructure. The CoreTrustSeal has been recognised [8] as an exemplar certification solution and as an enabler of FAIR [9] (Findable, Accessible, Interoperable and ReUsable) digital objects of all kinds.

I. INTRODUCTION

The release of CoreTrustSeal version 3.0 represents a significant step in Trustworthy Digital Repository (TDR) standards and certification; this work exists within a wider ecosystem of data services and community needs that are best addressed through open discussion and cooperation. The revised requirements are presented and further context is provided through the current preservation task force and community development that will support the wider vision of trustworthy research data infrastructures.

II. CORETRUSTSEAL v3.0

The CoreTrustSeal [1] Requirements [2] emerged [3] as the result of a goal set by the Research Data Alliance (RDA) [4] to deliver a single, sustainable, low barrier to entry set of 'core' TDR Requirements, and Certification process. Two [5] [6] prior assessment approaches were integrated and improved through an open community working group under the

A CoreTrustSeal certification is valid for three years, independent of any changes to the Requirements. The Requirements themselves are subject to community review and revision every three years; new and renewing applicants assess against the most recent version. The Revision process is led by the CoreTrustSeal Board, itself selected from the Community of Reviewers, each of whom represents a successful CoreTrustSeal repository. Submissions for the revision of Requirements are open to all individuals and repositories though bodies such as the RDA are key stakeholders and enablers of the process. Suggestions for changes must be evaluated in light of several, sometimes competing priorities: the rapidly changing nature of data and information management infrastructure, the presence of clear community expectations against which assessments can be made, and the need to deliver a low-barrier to entry and 'core' set of requirements. Like other TDR standards [10] [11], the CoreTrustSeal focuses on

sustainable organizational and technical infrastructure that provides for digital object management services that facilitate the long-term preservation of digital assets for a defined designated community of users.

Version 3.0 addresses a number of proposals raised by the Board but are subject to community review and feedback. In addition to structural and textual changes there are updates to improve clarity and maintain alignment with the repository and data infrastructure landscape. The panel will begin with the requirements, key changes and associated work to support reviewers and applicants being presented.

III. CORETRUSTSEAL IN CONTEXT

The CoreTrustSeal seeks to provide a sustainable global service for managing requirements and a certification process management for TDRs. Current challenges include support for a full range of generalist and specialist (e.g. disciplinary) repositories, interactions with a range of non-TDR data services, the machine-actionable assessment of TDR certification status and how more specific and detailed requirements can be developed around the 'core' of CoreTrustSeal.

The CoreTrustSeal's mission, and the way it approaches these challenges, must be addressed within the wider context of the communities providing digital object management and the other activities that seek to drive awareness and delivery of sustainable long term digital preservation. The other speakers' presentations emerge from the perspective of the European Open Science Cloud (EOSC) [12], where a number of projects [13] have offered support to CoreTrustSeal applicants and have provided valuable discussions on the topics of preservation [14]. Many aspects of the EOSC are reflective of a global trend for consolidation of research infrastructure and a greater need for shared expertise, outsourcing and interoperability.

IV. CORETRUSTSEAL, FAIR AND PRESERVATION

Trustworthy repositories are a critical dependency for the full lifecycle of interoperable research infrastructures and they also have a long history of involvement in the development of best practices, standards and assessment. But repositories are not the only actors in this space and there is ongoing work to define other types of data services [15] and

the degree to which their compliance with standards should be assessed and certified. [16]. The FAIR Principles themselves do not explicitly address preservation. Work to align CoreTrustSeal with FAIR [17] has demonstrated the need for active preservation to ensure digital assets of all types remain FAIR over time. The FAIR Forever report [18] concluded that the emerging EOSC vision lacks clarity around digital preservation, and made a number of significant recommendations for different actors. The first of these, directed towards the EOSC Secretariat, was to "establish a working party or task group, reporting directly to the EOSC Association Board with respect to digital preservation". The resultant EOSC Association Long Term Data Preservation (LTDP) Task Force [19] is now in place. Delivering on these goals will involve a range of international actors across policy makers, funders, repositories, and other data service providers. The FAIRsFAIR project has taken a significant first step in its Coordination Plan for a sustainable network of FAIR-enabling Trustworthy Digital Repositories [20].

A. LONG TERM DATA PRESERVATION TASK FORCE

Though the importance of preservation is referenced in the EOSC Strategic Research and Innovation Agenda (SRIA) [21] there is not yet an explicit strategy. The vision of the EOSC Association LTDP Task Force [22] will address the service infrastructure, financial implications, and stakeholder roles and responsibilities necessary to provide sustainable policies, practices and strategic execution.

B. NETWORK OF TRUSTWORTHY DIGITAL REPOSITORIES

The proposed network of FAIR-Enabling TDRs, (initially with a European scope) is currently at the exploratory stage. Such a network would benefit multiple stakeholders: the repositories themselves, the researchers, and the EOSC.

It can provide an active and unified voice for FAIR-enabling Trustworthy Digital Repositories in Europe who are key stakeholders of EOSC. The adoption of FAIR and TRUST [23] practices could be promoted through training and support programmes. This would increase repository compliance with the rules of participation of EOSC and it would facilitate the implementation of widely agreed and common TRUST and FAIR assessment frameworks. The

network could ultimately facilitate researchers' access to long term preservation by enhancing the connections between TDRs and data services and technology providers in EOSC.

V. OUTCOMES

Presentations from the three authors will provide the basis for discussion of the key topics raised. The input from expert audiences such as those represented at iPres are critical to guiding the scope and focus of efforts across the standards, task forces and community developments represented.

REFERENCES

- [1] CoreTrustSeal <https://www.coretrustseal.org/>
- [2] CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022 <https://doi.org/10.5281/zenodo.3638211>
- [3] L'Hours, H., Kleemola, M., & de Leeuw, L. (2019). CoreTrustSeal: From academic collaboration to sustainable services. *IASSIST Quarterly*, 43(1), 1–17. <https://doi.org/10.29173/iq936>
- [4] Research Data Alliance <https://www.rd-alliance.org/>
- [5] World Data System Regular Members <https://www.worlddatasystem.org/services/certification/>
- [6] Data Seal of Approval (DSA) Synopsis (2008–2018) <https://www.coretrustseal.org/about/history/data-seal-of-approval-synopsis-2008-2018/>
- [7] RDA/WDS Certification of Digital Repositories IG <https://www.rd-alliance.org/groups/rdawds-certification-digital-repositories-ig.html>
- [8] European Commission, Directorate-General for Research and Innovation, Turning FAIR into reality : final report and action plan from the European Commission expert group on FAIR data, Publications Office, 2018, <https://data.europa.eu/doi/10.2777/54599>
- [9] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
- [10] Consultative Committee for Space Data Systems (2011) Audit and Certification of Trustworthy Digital Repositories (Magenta Book, CCSDS 652.0-M-1). <https://public.ccsds.org/pubs/652x0m1.pdf>
- [11] nestor (2022) nestorSeal for Trustworthy Digital Archives, https://www.langzeitarchivierung.de/Webs/nestor/EN/Zertifizierung/nestor_Siegel/siegel.html
- [12] European Open Science Cloud (EOSC) <https://eosc.eu/>
- [13] FAIRsFAIR <https://www.fairsfair.eu/> SSHOC: <https://sshopencloud.eu/> EOSC-Nordic: <https://www.eosc-nordic.eu/>
- [14] L'Hours, Hervé, Kleemola, Mari, von Stein, Ilona, van Horik, René, Herterich, Patricia, Davidson, Joy, Rouchon, Olivier, Mokrane, Mustapha, & Huber, Robert. (2021). FAIR + Time: Preservation for a Designated Community (01.00). Zenodo. <https://doi.org/10.5281/zenodo.4783116>
- [15] Ramezani, Sara, Aalto, Tero, Gruenpeter, Morane, Herterich, Patricia, Hooft, Rob, & Koers, Hylke. (2021). D2.7 Framework for assessing FAIR Services (V1.0_DRAFT). Zenodo. <https://doi.org/10.5281/zenodo.5336234>
- [16] European Commission, Directorate-General for Research and Innovation, Jones, S., Aronsen, J., Beyan, O., et al., Recommendations on certifying services required to enable FAIR within EOSC, Genova, F.(editor), Publications Office, 2021, <https://data.europa.eu/doi/10.2777/127253>
- [17] Hervé L'Hours, Maaïke Verburg, Jerry de Vries, Linas Cepinskas, Ilona von Stein, Robert Huber, Joy Davidson, Patricia Herterich, & Benjamin Mathers. (2022). D4.6 Report on a maturity model towards FAIR data in FAIR repositories (1.0 DRAFT). Zenodo. <https://doi.org/10.5281/zenodo.6090389>
- [18] Currie, Amy, & Kilbride, William. (2021). FAIR Forever? Long Term Data Preservation Roles and Responsibilities, Final Report (Version 7). Zenodo. <https://doi.org/10.5281/zenodo.4574234>
- [19] EOSC Association Long Term Data Preservation Task Force <https://www.eosc.eu/advisory-groups/long-term-data-preservation>
- [20] Ilona von Stein, Hervé L'Hours, Linas Cepinskas, Benjamin Mathers, Ingrid Dillo, Maaïke Verburg, Mustapha Mokrane, Patricia Herterich, & Olivier Rouchon. (2021). D4.4 Coordination Plan for a sustainable network of FAIR-enabling Trustworthy Digital Repositories (1.0_DRAFT). Zenodo. <https://doi.org/10.5281/zenodo.5726691>
- [21] EOSC Strategic Research and Innovation Agenda (SRIA) <https://www.eosc.eu/sria>
- [22] Reporting to the EOSC Association advisory Group on Sustaining EOSC Association advisory group <https://www.eosc.eu/sustaining-eosc>
- [23] Lin, D., Crabtree, J., Dillo, I. et al. The TRUST Principles for digital repositories. *Sci Data* 7, 144 (2020). <https://doi.org/10.1038/s41597-020-0486-7>

DIGITAL STORYTELLING AS PRESERVATION

A Screening and Panel Discussion

Syreeta Gates

Gates Preserve
USA
syreeta.gates@gmail.com

Jamie A. Lee

University of Arizona
USA
Jalee2@arizona.edu
[0000-0001-6182-2372](tel:0000-0001-6182-2372)

James Lowry

City University of New York
USA
james.lowry@qc.cuny.edu
[0000-0001-9970-3846](tel:0000-0001-9970-3846)

Abstract – Mimi Onuoha's *Library of Missing Datasets* catalogues datasets that were never created and captured, illuminating how cultural values are represented in absences of care, of data and of memory. The “archival silence” trope is well recognized today, yet in the present some lives, experiences and events continue to be ghosted by record-making and institutional collecting and preservation. Digital storytelling is one set of techniques that are being used to transmit memories into the future. Digital storytelling encompasses a range of formats and methods, from geospatial and timeline media, to video essays and podcasts, to simple blogs. In the creation of these digital objects, which themselves require preservation, digital storytellers can preserve traces of analogue or lived culture and experience, so that preservation is enacted by and enacted upon digital stories. Exploring what lost media and memory mean for communities today and in the past, this panel presentation will present examples of digital storytelling work that is both a means for and an object of preservation. By positing digital storytelling as a preservation method, the panel will consider issues related to working in community, accessing tools and know-how, finding a home for digital objects, and sharing (and deleting) these products of memory work.

Keywords – Digital storytelling, analog preservation, memory work, community archives.

Conference Topics – Community

REFERENCES

- [1] M. Onuoha, “Library of Missing Datasets”
<https://github.com/MimiOnuoha/missing-datasets>

Is There a Right Way to Archive Email?

Ruby Martinez

University of Illinois at
Urbana-Champaign
USA
ruby1m2@illinois.edu

Christopher Prom

University of Illinois at
Urbana-Champaign
USA
prom@illinois.edu

Christopher Lee

University of North
Carolina Chapel Hill
USA
calle@ils.unc.edu

Abstract – As email and the systems used to preserve it continue to evolve, email archiving continues to do so as well. We are approaching a time of community maturity, with multiple institutions taking roughly similar approaches to preserve their email archives, but with different technologies at different scales and for different purposes. This panel will discuss the evolving landscape of email archiving and how projects on opposite ends of the spectrum are addressing the strengths and weaknesses of their approach. It will also provide attendees an opportunity to assess the projects' impacts on the email preservation community.

Keywords – Email Archiving, Sustainability, Community
Conference Topics – Community, Exchange

I. INTRODUCTION

When considering the long-term preservation needs of disparate user communities, email archiving (EA) workflows have not converged into a one-size-fits-all solution. Instead, a spectrum exists where 'practical' solutions can be found on one end, with 'exemplary' projects on the other. The former have fewer barriers to entry, and the latter require higher-level technical support.

Unfortunately, many organizations and institutions don't find themselves anywhere on the spectrum. Despite the refinement of current email archiving workflows or new ones emerging, a large number of institutions either don't archive email or have acquired it without an active plan for how they will preserve it, as noted in a recent survey of email archiving practice in the US State of Illinois [1].

The last two years have seen increased interest by researchers in digital preservation and efforts to address email preservation issues. The *Email Archives: Building Capacity and Community (EA:BCC)*

regrant program represents a growing network of numerous professionals in the archival, library, digital preservation, and museum fields that develop and address critically needed solutions to preserving email.

The expansion of the email archiving network allows for the continued development of resources and solutions. However, the challenges facing the community can be daunting, especially for institutions with limited resources and opportunities to discuss these challenges with others.

Panelists in this debate-style session will discuss the evolving landscape of email archiving and engage the audience in arguing the pros and cons of projects on opposite ends of the spectrum, e.g. those addressing 'practical' needs vs. those more 'exemplary' in nature. As such, attendees will participate in a discussion that addresses the strengths and weaknesses of alternative preservation approaches, while assessing their impact on the email preservation community and their own programs.

II. EXISTING WORK

The work being completed under the EA:BCC regrant program provides a backdrop to this session [2]. EA:BCC is funding work that expands a much-needed network of professionals from many fields that are addressing the need for email preservation. This includes the following projects:

- University of Maryland, Discovery environments for using email archives: Evaluating user needs with prototype version of Email CONTEXTualisation DIScovery Tool (EConDist)

- University of North Carolina Chapel-Hill, RATOM Functional, Interoperability and Reuse Extensions (RATOM-FIRE)
- 92nd Street Y, Love the Words: Preserving the Email Collection of 92Y's Unterberg Poetry Center
- Harvard University, Integrating Preservation Functionality into ePADD
- Council of State Archivists, Inc., CoSA PREPARE: Preparing Archives for Records in Email
- University of Chicago Library, Attachment Converter: Preserving the Context of Electronic Correspondence
- University of Albany, SUNY, Mailbag: A Stable Package for Email with Multiple Formats
- Columbia University, Creating Email Archives from PDFs: The Covid-19 Corpus

The EA:BCC work does not present any final solutions but rather sets out a vision for preserving email by encouraging collaboration and interoperability. Each project seeks to address the needs of a designated community. Some projects, which we might characterize as 'practical,' emphasize approaches that meld email archiving into existing repository tools and services. Others, which we might call 'exemplary' offer a high degree of customizability in workflow design, target formats, metadata interoperability, preservation systems, and dissemination options.

III. PRACTICAL SUSTAINABILITY

"Email Archiving in PDF (EA-PDF): From Initial Specification to Community of Practice," is an example of the former, 'practical,' approach. With funding from the Institute for Museum and Library Services, EA-PDF intends to open up email archiving possibilities to institutions that are currently not archiving email. To do so, the project is developing an EA-PDF file format, a prerequisite for low-barrier methods to produce authentic, renderable, and usable email packages in PDF form. Since PDF is a widely implemented file format, the ability to produce EA-PDF files would provide individuals and institutions a pathway to migrate email into the most widely used format for the distribution of text documents [3].

The use of PDF puts a layer of abstraction between encoding and rendering, potentially allowing the implementation of simple PDF-creation workflows (and dedicated viewers) without requiring

users to have an intimate understanding of underlying technical mechanisms. This approach would maximize scalability and sustainability for those institutions that use the format. But file format development also requires a significant upfront investment: the development of the specification, which is itself a prerequisite to enable the development of EA-PDF creation tools.

IV. EXEMPLARY SUSTAINABILITY

"RATOM Functional, Interoperability and Reuse Extensions (RATOM-FIRE)" is an example of the latter, 'exemplary' approach. RATOM-FIRE will address several essential email curation use cases through enhancements to software development through the Review, Appraisal and Triage of Mail (RATOM) project. RATOM has produced and bundled a suite of powerful tools to reliably and efficiently process email collections. The output of the software is designed to facilitate a wide range of curation activities, including review for sensitivity, appraisal and response to open records requests.

The RATOM-FIRE project will address identified archival community needs to integrate the tool output into existing and emerging digital curation workflows. This will include easier export of email messages as individual (EML) files, capturing more detailed preservation metadata, and expanding the public application programming interface (API) of the RATOM software library to facilitate easier integration into other tools (in addition to invoking RATOM tools through a command-line interface).

While RATOM-FIRE depends on a complex set of preservation tools and practices, the goal is to make it easier to collect, preserve and make sense of email archives at a wide range of institutions.

V. CONCLUSION

Because email represents such a large portion of digital information created and stored today, the need to develop and implement effective, sustainable strategies for its preservation is critical. While there has been a growing collection of specialized tools for archiving email, there are still many open questions the community is trying to answer: Will existing tools suffice if deployed with more care and expertise? Will email archives require even more powerful tools with capabilities beyond what is found in traditional archives? What is the best

balance between community-supported and commercial solutions?

Chris Prom and Christopher Lee will debate the pros and cons of respective approaches to email archiving in a discussion moderated by Ruby Martinez. This panel will also engage the audience on many more questions, challenges, and developments in email archiving.

REFERENCES

- [1] Ruby L. Martinez, "Email Archiving: Building Capacity and Community," iPres 2021, October 22, 2021, <https://www.scimeeting.cn/m/video/play/2?vid=173404>
- [2] Email Archives: Building Capacity and Community, News and Updates, <https://emailarchivesgrant.library.illinois.edu/blog/>.
- [3] EA-PDF Working Group. "A Specification for Using PDF to Package and Represent Email." January 2021. <https://www.ideals.illinois.edu/handle/2142/109251>.

LESSONS LEARNED DURING THE IMPLEMENTATION OF A DIGITAL PRESERVATION PROJECT

Experiences from Europe, USA and Asia

Jessica Knight

United States Holocaust
Memorial Museum
USA
jknight@ushmm.org

Nathan Tallman

The Penn State University
USA
ntt7@psu.edu
[0000-0002-5308-4100](tel:0000-0002-5308-4100)

Mark Hobbs

The Royal Horticultural
Society
UK
markhobbs@rhs.org.uk

Mui Huay Ho

Temasek Polytechnic
Singapore
HO_Mui_Huay@TP.EDU.SG

Driek Heesakkers

University of Amsterdam
Netherlands
h.j.heesakkers@uva.nl

Teo Redondo

LIBNOVA
Spain
teo.redondo@libnova.com
[0000-0001-6465-7771](tel:0000-0001-6465-7771)

Abstract – Before starting a digital preservation project (an OAIS-aligned long-term preservation of a digital repository), many things are taken for granted that are discovered during the implementation of the project.

Each implementation of a digital preservation project has its own peculiarities and characteristics, but also many similarities. Generally, preparation processes take much longer than expected, multiple teams within the organization need to be coordinated, and many project details need to be very well planned.

In this panel, representatives of institutions from different GLAM sectors from different countries (and even continents), will speak from their own experience about the lessons learned during the implementation of their digital preservation project.

Keywords – Digital Preservation, Implementation, Lessons learned, Digital Repositories.

Conference Topics – Community; Exchange.

I. INTRODUCTION

Tackling a long-term digital preservation project is not an easy task for any organization. Projects can vary in scope, from the complexity of implementing a digital preservation repository or as “simple” as a specific format-migration; regardless of their scope, all projects benefit from proper planning and resource allocations. There are some manuals such as the DPC Digital Preservation Handbook [1] that provides an internationally authoritative and practical guide to the subject of managing digital

resources over time and the issues of maintaining access to them, which are helpful to practitioners when approaching a digital preservation project. However, facing a real project involves much more planning and organization than anticipated. The work begins long before the selection of the system to be used, regardless of what particular DPS is to be chosen, identifying all the parties involved, the collections to be worked on, the current volume of collections and their scalability for the future, establishing the level of preservation (NDSA LoP [2]) to aim for, and many other things that are only identified once a real digital preservation project is implemented.

In this panel, people from the following institutions *U.S. Holocaust Memorial Museum*, and *Penn State University*, from USA; *The Royal Horticultural Society* and *The University of Amsterdam* from Europe; and *Temasek Polytechnic* from Asia, will exchange their experiences implementing a preservation project in different kinds of libraries and archives (a museum, a university, a herbarium), by sharing with the digital preservation community the lessons learned during the preparation and execution of the project, what works and what does not, and some useful insights for anyone in the same situation.

LIBNOVA is the common denominator among the different organizations, and its role will be to

serve only as a moderator of the panel session. Presentations will not dwell on LIBNOVA-specific details but instead focus on project management and lessons learned.

II. PANEL DISCUSSION TOPICS

The panel will discuss the following topics and questions:

- Key considerations when planning a digital preservation project, including overarching long-term aspects, such as scalability, funding, sustainability, etc.
- Who should be part of the preservation project and what should their role be? Identify the preservation team and the coordination between the different areas involved.
- The methodology used to select what to preserve and what not.
- Peculiarities of each type of content to be considered.
- General insights and future plans envisioned based on lessons learned.

III. PANELLIST PROFILES

Jessica Knight is Senior Advisor for Digital Ecosystem, Preservation and Discovery, at United States Holocaust Memorial Museum (USHMM). She manages the Museum's digital preservation project with a team of IT, INFOSEC, and digital access and preservation specialists. The Museum began its deployment of LIBSAFE in 2017 to house over 90 million digital files, close to a PB, of Holocaust-related records including born-digital and digitized oral testimony, film, documents, photographs, publications, and historical sound.

Nathan Tallman is Penn State University Libraries first digital preservation librarian, seeking to establish a robust digital preservation program, current efforts have been underway for about five years; LIBNOVA Advanced is Penn State's first digital preservation repository. Penn State currently manages about 250 TB of data that includes born-analog and born-digital personal papers and organizational records, general collections (monographs, serials, audio/visual), research data, software, and library publications. Penn State is configuring LIBSAFE Advanced in a flexible

implementation based on Levels of Digital Preservation Commitment rather than formats.

As project manager at the Library of the University of Amsterdam, **Driek Heesakkers** led the Digital Depot project. This project ran the European tender for a preservation grade depot for digital collections, suitable for both library, archive and museum collections, and subsequently the implementation of LIBSAFE from LIBNOVA. The digital depot was taken into production in May 2022, with links to the existing ArchivesSpace and image bank front-ends and Goobi digitization workflow system. In the next two years, the rare collections department of the library, known as 'Allard Pierson', will ingest around 250TB of digital objects currently stored in various media from previous digitization projects and a small but growing number of digital-born personal and institutional archives.

Mui Huay participated in two digital preservation project implementations and lived to tell the tale. As a cross-functional project team member, she was knee-deep from the get-go: from specifying system requirements and executing a preservation plan for each type of content to promoting the value of the project to stakeholders. Mui Huay heads the Archives and Reference division of the Temasek Polytechnic Library, and has been an active member of the implementation of the LIBSAFE platform and the migration from older technologies.

Mark Hobbs is Library Digital Collections Manager at the Royal Horticultural Society's Lindley Library in London, UK. The Lindley Library contains one of the world's most important collections of books on horticultural history, botanical artworks and the historic archives of the RHS and key figures in British horticulture. Mark coordinates the Library's ongoing digitisation programme, and plays a key role in ensuring the preservation of nearly 20 years of digitisation at the Lindley Library, through the implementation of LIBNOVA's LIBSAFE platform. These digitisation and preservation projects form part of a wider project to share the RHS's Library and Herbarium collections on LIBNOVA's open access platform in 2023.

Teo Redondo is the CTO and Head of Research & Development at LIBNOVA, where he leads several innovation projects about Digital Preservation solutions for Libraries, Archives and Museums, and

Research institutions, and also leads LIBNOVA Research Labs for the areas of future functionalities, most around implementing Artificial Intelligence techniques for better handling of research data and content. In this panel, Teo will act only as facilitator of the session.

REFERENCES

- [1] Digital Preservation Handbook, 2nd Edition, <https://www.dpconline.org/handbook>, Digital Preservation Coalition © 2015.
- [2] Levels of Preservation Revisions Working Group, "Levels of Digital Preservation Matrix V2.0," October 2019, <https://osf.io/2mkwx/>.

IT'S ALL IMPORTANT OF COURSE, BUT...

*A bloodless ~~fight~~ discussion about what is the **most** important aspect of digital preservation*

Paul Stokes

Jisc
UK

Paul.stokes@jisc.ac.uk

[0000-0002-7333-4998](tel:0000-0002-7333-4998)

Tamsin Burland

Jisc
UK

Tamsin.burland@jisc.ac.uk

[0000-0002-5129-979X](tel:0000-0002-5129-979X)

Abstract – a panel to shine a light on how different aspects of digital preservation are important to different practitioners and to provide an insight as to why those aspects are important to them.

Keywords – Sustainability, Cost, Value, Risk, Data loss.

Conference Topics – Resilience, Exchange.

I. INTRODUCTION

We all agree that digital preservation is "a good thing". We wouldn't be attending iPRES if we didn't. It is also probably a given that, should a room of digital preservationists be presented with a list of policies / features / characteristics / components / problems relating to digital curation, the consensus would be that they're ALL important...

And then someone would say "but"...

The truth is we (and by "we" I mean all those involved in digital preservation in one way or another) all have strong opinions about how those policies / features / characteristics / components / problems could and indeed should be ranked. Opinions shaped by our background and experiences.

"Preservation file formats are the most important". "Policy trumps everything". "Costs and values are obviously at the top of the list". "Without preservation systems you can't do anything"

We are an inclusive community and pride ourselves on listening to and learning from others. But we rarely have the opportunity to discuss these features / characteristics / components / problems in an open forum where we get the

opportunity to "get inside the heads" of other practitioners and understand why they feel (so strongly) the way they do.

II. THE PANEL

The proposed panel will bring together a small number of respected practitioners who are known to have strong (and potentially contrasting) opinions on core aspects of digital preservation policies / features / characteristics / components, along with some practitioners from historically overlooked or under represented communities, to present short provocations about what aspect of digital preservation is most important to them, and why. After the provocations there will be questions from the floor and an open discussion.

The panel will provide an opportunity to understand what drives colleagues. What their interests are and what they see as the primary problems in the field of digital preservation.

The panel has been designed to represent as diverse a range of viewpoints and communities as possible. At the time of writing, the following people attending iPRES 2022 (either in-person or virtually) have agreed to take part:

Name	Role
Elizabeth Thurlow	Digital Preservation and Access Manager at University of the Arts London
Alina Karlos	Assistant Archivist ILRC University of Namibia

Niklas Zimmer	Manager: Digital Library Services University of Cape Town
George McGregor	Institutional Repository Manager IS Scholarly Research Communications University of Strathclyde
Caylin smith	Head of Digital Preservation at Cambridge University Library
Kirsty Lingstadt	Director of Library, Archives and Learning Services Student and Academic Services University of York
Donna McRostie	Deputy Director, Research and Collection Stewardship University of Melbourne

The lead author, Paul Stokes—Product Manager (Preservation) at Jisc— will chair the session and also contribute.

Emerging themes for the most important aspect of digital preservation raised by the panel to date include:

- It's all about the money
- Advocacy trumps everything
- Designated communities are unimportant
- Designated communities ARE important
- Significant properties are unimportant

This session won't provide the definitive answer to the question relating to what is THE most important aspect of Digital Preservation, but it will engender wider understanding of digital preservation issues.

COMPUTATIONAL ACCESS TO DIGITAL MATERIAL

Exploring topics around engagement, ethics and resources

Leontien Talboom

Cambridge University Libraries
United Kingdom
lkt39@cam.ac.uk
[0000-0001-7408-5471](tel:0000-0001-7408-5471)

Jenny Mitcham

Digital Preservation Coalition
United Kingdom
jenny.mitcham@dpconline.org
[0000-0003-2884-542X](tel:0000-0003-2884-542X)

James Baker

University of Southampton
United Kingdom
j.w.baker@soton.ac.uk
[0000-0002-2682-6922](tel:0000-0002-2682-6922)

Sonia Ranade

The National Archives
United Kingdom
sonia.ranade@nationalarchives.gov.uk
[0000-0002-2674-8370](tel:0000-0002-2674-8370)

Abstract – Computational access is an innovative approach to access within the digital preservation community. This type of access is becoming more widely discussed, but a lot of uncertainties are present around the term. A guide has been created as part of a collaborative piece of work between the Digital Preservation Coalition (DPC) and Leontien Talboom, Software Sustainability Institute (SSI) fellow 2021 to help digital preservation practitioners to understand the topic and take some practical steps towards implementation. The creation and launch of this guide has led to much wider discussions on this topic, many of which are beyond the introductory scope of the guide. This panel session will therefore provide an opportunity and platform to discuss these issues in greater depth with a range of practitioners with practical experience in this area.

Keywords – computational access, collections as data, access, computational methods
Conference Topics – Innovation

I. INTRODUCTION

Within the digital preservation community, the term computational access is used with increasing frequency. Many practitioners are aware that this type of access may be beneficial - both for their own use of the digital material and that of their users. What is involved in establishing computational access to digital materials is sometimes harder to establish, and practitioners can be left unsure about which steps they should take to begin to explore

these technologies. In some cases there is also a lack of understanding of what the term 'computational access' actually means and how it interfaces with other related concepts such as artificial intelligence, machine learning and data mining.

Computational access, and the closely related term 'collections as data'[1], offers a new way of providing access to material to be used for computational methods. To explore this and related terms in more detail, and to provide the community with a way to get started, a beginner's guide has been created by Leontien Talboom (a Software Sustainability Institute (SSI) fellow) in collaboration with the Digital Preservation Coalition (DPC) and a broad network of experts drawn from the community[2]. This guide aims to demystify computational access and make it a more approachable topic for digital preservation practitioners.

The guide covers several themes. It starts by introducing and defining computational access and related terms, such as artificial intelligence and data mining. It discusses and describes the four main approaches to implementing computational access, discusses the benefits and drawbacks of using computational methods to provide access to digital materials and provides some useful first steps for practitioners who want to get started. A selection of

helpful case studies are also shared to demonstrate what is possible in a range of different organizations.

While in the process of creating this short guide, it was apparent that a huge amount of knowledge and experience has already built up within the community. Recognising that a beginner's guide can only go so far, this panel was brought together to enable continued discussion on the topic of computational access and to delve further into some of the details that were not included in the guide.

II. INTRODUCTION TO THE PANEL

The panel session began with a short introductory presentation providing background to the topic, to the beginner's guide and to the panelists. Computational access was described as any type of access in a digital environment that requires computational methods. Four approaches to computational access were introduced - terms of use, bulk dataset, API and platform - and it was noted that an organization may choose to employ more than one of these techniques. Further description of these approaches is detailed in the beginner's guide to computational access. It was noted that access is a key component of successful digital preservation and one that many institutions find challenging to implement for digital content at scale.

III. INVITED PANELISTS

The following panelists were invited to participate in this session ensuring a range of expertise and backgrounds were represented. Below a short description can be found outlining each panelist and their work. The session was facilitated by Jenny Mitcham, Head of Good Practice and Standards at the Digital Preservation Coalition.

James Baker is Director of Digital Humanities at the University of Southampton. He works at the intersection of history, cultural heritage, and digital technologies, and is currently researching histories of knowledge organization in twentieth century Britain. James is a Software Sustainability Institute Fellow, a Fellow of the Royal Historical Society, a member of the Arts and Humanities Research Council Peer Review College, a convenor of the Institute of Historical Research Digital History seminar, and a Trustee of the Programming Historian. Prior to joining Southampton, James held positions of Senior Lecturer in Digital History and Archives at the University of Sussex and Director of

the Sussex Humanities Lab, Digital Curator at the British Library, and Postdoctoral Fellow with the Paul Mellon Centre for Studies in British Art.

Dr. Sonia Ranade is Head of Digital Archiving at The National Archives (UK), with responsibility for digital services to records creators in government (for records selection and transfer), preservation of the digital Public Record and access to born digital records. Sonia's research interests include probabilistic approaches to archival description, digital preservation risk and developing new access routes for digital archives (both for 'readers' and for computational re-use). Sonia holds a PhD in Information Science.

Leontien Talboom is a collaborative PhD student at The National Archives (UK) and University College London. Her research focuses on the constraints faced by digital preservation practitioners when making born-digital material accessible. She currently also works as a web archivist on the Archives of Tomorrow project and a technical analyst at Cambridge University Libraries.

IV. TOPICS DISCUSSED

A discussion was facilitated on a range of topics related to computational access and these themes were further drawn out with questions from the audience. Discussion centered on the following topics and questions:

Resources and infrastructure - Computational access is a novel way of opening up collections and could offer many interesting future uses of collections, but how should organizations with limited resources and experience manage this? What is a potential way to get started?

Panelists touched on the importance of starting small and not trying to do much at once. When planning a computational access approach it is easy for the scope to creep and the ultimate plans to be unworkable. It is far better to start doing something and then build on this. The importance of sharing with the wider community and learning from each other was also highlighted. It was mentioned that the computational access guide was designed to help organizations get started with computational access, particularly the practical steps section. The 'Terms of Use' approach as detailed in the guide was flagged

up as the simplest way of enabling computational access if you have fewer resources available to you.

Communicating with stakeholders -

Establishing computational access to digital materials involves collaboration with a range of stakeholders from different disciplines and with different levels of knowledge and interest. There is a balance that needs to be found when communicating with different stakeholders (for example senior managers and colleagues in IT) and challenges that arise when communicating with these groups. How do we effectively build relationships with key stakeholders and how do we sustain computational access services and keep them running (particularly when they are often funded as short term projects)?

Discussion on this topic brought out some of the challenges of working in this area, but also some solutions. Challenges were noted around understanding some of the terms that practitioners working in computational access use and the need to ensure that all stakeholders understand what is being discussed was raised. There are also challenges in enabling stakeholders to visualize what plans for computational access might be. It may be hard for some people to get excited about an API, so it was considered to be more impactful to show your audience a more tangible example so they can really understand the benefits of the work you are proposing.

There was also some discussion about how computational access projects can be sustained, and indeed whether they always need to be sustained. When should a computational access project become a service or equally, when should a decision be made that it is no longer required? Whilst it was recognised by panelists that for innovative new approaches to access, funding is often more likely to be gained for exploratory projects rather than for longer term services, it is important to make the case for services rather than projects where necessary.

Engaging an audience - There is little benefit in facilitating computational access to digital materials if no-one uses these services. The panel discussed how to engage an audience and ensure the tools, services and resources they are providing meet users' needs. Panelists were asked to discuss to what extent their work is informed by the needs of users and how easy it is for users to influence the access options that are available?

The panel discussed the importance of building services around user needs. At The National Archives (UK), user research is a core part of establishing services, but it can be hard to find people to talk to about more specialist types of access such as computational methods. There can be a risk of a 'build it and they will come' mentality around computational access but The National Archives would prefer that their services are more firmly grounded in what their users need.

There is also a concern around providing too much for users, especially when doing computational methods for them. Some institutions have had mixed feedback on prepared datasets and computational access platforms that hide the complexity and choices that were made to provide the material in that way. There needs to be a balance between providing computational access and tools that ensure the practitioners do not end up doing the research for the users, but also in lowering the barrier for users who may not have the computational skills or confidence.

The importance of sustainability of services was also discussed. Services are unlikely to be sustained if they are not meeting users' needs. Without proper audience engagement at an early stage, it may be the case that services will not stand the test of time.

The point was also made that we shouldn't always try and do everything that our users want. Some users may have unrealistic expectations (for example suggesting we should keep everything or digitize everything), so whilst it is important to engage with users as much as we can, we have to strike a balance around how much we should try to facilitate. Users of computational access services are quite diverse in their needs and requirements and it is unlikely we can meet all of these needs. Whilst engagement with users is a key issue, thought needs to go into where to draw the line.

Ethics around using computational tools -

This is a key topic and one that can be approached from a number of different angles and perspectives - both from the user's and the digital preservation practitioner's point of view. Many of the processes associated with computational tools used to facilitate access are 'black boxes', therefore raising ethical concerns around the underlying processes of these tools. Questions discussed were around how considerations of ethics have impacted our work in

this area and to what extent they have been a barrier in embracing computational techniques.

Whilst the beginner's guide to computational access includes a short discussion on the topic of ethics, this panel session delved a bit further into some of the key issues. Firstly, the challenge of providing a useful set of results when searching digital content at scale was discussed. A keyword search may produce many thousands of results, but how do we strike a balance between providing a useful service for researchers and introducing bias. A researcher is unlikely to be able to look through tens of thousands of records, so ordering the records in some way might be helpful. However, we need to consider whether relevance ranking is helpful to our users or would introduce too much bias and lead their research in a particular direction. This is where transparency in our own approach may be helpful. At least if we document the bias of the approaches we are using, and make this information available to our users, they can make better informed decisions about their results.

Another similar example mentioned was the process of digitization. Why did we digitize this content but not that content? What did we leave out and why? Documenting this would enable users to understand the dataset in a more meaningful way. Working at scale highlights the need for documentation and we should be better as a community of documenting our decisions and making those decisions transparent to users.

Reference was made to an interesting case study relating to digitized colonial archives in Denmark, and the specific example of how a historic photograph of a crying child could be taken out of its original context[3]. This decontextualization of digital content highlights some of the ethical challenges that we face as a community when opening up wider access to our collections. The need to be able to retain an association between the digital content, its metadata and perhaps even the search terms used to originally find it was also discussed.

It was acknowledged that whilst there are many opportunities around computational access there is also an element of risk involved. Whilst allowing researchers to ask very different types of questions, computing over collections at scale increases the chances of revealing information that shouldn't be in the public domain. The National Archives (UK) are currently exploring how they might mitigate some of

these risks by re-introducing some of the natural friction that occurs when researchers order physical documents in the reading room. There is a course to be steered between the need to open up digital material for access and managing the resulting risks. Making all content openly available is unlikely to be possible but archives should also be wary of being too risk averse and closing access down without good reason.

V. CONCLUSION

This panel session provided a valuable opportunity to explore topics around resourcing, engagement and ethics as they relate to computational access. Some of the key messages to come out of the discussions were around the importance of getting started, the need to find ways to effectively sell the value of computational access to key stakeholders, the importance of engaging with users and the balance that must be found in meeting their needs, and the need to be open and transparent in our work whilst finding ways to manage the risk in providing access at scale.

REFERENCES

- [1] T. Padilla, L. Allen, H. Frost, S. Potvin, E. Russey Roke, and S. Varner, "Always Already Computational: Collections as Data," Final Report, 2018.
- [2] L. Talboom et al, *Computational Access: A beginner's guide for digital preservation practitioners*. Digital Preservation Coalition, 2022. <https://www.dpconline.org/digipres/implement-digipres/computational-access-guide>
- [3] T. Odumosu, "The Crying Child: On Colonial Archives, Digitization, and Ethics of Care in the Cultural Commons." *Current Anthropology* Vol. 61, no. S22, 2020. <https://doi.org/10.1086/710062>

WILL DNA FORM THE FABRIC OF OUR DIGITAL PRESERVATION STORAGE?

DNA Data Storage: A Panel Discussion

Daniel Chadash

Twist Bioscience
USA
dchadash@twistbioscience.com
[0000-0002-9712-6034](tel:0000-0002-9712-6034)

Paul Wheatley

Digital Preservation
Coalition
UK
paul@dpconline.org
[0000-0002-3839-3298](tel:0000-0002-3839-3298)

Sibyl Schaefer

University of California
USA
sschaefer@ucsd.edu
[0000-0002-7292-9287](tel:0000-0002-7292-9287)

Euan Cochrane

Yale University Library
United States
euan.cochrane@yale.edu
[0000-0001-9772-9743](tel:0000-0001-9772-9743)

DNA data storage is on the cusp of becoming an economical technology which could be adopted by the digital preservation community. Many long-lived media technologies have fallen by the wayside after failing to meet the needs of digital preservationists. Does DNA Storage have the potential to break this trend and offer new capabilities in long-term archiving? This panel will consider whether DNA data storage can meet the requirements of digital preservationists. It will discuss where DNA might fit into a storage strategy, and debate what needs to happen to transform the potential into a viable product for this community.

**Keywords – DNA Data storage, Storage mediums, Digital Preservation Requirements
Conference Topics – Innovation, Exchange**

I. LONG LIVED MEDIA - SOLUTION OR RED HERRING?

Over the course of the past two decades a recurring theme has been the emergence of new long-lived storage media technologies that have been proclaimed by their creators as the saviour of organizations undertaking long-term archiving challenges. A common characteristic of the announcements associated with these technologies has been a clear misunderstanding of the requirements of those organizations seeking to keep their data for long periods.

David Rosenthal referenced what in 2013 was the latest in a long line of these developments (in this case 5-dimensional quartz DVDs) and noted that "So far, announcements of very long-lived media have made no practical difference to large-scale digital preservation..." referencing primarily economics as the reason behind this failure. He also noted that DNA as storage technology showed a more promising future [1] as did Paul Bertone at iPRES in the same year [2].

II. DNA DATA STORAGE: ON THE ROAD TO REALITY

Storing data on DNA is not a new concept, having first been demonstrated several decades ago. Since then it has mainly been deployed in proof of concept scenarios due to the high cost of DNA synthesis (writing the data).

Synthesizing DNA on silicone is the breakthrough that could lower the cost to a competitive price compared to the TCO of current archival storage. Progress made in the past decades in semiconductor fabrication are driving the confidence that those advancements can be applied to DNA synthesis and bring it to a production level at an attractive price [3].

III. WHAT DOES THE DIGITAL PRESERVATION COMMUNITY NEED?

Digital preservation has a challenging set of requirements that need to be met by a storage solution it utilizes, for example as articulated in the NDSA levels, DPC RAM and the digital preservation storage criteria [4]. Referencing some of these as questions of a prospective storage solution illustrates points of discussion that will be considered by the panel:

Is the solution a robust and safe home for data and in what ways does it mitigate and counter the risks of data loss? What confidence is there that data really is safe? Is the way data is stored transparent and how can this be independently validated? Is the solution widely adopted by a community of users and suppliers, and if not then is it headed this way? What confidence is there that the solution will still be around in at least the medium term? Are there other uses and applications of the solution outside of digital preservation that will help ensure long-term sustainability and drive economies of scale? Are there specific use cases where the solution excels or fills a gap where there are few or no alternatives?

IV. THE PANEL

This panel session seeks to explore the current state of the art of DNA data storage and the requirements of the iPRES community for long-term preservation. It will seek to consider how DNA storage could work in practice in a long-term digital preservation context, what use cases it could fulfill and consider the key areas where both the technologists and the preservationists need to be on the same page to enable fruitful application of DNA storage in this field.

As such the panel will be composed of experts from a number of different sectors, brought together via a collaboration between the DNA Data Storage Alliance and the Digital Preservation Coalition, as well as the wider digital preservation community. The panel members are: Matthew Addis (Arkivum), Daniel Chadash (DNA Storage Alliance), Euan Cochrane (Yale University Library), Dave Landsman, (Western Digital), Sibyl Schaefer (University of California), Paul Wheatley (Digital Preservation Coalition) and Jenny Xiao, (Illumina)

V. FORMAT OF THE PANEL

The panel will focus on a question and discussion format, moderated by Paul Wheatley. As well as seeking to engage with the conference audience, the

panel will be posed the following questions with the aim of stimulating discussion on not just the viability and applicability of DNA data storage for long-term archiving but on *how* we as a community can seek to aid its successful adoption:

Is DNA the storage medium we have been waiting for? What are the challenges and opportunities for turning this technology into a viable solution?

Can DNA storage meet with the needs of digital preservationists? If it's not possible to have a "one medium fits-all", which digital preservation needs can DNA meet?

In an increasingly uncertain world, do we need to change our requirements for long term preservation storage? Would that further open the field for a long-lived storage medium such as DNA?

How can a new technology build trust with the preservation community in order to be adopted?

This new storage medium is fundamentally different from the current mediums. What are these fundamental differences and how can established approaches to utilizing storage be adapted?

There is a unique opportunity to involve the community in the early phases of the development of DNA data storage as a product. How can we make sure the community is sufficiently and effectively engaged in the process, so that we can be confident its requirements will be met?

REFERENCES

- [1] DSHR's Blog. <https://blog.dshr.org/2013/07/immortal-media.htm>
- [2] Bertone et al. *Towards practical, high-capacity, low-maintenance information storage in synthesized DNA*, Nature, <https://doi.org/10.1038/nature11875>
- [3] DNA Data Storage Alliance. *Preserving our Digital Legacy: An Introduction to DNA Data Storage*. [Online]. <https://dnastoragealliance.org/dev/wp-content/uploads/2021/06/DNA-Data-Storage-Alliance-An-Introduction-to-DNA-Data-Storage.pdf>
- [4] Schaefer et al. *Digital Preservation Storage Criteria*, [Online]. <https://osf.io/sjc6u/>

2021 NDSA STAFFING SURVEY

Digital Preservation Intent vs Reality

Lauren Work

University of Virginia
USA

lw2cd@virginia.edu
[0000-0002-0941-6921](tel:0000-0002-0941-6921)

Elizabeth England

US National Archives
USA

elizabeth.england@nara.gov
[0000-0002-6432-8123](tel:0000-0002-6432-8123)

Sharon McMeekin

Digital Preservation Coalition
Scotland

sharon.mcmeekin@dpconline.org
[0000-0002-1842-611X](tel:0000-0002-1842-611X)

Shira Peltzman

UCLA Library
USA

speltzman@library.ucla.edu
[0000-0003-0067-2782](tel:0000-0003-0067-2782)

Juana Suárez

NYU
USA

juana@nyu.edu
[0000-0002-4574-4738](tel:0000-0002-4574-4738)

Abstract – The 2021 Staffing Survey represents the third iteration of an expansive staffing survey to be carried out by the NDSA. The survey and its findings offer a unique perspective on digital preservation staffing provision and issues experienced. This panel aims to share key findings from the survey with an international audience, to place these in the context of the real-world experience of the expert panel, and to encourage attendees to engage in a dialogue around digital preservation staffing, organizational support, and workforce development within the field.

Keywords – staffing, organization, resources, skills, training

Conference Topics – community; exchange

PANEL PROPOSAL

I. INTRODUCTION

In 2012 the National Digital Stewardship Alliance (NDSA) surveyed [1] organizations worldwide about how they address digital preservation staffing and related issues. The survey provided a useful snapshot of the digital preservation landscape and insight into how its practitioners viewed the effectiveness of their organizational structures. A version of the survey was conducted again in 2017 [2], thereby establishing the only corpus of detailed longitudinal data that touches on how the field is staffed and organized. In 2021 a Staffing Survey Working Group was convened to gather new data on staffing practices and organizational trends [3].

The Working Group was co-chaired by Elizabeth England and Lauren Work and included 13 members

from the United States and the United Kingdom. Building from the 2012 and 2017 iterations, the Working Group extensively redesigned the 2021 NDSA Staffing Survey to incorporate new areas for data collection. The redesign was prompted by findings from the 2017 survey and developments in the field over the last decade. One of the most significant changes was that, in contrast to previous surveys, the 2021 survey was designed to be answered by individuals, not organizations, and there was no limit on the number of individual respondents per organization. Participation was open to any individual worldwide with current digital preservation responsibilities at their organization, ranging from practitioners to department managers to senior leadership, and membership in NDSA was not required for participation.

The survey was sent out via listservs in early November, 2021, and was open for a period of 32 days. During this time 269 individuals from 16 countries completed the survey, continuing a trend of increasing global participation in each iteration of the NDSA Staffing Survey. Data analysis was completed between January and March, 2022, with a written report to follow in the Fall of 2022. The survey findings will build on the body of Staffing Survey data already collected, which the digital preservation community can use to identify organizational and staffing trends within the field.

II. 2021 PRELIMINARY SURVEY FINDINGS

Analysis of the 2021 survey data has revealed a number of themes across individual perspectives and organizations of varying sizes and types: sustained funding and staffing levels for digital preservation are major challenges; staff with digital preservation responsibilities often have competing non-digital preservation responsibilities; and decision-making about digital preservation lacks coordination throughout organizations. These barriers to successful digital preservation programs are contrasted with the majority of survey respondents indicating agreement with the statement, "Digital preservation is a high priority for my organization," and suggests a strong disconnect between organizational intent and practitioner reality and resourcing.

Additional findings of interest are evident when cross-analyzing respondents' demographic data with their responses to questions about how digital preservation work is organized, prioritized, and understood within organizational structures. For example, respondents identified generic, non-digital preservation specific skills/abilities such as communication, collaboration, and analytical skills as important for digital preservationists, while specialized skills/abilities such as system or software procurement/maintenance, managing continued improvement (maturity modeling/ certification), and managing budgets were the most often identified as not important or applicable. When cross-analyzing this data with the roles of the respondents, senior staff were found to prioritize "big picture" skills such as developing policies and preservation planning, while practitioners were more likely to identify targeted skills such as workflow development/implementation and experience using digital preservation tools as important.

Another key finding shows that positionality within organizational structures affects opinions on adequate levels of administrative/executive support for digital preservation. For example, the level of disagreement with the statement, "My organization has the senior-level administrative/executive support needed to manage the content we steward" correlated with the organizational positioning of respondents. Those at an administrative/executive level tended to have low levels of disagreement with the statement, while respondents with digital preservation coordination, development, or activity responsibilities had higher levels of disagreement.

The disconnect between the outlook and responses of those with more positional or organizational power vs. those with less is revealing, and suggests possibilities for higher education, training, and advocacy opportunities that are specifically geared toward bridging this divide.

III. PANEL OBJECTIVES

The panel will share key findings from the Staffing Survey that will be of interest to the international audience of digital preservation practitioners, educators, researchers, and leaders of organizations with digital preservation responsibilities that will be in attendance at iPres 2022. In response to prompts posed by the session chair, including main survey themes around digital preservation staffing, training, organization, and activities, panelists will provide commentary and reflect on results from the Staffing Survey, informed by their differing roles and experiences of staffing, education, and workforce development in digital preservation. The audience will be invited to participate in the discussion through live polling on the issues addressed during the panel, as well as short question and answer opportunities.

This panel will serve as the first sharing of NDSA Staffing Survey results at an iPres conference and will provide a unique opportunity to draw on the findings to spark robust discussions around staffing issues and trends at a global conference.

IV. CONTRIBUTORS

Lauren Work, Digital Preservation Librarian at the University of Virginia and co-chair of the 2021 NDSA Staffing Survey Working Group, will chair the session.

Elizabeth England, Senior Digital Preservation Specialist at the US National Archives and Records Administration and co-chair of the 2021 NDSA Staffing Survey Working Group, will represent the work of the survey group.

Sharon McMeekin, Head of Workforce Development at the Digital Preservation Coalition and member of the 2021 NDSA Staffing Survey Working Group, will represent a professional development and workforce training perspective.

Shira Peltzman, Digital Archivist for Library Special Collections at the University of California, Los Angeles and member of the 2021 NDSA Staffing Survey Working Group, will represent a practitioner's perspective.

Juana Suárez, Associate Arts Professor and Director of the Moving Image Archiving and Preservation program at New York University and Latin American Media Scholar, will represent a higher education perspective.

The 2021 NDSA Staffing Survey Working Group members are Rachel Appel, Brenna Edwards, Elizabeth England (co-chair), Heather Heckman, Déirdre Joyce, Margaret Kidd, Julia Kim, Sharon McMeekin, Krista Oldham, Shira Peltzman, Jessica Venlet, Hannah Wang, and Lauren Work (co-chair).

PANEL DISCUSSION

V. INTRODUCTION

The panelists first introduced the survey and reviewed the five sections of the report: (1) Background Information, (2) Digital Preservation Activities and Planning, (3) Digital Preservation Organization and Staffing, (4) Staffing Qualifications and Training, and (5) Final Thoughts about Program Staffing and Organization. Panelists discussed key report findings structured around various perspectives and experiences, including education, experience as digital preservation practitioners, and workforce development, and answered several posed questions. The panel then engaged the audience for a short series of live questions, and concluded with audience questions.

VI. IMPLICATIONS FOR HIGHER EDUCATION

One of the questions asked in the report is “What particular types of staffing are needed to reflect organizational structures that are made up of individuals who collaborate on digital preservation?”¹ This is an important consideration in regards to higher educational and academic training programs, and especially to the significant role that internships often play in graduate curricula.² The report can serve as a helpful point of reference for curriculum assessment to ensure that the content and configuration of graduate-level courses, especially when paired with internship placements, are in alignment with the competencies that survey respondents overwhelmingly identified as being valuable for digital preservation work.

¹ There are many nuances to discuss in relation to this, but it is an important question to raise in the context of assessing the digital preservation landscape.

During the panel, discussion centered on how digital preservation instruction should not be limited to “hands-on” digital preservation-specific skills but also extend to generalizable skills (e.g., communication, collaboration, and analytical skills). Generalizable skills ranked highly as “essential” for digital preservationists when respondents were asked to rate the importance of various skills/abilities. Internships’ emphasis on “hands-on” skills needs to be balanced with an understanding of staffing and skills as addressed in the report.

VII. POSITIONALITY

One of the clearest trends to emerge from the survey data was that perceptions of digital preservation seemed to shift according to respondents’ roles within their organizations. The 2021 Staffing Survey included a question about organizational positionality that would allow this issue to be explored because it was one of the key findings in a 2020 study [4] that sought to understand what was causing the high and rising levels of dissatisfaction that practitioners reported in the 2012 and 2017 iterations of the NDSA Staffing Survey. The question described four types of roles that each correlated with different presumed levels of power or authority within the organizational hierarchy, and asked participants to select the role that best described their position. This enabled cross-comparisons that revealed observable trends in the data where responses to certain questions would change depending on where the participant placed themselves in the presumed organizational hierarchy.

Not only does this echo the findings in the 2020 study, it signals a clear disconnect between those who are most directly engaged in the routine, day-to-day work of digital preservation, and those who occupy more senior leadership roles. During the panel, discussion on this issue focused on whether a divergence of opinions about digital preservation between those with the most and least organizational power/authority was inevitable and normal, or whether it was important for the digital preservation community to bridge the gap in perceptions between these groups.

² For instance, in the Moving Image Archiving and Preservation Master’s program at New York University, every student completes three internships for a total of 770 practice hours before graduation.

VIII. TRAINING RESOURCES AND CONTINGENT STAFFING

While survey results indicated that most respondents receive resources to pursue digital preservation training, and most digital preservation jobs are permanent positions, closer examination of responses revealed concerning trends.

Cross-comparing survey respondents' positionality to their answers to questions about resources for training—funding, access, and time—highlighted that those respondents presumed to have the least organizational power/authority felt the least supported by their organizations with regards to training. For example, 28% of respondents in the role “someone who performs specific digital preservation activities” said they don't receive professional development funding they could use for digital preservation, compared to 13% of respondents overall.

In the survey, “contingent” was framed as a position with an end date, including interns and student workers. While the majority of respondents reported no contingent staff in any of the four types of roles, 36% of respondents reported one or more contingent workers in the “someone who performs specific digital preservation activities” role at their organization. The prevalence of contingent staffing for the work that typically pays the least and is likely to attract recent graduates is of concern. These are the same survey respondents who are most likely to note they're not given enough funding or access to digital preservation training they need.

The panel conversation considered a common scenario presented by these circumstances in which practitioners need to get the training, in order to get the skills, in order to get the job that (hopefully) isn't contingent. Additional discussion focused on actions that can be taken to better support junior-level and/or contingent staff.

IX. PROGRESS AND PRIORITIZATION

The most frequently recurring theme amongst the responses to the survey was the desire for more staff. When asked to indicate their level of agreement with the statement “My organization has the staffing needed to manage the content we steward,” 70% of respondents “disagreed” or “strongly disagreed.” In free text responses to the question “What is one thing your organization could do to improve digital preservation?” the largest group of respondents commented on staffing levels and responsibilities,

including the need for more time to dedicate to digital preservation and for staff to have the necessary decision making authority to facilitate progress.

Several responses also mentioned the issues caused by digital preservation being “collateral duties,” secondary to other tasks and responsibilities, that their organization lacked dedicated digital preservation roles, and the toll that the constant advocacy burden of digital preservation takes on practitioners. Responses such as “we are stretched so thin” and “staff is taxed with many demands on time” raise concerns about stress and burnout within the profession. Evidence of this problem has been anecdotal to date, but the Staffing Survey provides the first quantifiable confirmation of this problem.

The lack of the opportunity to prioritize digital preservation and insufficient staff numbers are clear barriers to progress, and may also be detrimental to the wellbeing of practitioners. Of the four staffing roles identified in the survey, the role presumed to have the least organizational power/authority, “someone who performs specific digital preservation activities,” was the role for which most respondents indicated a need for increased staff. The median number of additional staff ideally needed in this role was two, illustrating that the changes needed to address these issues are not extreme.

X. AUDIENCE PARTICIPATION

During the panel session the speakers utilized the tool Mentimeter [5] to engage with those attending in-person and online, repeating questions asked as part of the 2021 survey. Each question posed in the panel was answered by 102 attendees.

The first Mentimeter slide combined questions 13, 25, 24, and 27, where attendees were asked to indicate their level of agreement (strongly disagree, disagree, neither agree nor disagree, agree, strongly agree) with a number of statements covering if digital preservation was prioritized, if it had executive support, if the implementation worked well, and if there was sufficient staffing (Fig. 1). The responses mirrored those in the main survey, with a positive skew in relation to prioritization, but with the level of agreement decreasing through executive support and implementation, and staffing levels having a negative skew.

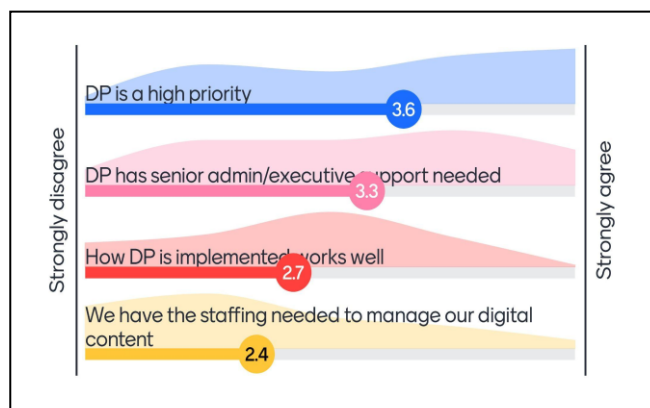


Figure 1. 1st Mentimeter slide responses

The second Mentimeter slide asked attendees to select the action that would have the greatest impact on helping to progress their digital preservation program. The options were:

- Hire dedicated/more staff
- Leadership buy-in/prioritization
- IT support/improve technology and infrastructure
- Sustained operational funding
- Give staff time/job support to focus on digital preservation
- Something else

Again, as can be seen in Fig. 2, the responses echoed those of the survey, with more than half of those responding (54 attendees, 27 for each option) selecting one of the two answers relating to staffing issues. In line with the previous question, the option relating to implementation was the next most selected (18 attendees), and those relating to leadership support and prioritization received fewer votes. Only four attendees chose the “something else” option, and while asked if anyone would like to expand on their choice, no comments were forthcoming during the panel session.

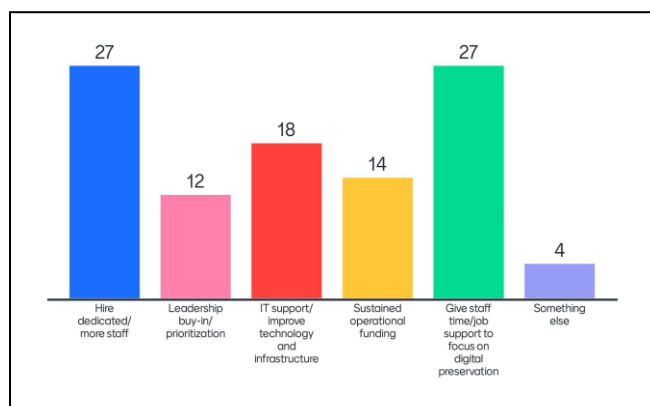


Fig. 2 - 2nd Mentimeter slide responses

XI. CONCLUSION

The panel concluded with questions from both the online and in-person attendees for the session. Topics ranged from how those in senior administrative/executive positions can best advocate for digital preservation staffing at their organizations, to further engaging with the discussion around staff health and wellbeing within digital preservation and related fields. The 2021 Staffing Survey survey and report, as well as the panel and audience discussion, highlighted digital preservation staffing issues in the field. These important issues should continue to inform future areas of study and organizational strategy for staffing.

REFERENCES

- [1] Staffing for Effective Digital Preservation: An NDSA Report, NDSA. <https://osf.io/a84dh/>.
- [2] Staffing for Effective Digital Preservation 2017: An NDSA Report, NDSA. <https://osf.io/mbcxt/>.
- [3] 2021 Staffing Survey Report, NDSA. <https://osf.io/2rb7k/>.
- [4] What's Wrong with Digital Stewardship: Evaluating the Organization of Digital Preservation from Practitioners' Perspective. Blumenthal et al. <https://elischolar.library.yale.edu/jcas/vol7/iss1/13/>.
- [5] Mentimeter. <https://www.mentimeter.com/>.

KEEPING UP WITH THE DATA:

Reflections on Fixity and Data Visualizations

Angela Beking

Privy Council Office

Canada

Angela.Beking@pco-

bcp.gc.ca

[0000-0002-6165-5970](tel:0000-0002-6165-5970)

Seeking to use data to drive decision-making in real-time, organizations are revolutionizing how they understand, produce, and use information. To sustain the value of insights derived from data, digital preservation professionals must challenge a core assumption: that data and information workflows will automatically produce static objects that can represent a complete record of our times. Using the specific example of Tableau data visualization software, I will suggest that the only way to support accountability, transparency, and justice via digital preservation is to become more active in the data and information lifecycle. To succeed, we must invest in collaboration with a broader community: specifically, data and information management professionals.

**Keywords – Data Management; Information Management; Data Visualization; Collaboration; Fixity
Conference Topics – Community; Innovation**

I. INTRODUCTION

Fixity is a core concept in digital preservation. It means assurance that a digital file has remained unchanged. Establishing and maintaining fixity demonstrates chain of custody, which allows us to ensure that digital materials are authentic [1]. But what if there is no digital file (or otherwise static preservation object) on which to establish fixity? What if data and information workflows no longer naturally create static objects?

This poster will present data visualization software as an example of a technology that does not necessarily generate a complete “record” by default. While Tableau is the example discussed here, there may be similar challenges with other tools such as Infogram, ChartBlocks, Datawrapper, or Power BI.

A. “Analytics for Everyone”

Tableau is a visual analytics platform that provides “analytics for everyone.” Tableau seeks to fuel *data culture*: the “collective behaviors and beliefs of people who value, practice, and encourage the use of data to improve decision-making” [2]. Users can visually express data by drag and drop actions into data queries through a GUI; advanced coding knowledge is not required. The resultant data visualizations are completely interactive and can be refreshed at different intervals, such as hourly, daily, weekly, or monthly. This changes the information product in real time, as the data is updated (e.g., by adding new sales figures to the connected data source) [3].

II. THE CHALLENGE

Tableau’s default file format (*.twb) does not contain the actual data that is used to create the visualization(s). This means that when the data connection changes (such as post transfer to an archival repository), the file will fail to open. While Tableau can also produce a “packaged workbook” (*.twbx) that saves local file data, this requires manual intervention. It may also be possible to create an “archive solution” in Tableau via the REST API functionality or Tabcmd; this is a current topic of research [4].

What is important is that none of these potential solutions are “out of box” functionality of the software. A software cannot predict which data and information need to be maintained, for how long, or for what purpose. Defining this is a “people and process” issue as much as it is a technical issue. The content that is required to maintain accountability,

transparency, and justice will not necessarily be created without active intervention. This poster will explore how that intervention could succeed as a broad community effort between data management, information management, and digital preservation professionals.

REFERENCES

- [1] Digital Preservation Coalition, "Fixity and checksums," *Digital Preservation Handbook*, <https://www.dpconline.org/handbook/technical-solutions-and-tools/fixity-and-checksums>
- [2] "What is Tableau?" *Tableau: A Salesforce Company*, <https://www.tableau.com/why-tableau/what-is-tableau> and "What is Data Culture?" *Tableau: A Salesforce Company*, <https://www.tableau.com/why-tableau/data-culture>
- [3] "Schedule Refreshes on Tableau Online," *Tableau: A Salesforce Company*, https://help.tableau.com/current/online/en-us/schedule_add.htm#:~:text=Daily%3A%20The%20available%20frequencies%20are,Intervals%20or%20once%20a%20day
- [4] Atul Bhagwat, "Create an archive solution in Tableau," *Tableau: A Salesforce Company*, <https://community.tableau.com/s/question/0D54T00000C5QaRSAV/create-an-archive-solution-in-tableau>

SUPPORTING PRESERVATION OF VETERAN PERSONAL ARCHIVES

Development & Use of the Virtual Footlocker Project Curricula

Edward Benoit III

Louisiana State University
United States
ebenoit@lsu.edu
[0000-0002-6707-0623](tel:0000-0002-6707-0623)

Allan Martell

Louisiana State University
United States
amarte6@lsu.edu
[0000-0001-7768-7822](tel:0000-0001-7768-7822)

Abstract – The Virtual Footlocker Project (VFP) developed online curricula supporting the preservation of contemporary veteran personal records based on a series of in-depth focus groups, with one curriculum directed at veterans and the other for cultural heritage workers. This poster outlines the curriculum development, implementation, and use focused on the digital preservation aspects of its design. The poster also demonstrates the application of the curriculum beyond the U.S. military to include broader personal digital archives and the use of the VFP curriculum in community and participatory outreach projects.

Keywords – Personal digital archives, military records, community outreach, training

Conference Topics – Community; Exchange

I. INTRODUCTION

For generations, soldiers documented their wartime experiences in personal diaries, photographs, and correspondence. Often veterans kept these treasured personal collections long after their service and handed them down to family members, with some eventually donated to archives and museums. These personal military service accounts are vital in humanizing wartime sacrifices and experiences. The contemporary 21st-century soldier no longer creates and maintains the same analog personal archives with the shift towards digital technologies over the past twenty years, thereby creating a critical future gap in the record [1].

The Virtual Footlocker Project (VFP) is an Institute of Museum and Library Services grant-funded project whose primary goal is to support active-duty military and veterans in preserving their personal military records [2]. Based on in-depth focus groups, the VFP developed a set of curricula providing veterans with the tools and training needed to identify important records, organize, store, and preserve their collections. Contemporary veterans utilize a broader array of platforms to document their time in service, incorporating both analog and digital worlds. As such, the VFP curricula support both analog and digital materials.

This poster focuses on the digital preservation aspect of the curricula design and implementation. The poster will present the key issues and challenges of personal military records, best practices for working with military members, and the adaptability of the curriculum for broader personal digital archival training and outreach projects.

II. CURRICULUM DEVELOPMENT

The VFP team conducted 22 focus groups with 99 members of the different branches of the U.S. military who served during the past 15 years. Open coding analysis of the focus group findings identified 14 major headings and 225 sub-headings with over 3,000 unique codes. The data included concerns over preserving both analog and digital objects, privacy, storage, record loss, and other concerns. The VFP team utilized the findings to create two sets of online

curricula—one for active-duty military and veterans and the second for archivists and other cultural heritage workers who wish to work with veterans.

The curricula each include four individual modules that can be completed entirely online or implemented in an in-person workshop format with a combination of audiovisual and textual content. In addition to the provided content, each module includes opportunities for participants to apply their knowledge through applied exercises.

The veteran's curriculum includes the following modules: (1) Introduction to the preservation of personal military records; (2) Organization & storage of personal military records; (3) Preservation of analog and digital personal military records; and (4) Additional resources and donation of personal military records.

The archivist's curriculum includes the following modules: (1) Introduction to personal military records; (2) Working with active-duty military and veterans; (3) VFP Curriculum for active-duty military and veterans; and (4) Creating and implementing outreach projects

III. DIGITAL PRESERVATION IN VFP CURRICULA

While the focus group data indicated veteran records include both analog and digital materials, the participants indicated significant concerns with preserving the latter. Storing digital materials remains a significant challenge for most veterans, with many noting they retained old hard drives, cell phones, and other storage devices without the ability to access the data due to missing passwords, lack of proper hardware, and other issues. Additional challenges included cloud-storage companies going out of business, social media account hacking, and privacy concerns over potentially classified information.

The VFP curricula address these concerns by introducing fundamental digital preservation approaches throughout all four modules. This includes, but is not limited to: digital storage challenges, local and cloud-based storage options, format migration, file naming conventions, IPTC social media testing, best practices for digitizing analog material, downloading and archiving email, app-based chats, and social media accounts, and locating additional resources or professional assistance.

IV. OUTREACH & ADAPTABILITY

As noted earlier, participants may complete the VFP curricula entirely online; however, a series of in-person workshops with each curriculum will be offered to veterans and archivists in 2023. Additionally, the veteran curriculum is provided as a downloadable package with PowerPoint decks, scripts, video files, PDF handouts and worksheets, and sample exercise files. The archivist's curriculum also includes a module focused on adapting the veteran's curriculum for outreach and community-based projects utilizing the example of an event supporting the digitization of service members' paper-based records. The poster presents workflows for using the veteran's curriculum in the example community event.

Although the VFP curricula remain focused on supporting military members, elements of both curricula can be easily adapted for use with non-military audiences interested in personal digital archiving. In addition to the curricula, the VFP provides access to the transcripts from all of its focus groups for future use in other projects.

REFERENCES

- [1] E. Benoit, III, "Digital V-Mail & the 21st century soldier: preliminary findings from the Virtual Footlocker Project," *Preservation, Digital Technology & Culture*, vol. 46, no. 1, 2017, pp.17-31.
- [2] <https://www.virtualfootlocker.com/>

EXPLORING SOFTWARE, TOOLS AND METHODS USED IN WEB ARCHIVE RESEARCH

Schmid, Katharina

Bayerische Staatsbibliothek
Germany
[0000-0001-6057-6640](https://nbn-resolving.org/urn:nbn:de:hbz:5:1-63664-p0000-9)

Healy, Sharon

Maynooth University
Ireland
[0000-0003-3493-0938](https://nbn-resolving.org/urn:nbn:de:hbz:5:1-63664-p0000-9)

Byrne, Helena

UK Web Archive
United Kingdom
[0000-0002-0966-4685](https://nbn-resolving.org/urn:nbn:de:hbz:5:1-63664-p0000-9)

Abstract – This paper is one part of a larger research project, titled, **Web Archives - Researcher Skills and Tools (WARST)**. In this poster we focus on the data from the WARST study which examines the software, tools and methods used in the web archive research lifecycle.

Keywords – web archive research, web archiving, web archive creators, web archive users

Conference Topics – Innovation; Community

I. INTRODUCTION

In this poster we explore the landscape of software, tools and methods used in web archive research. We consider web archive research to be inclusive of web archiving, curation, and the use of web archives and archived web content for research or other purposes [1]. We maintain that web archive research is representative of the processes and activities described in the Archive-It's web archiving lifecycle model from appraisal and acquisition, to replay, access, and use [2]. We suggest that there will always be a need to keep examining the roles of skills, tools, and methods associated with the web archiving lifecycle as long as internet, web and software technologies keep advancing, upgrading, and changing.

A. Background

This poster is one part of a larger research project, titled, **Web Archives - Researcher Skills and Tools (WARST)**. The WARST project focuses on individuals around the globe who participate in web archive research, and explores the skills, tools and knowledge ecologies in the web archive research lifecycle. Please see Healy et al. for a full documentary of their methodology [1]. This poster focuses on the data from the WARST study which examines the tools, software and methods used in

web archive research. We use Gephi, to show a network analysis of the software, tools and methods in line with two communities of practice (i) libraries, archives, and web archive environments (n=30) and (ii) academic, scholars, students, or professionals working in IT/web environments (n=14). Through a network analysis we provide some understanding of the environment and its connections.

B. Related Literature

Several other studies have done substantive work in this area focusing on web archiving initiatives and practises, users of web archives, awareness and engagement with web archives, scholarly use of web archives, or examining both web archiving practises and the challenges and opportunities for using web archives for research [3] [4] [5] [6] [7]. We build on these studies to foster discussion about the current state of collaboration and communications between web archiving initiatives and users/researchers.

II. FINDINGS & DISCUSSION

A. Data Collection

Overall there is significant overlap between the types of tools and methods the two groups of respondents use to collect data. In the library, archive and web archive environment, however, crawling software which produces data in the standard WARC format clearly dominates. The tools and methods used by participants from a scholarly or academic environment seem to be more diverse as they are influenced by the specific research question and methodology, for example when data is collected manually for close reading.

Changes in web technologies also clearly influence tools and methods for data collection.

Social media platforms, for example, are difficult to archive with traditional crawling software and generally require platform-specific software. This is reflected in the use of tools like Instaloader and Twarc to collect data directly from an API. Additionally, both groups use browser-based crawling software alongside traditional crawlers like Heritrix to capture dynamic websites that rely heavily on technologies like JavaScript.

B. Data Analysis

In the responses of both groups, we see a broad range of methods of analysis. They include manual and computer-assisted forms of analysis as well as qualitative and quantitative analyses, sometimes used in combination. From the software mentioned we can infer that input includes text and network data as well as metadata from the crawls. Tools for processing visual, audio or audio-visual data are not mentioned explicitly.

In the responses from the library, archive or web archive environment, there is a clear focus on tools and methods for search and information retrieval. They include tools for metadata search like CDX queries but also full-text search like Apache Lucene or Apache Solr. This reflects ongoing efforts to improve search capabilities and turn “web archives [...] from mere document repositories into accessible archives” [8].

Respondents from the library, archive and web archive community also refer to tools for digital forensics and digital preservation, which are not mentioned by respondents from the academic and research community. The same is true for software used specifically to process large amounts of data. This may point to fields of expertise in the library and archive community, from which other communities could benefit.

Notably, respondents from an academic or scholarly environment did not report using any of the user interfaces offered by web archiving institutions. These include tools for replaying archived web content as well as user interfaces that offer limited analytical functionalities like the SolrWayback. Further research is required to find out whether this observation holds true beyond the scope of this survey and to determine the underlying causes.

Instead, respondents seem to prefer stand-alone tools that are not specific to web archive content. As

the responses show, tools like Voyant Tools, IramuteQ and Gephi that have been developed and are widely used in the digital humanities and social sciences are also in part taken up by the library and archive community. This indicates an ongoing fruitful exchange between the two communities.

Spreadsheets are another type of standard software that is used by both communities of practice. Respondents report using it to collect and manage data as well as to conduct analyses. Both communities could therefore benefit from collaborations in developing training materials for spreadsheet software.

III. CONCLUSION

In this poster presentation we specifically focus on data from the WARST project which examines skills tools and methods used in web archive research. We surmise that the landscape is heavily influenced by changes in web and software technologies and therefore merits continuous reappraisal through studies like this. The WARST project highlights shared practices and commonalities between different communities that are involved in web archive research. By visualizing the data through a network analysis, we can examine the environment to see exactly where the commonalities are in terms of software, tools and methods. This could serve as a starting point to foster discussion for the development of training in skills, tools and methods for web archive research. For example, the findings hint at further opportunities for collaboration and knowledge exchange with regard to user interfaces for web archive collections, and training in the use of spreadsheet software for both collection and analysis.

REFERENCES

- [1] Healy *et al.*, *Skills Tools and Knowledge Ecologies in Web Archive Research*. WARCnet Special Report, 2022. WARCnet Aarhus, Denmark, <https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>
- [2] M. Bragg and K. Hanna, ‘The Web Archiving Lifecycle Model’, The Archive-It Team, Internet Archive, USA, Apr. 2013.
- [3] J. Bailey *et al.*, ‘Web Archiving in the United States: A 2013 Survey’, NDSA Report, National Digital Stewardship Alliance (NDSA), USA, , Sep. 2014. Available: <https://osf.io/h4e6z/>.
- [4] M. Costa and M. J. Silva, ‘Understanding the information needs of web archive users’, in *Proceedings of the 10th International Web Archiving Workshop (IWAW 2010)*, Vienna, Austria, September 22-23, 2010, pp. 9–16. Available:

<https://web.archive.org/web/20110723173820/http://www.iwaw.net/10/IWAW2010.pdf>.

- [5] M.-D. Costea, 'Report on the Scholarly Use of Web Archives', NetLab, Aarhus, Denmark, 2018. Available: http://netlab.dk/wp-content/uploads/2018/02/Costea_Report_on_the_Scholarly_Use_of_Web_Archives.pdf.
- [6] H. Hockx-Yu, 'Access and Scholarly Use of Web Archives', *Alexandria*, vol. 25, no. 11, pp. 113–127, 2014, doi: [10.7227/ALX.0023](https://doi.org/10.7227/ALX.0023).
- [7] Truman, G., 'Web Archiving Environmental Scan', Harvard Library, Massachusetts, USA, 2016. Available: <https://dash.harvard.edu/handle/1/25658314>.
- [8] M. Costa, 'Full-Text and URL Search Over Web Archives', in *The Past Web: Exploring Web Archives*, D. Gomes *et al.*, Cham: Springer International Publishing, 2021, 71–84. doi: [10.1007/978-3-030-63291-5_7](https://doi.org/10.1007/978-3-030-63291-5_7).

TOWARDS A COLLECTIONS MODEL FOR PRESERVATION PLANNING AT THE BRITISH LIBRARY

Michael Day

The British Library
London, United Kingdom
Michael.Day@bl.uk
[0000-0002-1443-5334](tel:0000-0002-1443-5334)

Maureen Pennock

The British Library
Boston Spa, Wetherby, United Kingdom
Maureen.Pennock@bl.uk
[0000-0002-7521-8536](tel:0000-0002-7521-8536)

Abstract – The development of a framework for preservation planning at the British Library has highlighted the need for a more-structured understanding of its digital collections, in particular with regard to identifying the specific sets of objects that would be the focus of preservation plans. Work has recently commenced on developing a model of the Library's collections to support this.

Keywords – Preservation planning, collection models

Conference Topics – Innovation

I. INTRODUCTION

The British Library is currently in the process of developing a framework for preservation planning, e.g. through projects like the Integrated Preservation Suite (IPS) [1]. One of the things that has emerged from this activity is the need for a more-detailed understanding of the Library's digital collections, especially with regard to identifying the specific collections (or object groups) that would need to be the focus of preservation plans. This poster presentation introduces an initial attempt to model the Library's collections to support preservation planning and other digital preservation activities, including repository ingest and migration.

II. COLLECTIONS AS SPECIFIC SETS OF OBJECTS

Becker, *et al.* [2] have distinguished between *preservation planning* at an abstract level and how the concept might be implemented in practice in the form of *preservation plans*. They defined the latter as, "specifying an *action plan* for preserving a specific set of objects for a given purpose." They refer to that *specific set of objects* elsewhere as a "collection," defining them at first in neutral terms as, "the set of digital objects or records for which a preservation plan is created," adding that in technical terms, however, a collection would be "all of the objects that

shall be treated with the same tool with identical parameter setting during the application of preservation actions."

Preservation plans, therefore, can only really work when they are applied to sets of objects with (at least) some features in common, for example: a collection of eBooks in EPUB format.

III. THE BRITISH LIBRARY CONTEXT

The British Library has to date developed an understanding of its digital collections from two main directions. The first might be seen as a 'bottom-up' approach, focused on the practical needs of individual ingest streams or of specific collections or projects. The main problem with this approach is that it can be difficult to link these isolated collections and sub-collections into an integrated whole. The second is a 'top-down' approach based on the production of collection profiles for all of the Library's major digital content types [3]. These profiles were designed to be a way for the digital preservation team to work with curators and collection owners across the Library to identify all relevant collections, to explore high-level digital preservation requirements, and to help specify preservation intent. The profiles are also reviewed on a periodic basis to ensure that they remain up-to-date and in-line with curatorial and user expectations.

In parallel, the IPS project has been designing and implementing a technical infrastructure based on a web-based 'workbench,' which in turn provides interfaces to: a 'knowledge base' of information about file formats and software, a repository for preserving software, and a facility for storing Library-specific preservation information, including policies, preservation plans, and collection profiles [1]. When

the project was developing an initial template for preservation plans, it soon became clear that there was a need to be able to specify collections in a more specific way than that which was made possible by the Library's collection profiles. The profiles do contain tables listing collections at a lower level, but these are typically not specific enough to form the basis of a preservation plan.

IV. MODELING THE LIBRARY'S COLLECTIONS

In order to help identify the appropriate collection levels at which to apply preservation plans, therefore, the British Library is now attempting to produce a model of its digital collections.

There was relatively little prior work to base this upon, except for a few attempts to develop formal models and ontologies for digital library services [4, 5]. Perhaps more directly applicable were the metadata models and schemas developed for collection-level description (CLD) in the late 1990s [6]. These initiatives aimed to integrate heterogeneous collections into a single discovery framework and were based on a comprehensive analytical model of collections and catalogues developed by Heaney [7].

The Library's initial ambitions were far more modest. The aim was to break down the high-level collection areas, e.g. as described in the collection profiles, and link them to collections at a lower level in diagrammatic form using the MS Visio tool (Fig. 1). These can then be used as a basis for further analysis.

This work has only just started, but the modeling so far has established at least four (approximate) layers of hierarchy. The top layer represents high-level collection areas (books, newspapers, sound content, etc.), which is then divided into 'born-digital' and 'digitized' categories (the Library's collection profiles had expressly tried to integrate these, but the distinction immediately re-emerged when the model incorporated collections at a lower-level of granularity). Individual collections and sub-collections then feature in the lower levels, which will in turn need to be broken down further in order to identify those specific sets of objects that could be the focus of a preservation plan. This is the part of the modeling process that will merit the most attention in the future and which will determine the success of the approach.

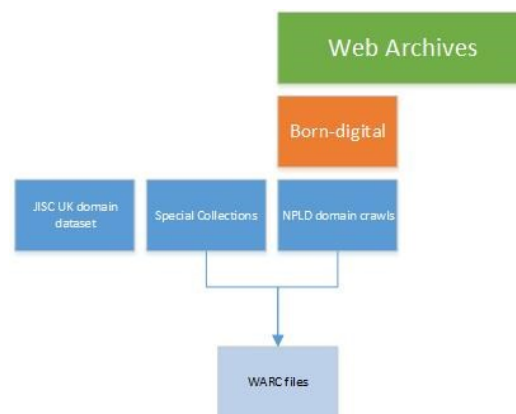


Fig. 1: Web archive collections

V. FUTURE WORK

The current focus of the collections modeling activity is pragmatic, intended to support IPS and to amplify the Library's collection profiles. It will also provide input into a major project that is underway to migrate the Library's digital collections to a new preservation repository system. The modeling of the Library's collections is still very much a work-in-progress and it will be interesting to see where it might lead next.

REFERENCES

- [1] P. May, M. Pennock, and D. Russo, "The Integrated Preservation Suite: Scaled and automated preservation planning for highly diverse digital collections," 16th International Conference on Digital Preservation, iPRES 2019, Amsterdam, The Netherlands, October 2019.
- [2] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman, "Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans," *Int. J. Digit. Libr.*, vol. 10, pp. 133-157, 2009. DOI: 10.1007/s00799-009-0057-1.
- [3] M. Day, M. Pennock, A. Kimura, and A. MacDonald, "Identifying digital preservation requirements: Digital preservation strategy and collection profiling at the British Library," 11th International Conference on Digital Preservation, iPRES 2014, Melbourne, Australia, October 2014.
- [4] L. Candela, D. Castelli, N. Ferro, Y. Ioannidis, G. Koutrika, C. Meghini, et al., *The DELOS digital library reference model: Foundations for digital libraries*, Pisa: ISTI-CNR, 2007.
- [5] M. A. Gonçalves, L. T. Watson, and E. A. Fox, "Towards a digital library theory: A formal digital library ontology," *Int. J. Digit. Libr.*, vol. 8, pp. 91-114, 2008. DOI: 10.1007/s00799-008-0033-1
- [6] A. Powell, M. Heaney, and L. Dempsey, "RSLP Collection Description," *D-Lib Magazine*, vol. 6, no. 9, September 2000. DOI: 10.1045/september2000-powell
- [7] M. Heaney, *An analytical model of collections and their catalogues*, 3rd issue, rev., Bath: UKOLN, January 2000. URL: <http://www.ukoln.ac.uk/metadata/rsdp/model/>

DIGITAL PRESERVATION IN A LUNCHBOX

Launching a community of practice

Émilie Fortin

Université Laval
Canada
emilie.fortin@bibl.ulaval.ca
[0000-0002-9717-6840](tel:0000-0002-9717-6840)

Mireille Nappert

HEC Montréal
Canada
mireille.nappert@hec.ca

Abstract – This poster presents the early days of a French-speaking digital preservation community of practice in Quebec, Canada. It showcases various considerations for defining its scope, organization, and development.

Keywords – Community of practice, knowledge sharing.

Conference Topics – Community; resilience.

I. FEELING HUNGRY [INTRODUCTION]

Even before the confinement era, the digital preservation community tried to find ways to connect. Mailing lists and conferences exist, but when we live and work in a different language than English, we long to discuss the topic in our own language.

In November 2021, a French-speaking digital preservation community of practice was launched in the province of Quebec, in Canada. Here will be discussed challenges, successes, and the future of this community.

II. LAUNCHING A COMMUNITY OF PRACTICE

A. *Lunch Prep [The Idea]*

The need was there. Every time two or more French Canadian professionals met, they were so excited to find someone that understood them. Every time, the words “We must keep in touch!” or “Let’s get together soon!” were said, but little happened afterwards.

In one instance, a few archivists were planning to meet regularly, and create a public forum using Google, but most of them could not publicly associate with their employer. Time constraints did the rest.

Same as anywhere else, Quebec practitioners (and aspiring ones) are scattered amongst different professions and organizations types. Which meant that there wasn’t enough incentive for any of the professional associations to create chapters, committees or groups focused on digital preservation.

So, how to get in touch with people from universities, museums, archival societies, and libraries then?

B. *Great (?) Expectations*

Quite simply, we wanted to be able to periodically bring together Canadian francophones involved with digital preservation to discuss the issues we are facing in our workplaces. This potential “audience” was targeted as we figured that there was no shortage of English-speaking digital preservation communities, and that Quebec and Canada’s legislation and structures will differ from other countries, such as France or Morocco.

The goal was obviously to share knowledge and tools, but also to break the isolation. We wanted to create a welcoming space that would nurture discussion, but where no one would feel required to intervene, most importantly, we didn’t want to burden anyone.

C. *Let’s Grab Lunch! [The Beginning]*

The impulse to launch the community was made by three people working in universities, and we invited people we know, thinking that they would invite people they know. It more or less worked, precisely because many of the professionals are isolated. We didn’t send an invitation to discussion lists because we wanted to gather practitioners, not

just people interested in the topic. We also aimed to keep a balance between institution type. However, as of March 2022, we are about twenty and we don't have anyone from archival societies or cities.

There has been no lack of practical considerations... Online or on site wasn't a question, with virtual meetings being the new normal, but how frequently and how long? Did we need a moderator? Reports? Documentation? Discussion list? Subjects?

At the first two meetings, we discussed those questions. We decided on 45 minutes meetings the first Thursday of each month at lunch time since it's usually the only free time left. Was it long enough or too long? At each meeting, we must cut short the discussions.

We started by using the institutional Zoom account of one of the participants. As for communication channels we created a discussion list, knowing it shouldn't be too intrusive.

We also have reports and different files on Google Drive, though we haven't had much success implementing collaborative note taking for the meetings.

Rapidly, we noticed two things: if we wanted useful and fluid discussions, we needed topics and moderation. Hence, we have been using Zoom's survey function at the beginning of each meeting to decide on the next month's discussion topic.

At the moment, facilitation is done by the participant who sent the Zoom invitation, but this is being questioned: meetings can't depend on one person and perhaps other people are better suited or interested to do it.

III. FROM A LUNCH TO A BANQUET [THE FUTURE]

Oddly enough, given our common interest for long-term goals, the biggest challenge will be the sustainability of the group. As for any community project, the success comes if many people are involved. Unfortunately, another element for success appears to be that the community shouldn't ask too much from its participants. Most of us don't have any professional incentive for such involvement, and in some cases, digital preservation is just one of many undertakings.

We'll also need to question the functioning of the group to be sure it works and remains relevant. To get there, we involving other members of the group to deal with the logistics, and to reflect with us the

need for further structure, such as a Code of conduct. Though we have not dived into the community of practice "best practices" literature yet, we believe that resources like the Open Science community starter kit¹ can be most relevant to our journey.

Finally, there is the question of the growth of the group. Starting small, by word of mouth, felt like the right choice, but we've already realized that the group has representation issues. So we need to find a way to reach out to those who aren't at the table.

We know the global digital preservation community is generous and altruistic, yet it was a bit of a (lovely) surprise to see displays of mutual support appear so early in the existence of our group. It's certainly a good omen for this young adventure!

The table has been set, guests have started filling it and sharing their dishes. So let's have a toast for present and future friendships!

REFERENCES

- [1] « Open Science Community Starter Kit », 2022.
<https://www.startyourosc.com/>.

AIA/ OLIVER WITTE COLLECTION:

A digital preservation workflow

Elisabeth Genest

Canadian Center for Architecture
Canada
egenest@cca.qc.ca

Abstract – In 2017, the Canadian Center for Architecture (CCA) collaborated with the American Institute of Architects (AIA) and Yale University to help preserve a collection of architectural software solutions from Oliver Witte, a writer and reviewer of CAD and CAD related software. The project consisted of disk imaging over 700 obsolete software, migrating them in the OAIS-compliant software preservation system, Archivematica, and finally using SCOPE, a born-digital archives access interface built by the CCA and Artefactual Systems. In addition, the CCA aims to use the Emulation as a Service Infrastructure (EaaSI) which would help view documents in their legacy environment. What makes this project unique is the CCA's collaboration with the AIA, and the considerable volume of the software collection.

Keywords – digital preservation, born digital archives, disk imaging, archival access, eaaS

Conference Topics – community, exchange

develop workflows, create specific digital forensic tools, and set up a digital lab.

In connection, the American Institute of Architecture (AIA) has amassed a collection of over 700 architectural software from Oliver Witte; writer and reviewer of such software. These mostly came in formats that are now obsolete, such as floppy disks, CD's, audio and VHS tapes. As a result, this led to an agreement in which the AIA would lend the collection to the CCA to stabilize the software and give access to the collection through SCOPE, a born-digital archives access interface built by the CCA and Artefactual Systems. In addition, the CCA aims to use Emulation as a Service Infrastructure (EaaSI), headed by Yale University, to access files in their legacy environment, instead of opening the files with modern software.

I. BACKGROUND

Founded in 1979, the Canadian Center for Architecture (CCA) is an international research institution and museum whose central premise is that architecture is a public concern. As a result, the CCA puts on exhibitions, produces publications, shares its' collection, supports research, and offers public programs related to the advancement of architecture.

In 2013, the CCA presented a series of expositions titled *Archaeology of the Digital* which focused on the development, use and impacts of digital technology in architecture. The project led to the acquisition of twenty-five archives with a significant born-digital component. In addition, a digital archives team of five was put in place to describe the archives,

II. POSTER

The poster aims to provide an overview of the workflow, tools and best practices that were used throughout the AIA/ Oliver Witte project. A description of each tool would be summarized underneath an illustration that would represent the workflow and tools used for each task. The goal is to create learning opportunities and to exchange with other information science professionals on the different ways to use technology in digital preservation.

INCORPORATING DIGITAL PRESERVATION AND ACCESS MATURITY MODELS INTO WIDER ASSESSMENT PROGRAMMES

Archive Service Accreditation and the Levels of Digital Preservation and Born-Digital Access

Melinda Haunton

*The National Archives
UK*

*Melinda.haunton@national
archives.gov.uk
[0000-0002-4885-6313](tel:0000-0002-4885-6313)*

Abstract – this poster explores how and why two maturity models, the NDSA Levels of Digital Preservation and the DLF Levels of Born-Digital Access, have been embedded into a broad management standard for memory institutions (the UK Archive Service Accreditation programme), and the results of this approach to date. The poster explains the different approaches taken to preservation and access in this embedding, reflecting different broad levels of maturity within the UK archive sector's digital activity. Finally the poster outlines early findings from the implementation and potential future developments.

Keywords – Memory institutions, maturity modelling, standards,

Conference Topics – community; resilience

I. INTRODUCTION

UK Archive Service Accreditation is a partnership programme supported by professional organisations and memory institutions across the archive sector in the home nations of the United Kingdom [1]. The programme overall aims to increase the visibility of archives and their needs within their parent organizations, to improve practice and to recognize sustainable and effective activity which meets the mission and purpose of each archive service, whether private or public sector.

II. POSTER CONTENT

Live-launched in 2013, the Archive Service Accreditation programme initially incorporated only limited content relating to management, preservation and access to digital records. This was an acknowledged gap reflecting a lack of maturity in accepted standards across the UK archives sector, and the relatively limited progress realistically to be expected of individual archive services, across the 400 or more eligible institutions.

This poster explores how the gap has been filled from 2018 onwards by incorporating external maturity models into the assessment process, and the findings to date on the impact of this work.

The development of Archive Service Accreditation content initially focused on digital preservation only, mapping against existing preservation standards and reference models. Comparison of the approaches taken by ISO16363 and the then Data Seal of Approval demonstrated strong parallels to Archive Service Accreditation. The decision was taken not to use these parallel standards as a reference point to avoid creating repetitious requirements for applicants.

Instead, the Archive Service Accreditation Committee decided to embed version 1 of the NDSA

Levels of Digital Preservation [2] as a risk-assessment matrix within the existing structure of the Accreditation assessment, alongside risk-assessment approaches already in use for analogue archive collections. The model was reframed with a new level 0, to indicate a lack of ability to deliver against the risk at a basic level. Vitally for assessment purposes within a larger standard which emphasizes responses to institutional context, the model was not simply used as a scored assessment, but to spark discussion and understanding of current practice, barriers and institutional capacity.

As version 2 of the Levels of Digital Preservation was published, this was further incorporated into Archive Service Accreditation. The Digital Library Federation's Levels of Born-Digital Access [3], developed in response to the Preservation model, have also been incorporated, but more lightly, as this is an area where many archive services in the UK have made limited progress. The Access levels are used as a reference point, rather than an assessment question.

III. FINDINGS TO DATE

Embedding externally-managed maturity models into Archive Service Accreditation has been a productive exercise overall. The incorporation of maturity models within a wider standard has raised the profile of these development tools across UK archives. Using the Levels in this way has made it possible for archive services to demonstrate an approach to development while clearly signaling that progress may be incremental.

The areas of challenge have been in communication and clarity, with a need to emphasize that not all archive services are expected to reach the highest levels across the models, depending on their resources and capacity. It has also been essential to keep abreast of changes to the models – the transition to version 2 of the Levels of Digital Preservation had to be managed carefully.

Reviewing reported performance against the Levels within a broader assessment context has shown significant variation among the applicant archive services in their understanding of the Levels and expectations of what reasonable performance looks like. This has not produced a statistically-robust evidence base to generalize about current performance across archive services. However, using

an external matrix has been a productive approach to understanding risk within context.

An emerging theme is that within the UK the publication of the Digital Preservation Coalition's Rapid Assessment Model has led to this being adopted by a growing number of archive services. It is possible therefore that the maturity model used will change in future. Archive Service Accreditation will continue to respond to the development of capacity within the UK archive sector and the development of standards and models relevant to digital content at memory institutions.

REFERENCES

- [1] Archive Service Accreditation homepage <https://www.nationalarchives.gov.uk/archives-sector/archive-service-accreditation/>
- [2] National Digital Stewardship Alliance, Levels of Digital Preservation, <https://osf.io/nt8u9/>
- [3] Digital Library Federation, Levels of Born-Digital Access, <https://osf.io/r5f78/>
- [4] Digital Preservation Coalition, Rapid Assessment Model <https://www.dpconline.org/digipres/implement-digipres/dpc-ram>

PRESERVING COLLECTIONS ON TAPE AT THE NATIONAL LIBRARY OF SCOTLAND

From Business Case to Bytes

Bell, Alistair

National Library of
Scotland
UK
a.bell@nls.uk

Hibberd, Lee

National Library of
Scotland
UK
l.hibberd@nls.uk

Russell, Alan

National Library of
Scotland
UK
a.russell@nls.uk

The National Library of Scotland collects, preserves and promotes access to films capturing Scotland and her people from the early days of film-making to the present day. Around 12 thousand items in the collection are video tapes that will soon become inaccessible as playback devices become increasingly obsolete. The Library started the Collections On Tape project in 2022 to preserve access to this rich culture through an ambitious digitization project due to finish in 2025. Around 700 terabytes of data will be created that takes advantage of the ffv1 video codec. The Library will provide unprecedented access to our visitors and video specialists and is honored to share key aspects of this project with the preservation and moving image community.

Keywords – Digital Preservation, Business Case, Audio Visual, Legacy Media, National Collections
Conference Topics – Resilience; Community

I. INTRODUCTION

The National Library of Scotland collects, preserves and promotes access to films capturing Scotland and her people, from the early days of film-making to the present day. We hold over 46,000 moving items and share details of a project to preserve access to all of our video tapes for the benefit of the nation and the world beyond.

To date over 2000 video tapes have been digitized but to protect the entirety of the Library's video tape collection we will need to digitize an additional 10,000 over the next 3 years. If we don't act quickly, it will become too expensive to do so, and the content – a large part of Scotland's moving

image heritage from the mid-1950s to late 2000s will be lost because it will be trapped in formats which are effectively obsolete, unsupported, and will be unplayable in the near future.

II. COLLECTIONS ON TAPE PROJECT

The goal of this project is to prevent these collections from becoming inaccessible. The Collections on Tape project is making all video (and audio tapes) available to watch and listen at the National Library through a process of digitization for long-term preservation that creates a combination of ffv1/mkv and web friendly files. The project started in early 2022 and will run until 2025 with a cost of around 350 thousand GBP plus an ongoing commitment to store and serve around 700 terabytes of digital video assets created by the project.

III. THE POSTER

The poster is divided into the following sections:

- A. *Introduction To The National Library of Scotland's Moving Image Collections*
- B. *Business Case For Access And Preservation*

Highlights from the business case for digitization for preservation and the research that supported it.

- C. *Collections On Tape Project.*

An overview of the project including size, duration, cost, progress to date, the project team.

D. Workflows From Capture To Preservation

Including content description, rights clearance, capture, file processing, storage, preservation, delivery, automation, throughput volumes, hardware and software used.

E. Providing Access To The Video

A depiction of the different methods of providing access to the general public and video specialists.

F. Troubleshooting and Future Plans

List of some of the issues we have or haven't resolved and more good plans for the future.

FOSTERING A DATA INFRASTRUCTURE FOR THE HUMANITIES AND SOCIAL SCIENCES

A Case Study in Japan

Ui Ikeuchi

*Japan Society for the Promotion of Science (JSPS),
Bunkyo University
Japan
ikeuchi@bunkyo.ac.jp
[0000-0002-5680-1881](tel:0000-0002-5680-1881)*

Shinsuke Ito

*Japan Society for the Promotion of Science (JSPS),
Chuo University
Japan
ssitoh@tamacc.chuo-u.ac.jp
[0000-0002-5239-7391](tel:0000-0002-5239-7391)*

Yukio Maeda

*Japan Society for the Promotion of Science (JSPS),
The University of Tokyo
Japan
ymaeda@iss.u-tokyo.ac.jp
[0000-0001-8934-0420](tel:0000-0001-8934-0420)*

Kiyonori Nagasaki

*Japan Society for the Promotion of Science
Japan (JSPS),
International Institute for Digital Humanities
nagasaki@dhii.jp
[0000-0002-5485-0567](tel:0000-0002-5485-0567)*

Takeshi Hiromatsu

*Japan Society for the Promotion of Science
Japan (JSPS),
The University of Tokyo
sna99237@biglobe.ne.jp*

Abstract – For research institutions in the humanities and social sciences in Japan, it is difficult to maintain and operate data archives on a long-term basis. Therefore, the Japan Society for the Promotion of Science (JSPS) launched the "Program for Constructing Data Infrastructure for the Humanities and Social Sciences" in FY 2018 with a five-year timeframe. The program provides (1) the Japan Data Catalog for the Humanities and Social Sciences (JDCat) to enhance data discoverability, (2) "A Guide to Data Sharing in the Humanities and Social Sciences", (3) funding and consultation to five institutions, and (4) an online data analysis system. Through these activities, the program aims to promote long-term preservation and re-use of data in the humanities and social sciences.

Keywords – Data Infrastructure, Humanities and Social Sciences (HSS), Cross-search System

Conference Topics – Community

I. INTRODUCTION

There are various types of data in the humanities and social sciences (henceforth, HSS), including individual data from social surveys, statistical tables from official statistics, texts of historical materials, image data, and many other types of data. The "Program for Constructing Data Infrastructure for the Humanities and Social Sciences" aims to promote collaborate research domestically and internationally, thereby promoting HSS through building a comprehensive data infrastructure that researchers can utilize to share data on HSS research across disciplines and countries while fostering a shared culture among researchers and institutions for HSS by funding and consultation.

II. FOSTERING A DATA INFRASTRUCTURE

A. Japan Data Catalog for the Humanities and Social Sciences (JDCat)

In July 2021, JSPS and the National Institute of Informatics (NII) launched Japan Data Catalog for the Humanities and Social Sciences (JDCat)¹, a cross-

¹ <https://jdcats.jp/>

search system for social sciences data from four research institutes: JGSS Research Center at Osaka University of Commerce²; Panel Data Research Center at Keio University³; Center for Social Research and Data Archives, Institute of Social Science, the University of Tokyo⁴; and Institute of Economic Research, Hitotsubashi University⁵. From November 2021, JDCat added humanities data from Historiographical Institute, the University of Tokyo⁶ and began full-scale operation as a cross-search system for both HSS data. As shown in Fig. 1, JDCat has a faceted search function, which allows users to find data without knowing the technical terms. Prior to it, the project created JDCat schema derived from schemata of JPCOAR (Japan Consortium for Open Access Repository) and DDI (Data Documentation Initiative).

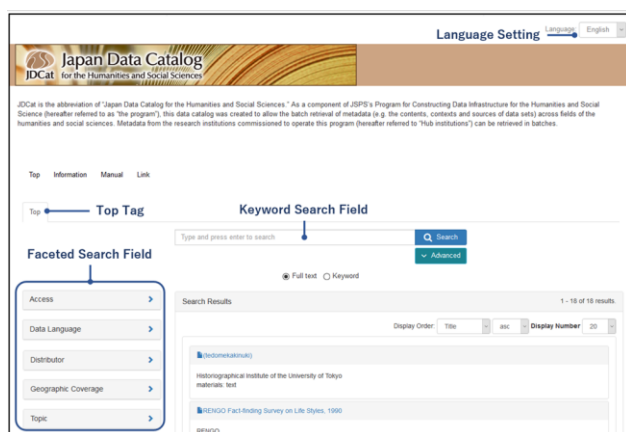


Figure 1: JDCat search screen

B. *"A Guide to Data Sharing in the Humanities and Social Sciences"*

The program published *"A Guide to Data Sharing in the Humanities and Social Sciences"*[1] in 2021. The purpose of the Guide is to help advance research in fields of the HSS. The Guide covers standards for metadata and formatting, Data Management Plan (DMP), data preservation, and data confidentiality. It helps young researchers and graduate students in fields of the HSS to effectively manage data for sharing and long-term preservation.

C. *Funding and consultation*

Under this program, five institutions were selected through an open call for proposals. The

program has provided funding and advice for the long-term operation of data archives. The development of JDCat was carried out in cooperation with five institutions, meeting on a regular basis.

D. *Online data analysis system*

The JSPS and NII have developed an online data analysis system. This system allows users to create and run R and Python programs to analyze data retrieved from JDCat without having to install statistical software or download data. Analysis programs and results can be published and shared. The system is expected to be used for collaborative research and education.

III. OUTCOMES AND FUTURE CHALLENGES

As a result of four years of activities, over 7,000 metadata from five institutions have been harvested and can be cross searched in JDCat. The cross searching of data in the fields of HSS by a wide range of users will lead to the use of data in combinations that have never been seen before and to the creation of new research results and joint research that transcends disciplines.

Future challenges for this program are to create a structure that allows institutions interested in data archiving to participate, to collaborate with data archives internationally, and to develop human resources with expertise and skills in the long-term preservation and utilization of data.

REFERENCES

- [1] Program for Constructing Data Infrastructure for the Humanities and Social Sciences, Steering Committee Working Group. *A Guide to Data Sharing in the Humanities and Social Sciences*, Japan Society for the Promotion of Science Japan, 2021. https://www.jsps.go.jp/j-di/data/guide/tebiki_p.pdf

² <https://jgss.daishodai.ac.jp/english/index.html>

³ <https://www.pdrc.keio.ac.jp/en/>

⁴ <https://csrda.iss.u-tokyo.ac.jp/english/>

⁵ <https://www.ier.hit-u.ac.jp/English/index.html>

⁶ <https://www.hi.u-tokyo.ac.jp/index.html>

LEVERAGING AI FOR VIDEO APPRAISAL

A Case Study at the World Bank Group

Paloma Beneito Arias

World Bank Group Archives

USA

Pbeneitoarias@worldbankgroup.org

[0000-0003-2509-4149](tel:0000-0003-2509-4149)

Jeanne Kramer-Smyth

World Bank Group Archives

USA

Jkramersmyth@worldbankgroup.org

[0000-0002-5689-8409](tel:0000-0002-5689-8409)

Abstract – The World Bank Group (WBG) Archives, in partnership with the WBG IT Department, has developed a tool to support the appraisal, selection and disposition of video recordings. The Archives Video Appraiser (AVA) leverages Machine Learning (ML) to make recommendations on which recordings to keep or destroy after having learnt from a trusted set of training data. Decisions are validated or corrected by an archivist so that AVA can continue to learn from its mistakes. AVA’s current predictions are 85% accurate and the use of AVA has resulted in 1.5-person-day savings per month and in the reduction of manual mistakes.

Keywords – Artificial intelligence, machine learning, video appraisal, digital preservation

Conference Topics – Innovation

I. INTRODUCTION

The World Bank Group (WBG), founded in 1945, is the oldest and largest multilateral development bank in the world. It is one of the largest sources of funding and knowledge for developing countries; a unique global partnership of five institutions dedicated to ending extreme poverty, increasing shared prosperity, and promoting sustainable development. With 189 member countries and more than 12,000 staff worldwide, the WBG works with public and private sector partners, investing in groundbreaking projects and using data, research, and technology to develop solutions to the most urgent global challenges.

The WBG Archives provides the public with access to the archival holdings of the WBG along with engaging tools that enable the discovery of historical information. It also has fiduciary responsibilities for current records management and information governance within the WBG, made possible by policies, programs and services provided internally.

One of the responsibilities of the WBG Archives is to appraise the business, legal and research value of WBG records to identify records with permanent value and records that can be destroyed when no longer needed. The ever-expanding volume of uncategorized digital records makes the manual appraisal and selection of records increasingly labor-intensive- and prone to mistakes. In 2021, WBG archivists and their IT counterparts developed a solution to test the effectiveness of using Machine Learning (ML) technology to assist the Archives in the appraisal and selection of born-digital moving image records.

Our poster will introduce the results of this case study: the design of the so-called Archival Video Appraiser (AVA), a ML tool that generates recommendations on which video recordings should be permanently preserved or destroyed based on a set of pre-defined appraisal criteria. Furthermore, AVA also automatically transfers the records to a designated folder based on the decisions.

The poster will be structured as follows:

- The problem we were struggling with
- Goals of the project
- Results
- Technology used
- Lessons learned

II. PROBLEM STATEMENT

The exponential growth of digital content makes it hard, if not impossible, for the WBG Archives to appraise the vast volume of unclassified records that we receive into custody. An example are video recordings, of which we receive an average of 200 per month. Many of these recordings have permanent value according to our Retention Policies,

while others need to be destroyed. Some recordings are transferred to our custody with promising and descriptive titles, but contain no sound or content (e.g., a meeting scheduled to be automatically recorded via Webex is cancelled and recorded anyway).

The current process of human-driven appraisal of recordings requires the visual review of a one-minute segment of every video to ensure that there's content and sound and to determine which retention rule applies. Archivists also use metadata, such as the title or the date of the recording, to support the decisions. Once decisions are documented, the archivist manually transfers the videos to different locations depending on whether they are eligible for ingest into the WBG's digital preservation platform (internally named Digital Vault) or are ready to be destroyed.

The process takes about 2-person days per month and is prone to errors due to its very manual nature.

III. PROJECT GOAL

Our goal was to develop a tool to support the appraisal of WBG moving image records and the automated staging of the videos after appraisal for ingestion into the Digital Vault, or their destruction when appropriate. Our objectives were to speed-up decision making and transfer, to increase the accuracy of the appraisal decisions.

IV. RESULTS

AVA scans designated network folders for new videos and stores the information in a database. AVA then extracts and analyzes the filename and a generated transcript, identifies empty and soundless recordings, and provides recommendations for archival retention or destruction. AVA's preliminary decisions are validated by an archivist. Thereafter, the videos identified for archival retention are transferred to a digital preservation staging area and those identified for destruction are destroyed. Audit reports are generated and automatically captured in the WBG's Electronic Document and Records Management System.

AVA's accuracy currently averages 80%. AVA is also successful in detecting empty and soundless videos. The tool requires approximately 10 minutes to make the recommendations for 200 videos. Once the recommendations are available, an archivist

needs about 30 minutes to verify and take any required corrective actions.

AVA's implementation has resulted in 1.5-person day savings per month, a reduction of manual errors, and an increase in appraisal decision accuracy.

V. TECHNOLOGY USED

M365: PowerApps, PowerApps API and Power Automate (Flow).

AWS: API Service, Python, Flask, Celery, ECS Lambda, Rekognition and RabbitMQ.

VI. LESSONS LEARNT

Identifying a representative set of training data requires a high initial time investment for the archivists. To support AVA's learning process, the archives provided 3000 carefully verified appraisal decisions that were used to teach the tool how to differentiate between permanent and temporary content. In addition to the videos selected by AVA, it is important to ingest any available metadata about the full set of videos, such as dates, meeting titles, meeting room, participants, and others.

A human driven iterative training process is still required to continue training the tool and increase accuracy. Archivists need to plan for future developments and ensure that those plans reflect the current technology and are funded appropriately.

Our current use case is relatively simple, requiring evaluation of straightforward criteria. Scaling it to larger collections of mixed formats, etc. will likely present greater challenges, such as more AI bias because of human errors on the training data or because the training data is not representative enough.

JOURNEYS TO PARTICIPATORY DIGITAL PRESERVATION:

Challenges for the Gold Museum as a Public Museum Engaging Communities with Geopark

Yi-Ting Lin

Information Studies, University of Glasgow

United Kingdom

2542320L@student.gla.ac.uk

[0000-0001-9605-0296](tel:0000-0001-9605-0296)

Abstract – Using the Gold Museum in the Shui-Chin-Chiu Geopark as an example, this research simulated user journey maps to unpack usability issues and related problems in participatory digital preservation. This is in response to the Taiwanese government attempts to integrate digital content across museums.

Keywords – Community-Engagement, Geopark, Sustainability, Participatory-Digital-Archive, Museum-Communication

Conference Topics – Community

I. INTRODUCTION

The participation of community members in the digital preservation process can potentially change the policies and decision-making system in a museum. The community-based exhibit has been in museum practice since the 1990s [1], but museums have been improving community engagement strategies. For example, the Gold Museum in Taiwan has been a member of the Chiufen-Chinkuashih-Shueinandong Geopark (a.k.a. Shui-Chin-Chiu Geopark) since 2020 [2]. Before this, the Gold Museum has been practising a participatory budget, allowing local people to proffer their opinion on budget distribution since 2016. These new strategies bring new challenges to digital preservation, and this research aims to detect them by simulating user journey maps.

II. BACKGROUND

It is possible to conduct digital preservation in a geopark. Successful cases have been seen in Britain, such as the Shetland Museum and Archives. Located

in Geopark Shetland, a global geopark in Scotland, the Shetland Museum and Archives annually welcomes 83,000 visitors [3] and has built an online catalogue. Its online catalogue currently contains more than 100,000 entries and still growing [4].

The Gold Museum has digitised its collection since 2010 [5], [6] as well. Until 2020, the Gold Museum had 3D scanned 12 contemporary metal crafts [6] and digitised 1,045 items with 2,907 images and 192 historical documents, including 20,375 pages and 25 individual pictures [5].

However, the preservation in the Gold Museum is a governmental instituted process. These created a delicate situation for the Gold Museum to participate in the Shui-Chin-Chiu Geopark. On the one hand, developing a participatory archival platform in the Gold Museum is difficult with limited funding. On the other hand, the Gold Museum needs to follow national policies, delivering digital collections to the public via the Consolidated Cultural Archives System (CCAS) [6, p. 19], which has poor usability [7], [8]. This arrangement might be why the Gold Museum only authorised seven cases to use their collection from 2018 to 2020 [5], [6], [9].

III. METHODOLOGY

This research adopted the DCC curation lifecycle model [10] to illustrate how controversies developed while the museum balanced the need to engage with the public and follow the policies. In terms of detecting potential challenges in usability issues, this

research borrowed the persona created by the National Archives to simulate user journey maps.

IV. CONCLUSION

This study is a pilot study in my PhD research. By scoping the problem with user journeys, this study provided a basis to understand users from the source community and the role of the museum. The result will continue to be examined in subsequent interviews. The observation from this study can assist museums in avoiding risk and conflict while developing a socially sustainable digital preservation plan. Specifically, this study discovered a range of subjects for museums to communicate with stakeholders and investigate in user research. In this way, the museum can effectively respond to conflicting requests from the local communities and the governmental agencies in the context of a geopark.

REFERENCES

- [1] R. B. Phillips, "Part 3: Community Collaboration in Exhibitions Section Introduction," in *Museums and source communities: a Routledge reader*, L. L. Peers and A. K. Brown, Eds. New York: Routledge, 2003, p. 280.
- [2] C.-C. (黃家俊) Huang, "The Role of Gold Museum in the Chiufen-Chinkuashih-Shueinandong Geopark (黃金博物館在九份金瓜石水湳洞地質公園的角色定位)," *臺灣博物季刊 大地的變奏曲—講寫人與自然的新關係 / Var. a Theme Earth Compos. New Human-Nature Relatsh.* /, vol. 150, no. 40, pp. 12–21, 2021, doi: 10.29879/TNS.
- [3] Shetland Amenity Trust, "Shetland Museum and Archives," *Official Website of the Shetland Amenity Trust*, 2022. <https://www.shetlandamenity.org/shetland-museum-and-archives> (accessed Mar. 07, 2022).
- [4] Shetland Museum & Archives, "Online Resources," *Official Website of Shetland Museum & Archives*, 2022. <https://www.shetlandmuseumandarchives.org.uk/collect-ions/archive/resources> (accessed Mar. 07, 2022).
- [5] N. T. C. G. (新北市立黃金博物館) Gold Museum, "The Annual Report of the Gold Museum: 2020(2020新北市黃金博物館年報)," New Taipei City, 2020. [Online]. Available: https://www.gep.ntpc.gov.tw/files/file_pool/1/0L167693925174268441/2020年新北市立黃金博物館年報.pdf.
- [6] N. T. C. G. (新北市立黃金博物館) Gold Museum, "The Annual Report of the Gold Museum: 2019 (2019新北市黃金博物館年報)," New Taipei City, 2019. [Online]. Available: https://www.gep.ntpc.gov.tw/files/file_pool/1/0L167694738610611595/新北市立黃金博物館2019年報.pdf.
- [7] S.-C. Chang, "Cataloging of Cultural Objects - Issues and Suggestions for Improvement: Example of the National Museum of Taiwan History (博物館藏品編目實務面向的問題與改善建議—以國立臺灣歷史博物館為例)," *博物館學季刊*, vol. 32, no. 3, pp. 63–79, 2018, doi: 10.6686/MuseQ.201807_32(3).0004.
- [8] S.-J. (陳叔偉) Chen, "文化部文物典藏共構系統中分類欄位的標準化問題," *國立自然科學博物館館訊*, vol. 389, pp. 4–5, Apr. 2020, Accessed: Jun. 14, 2022. [Online]. Available: <http://edresource.nmns.edu.tw/ShowObject.aspx?id=0b81a1f8c90b82107e75>.
- [9] N. T. C. G. (新北市立黃金博物館) Gold Museum, "The Annual Report of the Gold Museum: 2018 (2018新北市黃金博物館年報)," New Taipei City, 2018. [Online]. Available: https://www.culture.ntpc.gov.tw/files/file_pool/1/0L167698121208750637/新北市立黃金博物館年報2018.pdf.
- [10] S. Higgins, "The DCC Curation Lifecycle Model," *Int. J. Digit. Curation*, vol. 3, no. 1, pp. 134–140, Dec. 2008, doi: 10.2218/IJDC.V3I1.48.

STRENGTH IN NUMBERS

Sharing digital preservation good practice in the United Kingdom through community networks.

Laura Peaurt

University of Nottingham
United Kingdom
Laura.Peaurt@nottingham.
ac.uk

Rachel MacGregor

University of Warwick
United Kingdom
Rachel.MacGregor@warwic
k.ac.uk
[0000-0002-4296-6159](tel:0000-0002-4296-6159)

Abstract – Building a community of practice was a key driver for the establishment and development of the Midlands Digital Preservation Network. This poster outlines how we went from the happy accident of creating an online network and developed a safe space to share problems, successes and failures.

Keywords – collaboration, networks, advocacy, education

Conference Topics – Community; Resilience

I. INTRODUCTION

Making progress in digital preservation is about standing on the shoulders of giants and our network, MidiPres (Midlands Digital Preservation Network) came about directly inspired by hearing the fantastic work already being done by our fellow practitioners in the community, notably Australasia Preserves[1]. After hearing their story at iPres 2019[2] we reflected on our position of privilege in being able to attend an international conference and the irony of traveling abroad to exchange ideas with near neighbors. We decided to establish an informal network of anyone in our geographic region (central England) who had an interest in digital preservation - with an emphasis on reaching out to those who had little or no practical experience in this area. What began as an idea for an in-person meetup was forced by the pandemic to go online which with hindsight was a key to its success. Two years on we have a vibrant and accessible forum where members share questions, problems, successes and failures.

II. HOW WE WORK

A happy accident of being forced online became a key strength. Members no longer needed to give

up half or a whole day to travel (even locally) and everyone became accustomed to working collaboratively in this space. A further strength is the diversity in membership and the vendor agnostic nature of the forum. We strive to create a “safe space” where people can ask questions and admit failures that they would not do in a public forum. It also allows for informal benchmarking and comparing progress with others in a similar position. Members can share examples of achievable good practice rather than the more innovative and resource hungry solutions which may be out of reach when taking early steps on the preservation journey.

III. OUR STRENGTHS

Having a community which is local means that many (although not all) of us knew one another - this made things easier both at the start, particularly when holding an inaugural meeting in an online format, and as we have continued to develop. A relationship was already established to allow a friendly and respectful environment within which to discuss a wide variety of topics. The diversity of experience is important - there are those of us who are able to share expertise and knowledge - but too much expertise can feel intimidating and overwhelming for an individual at an institution that is still at the beginning of its digital preservation journey. Our aim is to build confidence through mutual support and reassurance. For example, we are able to share knowledge and practical demonstrations of new tools and learning opportunities which many members would not otherwise have the time to experiment with

themselves. The ability not just to gain subject specific knowledge but to keep it current and relevant is vital in digital preservation, as outlined in the DigiCurv framework[3] but this is time consuming and extremely difficult for those for whom digital preservation is only a part of their overall responsibilities. MidiPres directly promotes knowledge sharing and connecting with the wider digital preservation community.

IV. CHALLENGES

We want to grow and flourish as a community but how do we do this without losing what we have built?

1) *Sustainability*: The group is heavily reliant on the two founders - if either moves on the group currently risks having the organizational ability to continue. Is it possible to link this network in with other similar existing networks without raising barriers to entry such as membership costs?

2) *Growth*: Membership currently stands at around 30 people and meetings usually attract about two thirds of these. This works well in the format we have - successful recruitment of new members might jeopardize this.

3) *Geographical remit*: This was set originally as we had envisaged an in person local meet up. We have already extended our welcome a little to members outside the region - do we retain the limit to help keep the focus or throw open the doors wider?

V. CONCLUSION

The group arose when the founders perceived a gap for themselves and other isolated digital preservation practitioners for practical experience-based knowledge sharing, and on the ground support at a local level. We recognised that financial and staffing barriers often stood in the way of membership to some of the excellent existing support networks. The ability to share workflow experiences, successes as well as failures and pool our professional knowledge in a world of rapidly changing technologies, platforms, and limited resources has proved extremely beneficial. The diverse membership and wider networks we are linked to means that problems which are beyond our resources to address can be shared out with the wider community linking our little group with the rest of the digital preservation world.

REFERENCES

- [1] Australasia Preserves.
<https://www.australasiapreserves.org/p/australasia-preserves.html>
- [2] The Australasia Preserves Story: Building a digital preservation community of practice in the Australasian region, iPres 2019. <https://osf.io/njsyh/>
- [3] DigCurV Practitioner Lens on Digital Skills, DigCurV.
<https://digcurv.gla.ac.uk/practitionerLens.html>

RESEARCH WEEKS

Create Time to Create Change

Peter May

British Library

London, UK

Peter.May@bl.uk

[0000-0001-8625-9176](tel:0000-0001-8625-9176)

Abstract – Technology companies often run ‘hack weeks’ allowing staff to spend some time exploring new ideas, developing, learning, and collaborating on something of personal interest. The organizational incentive is to reap the benefits derived from allowing such freedom to work on a personal project. This practice has been used within the digital preservation domain before, the AQUA and SPRUCE projects for example, which brought together content holders and developers to quickly develop solutions to content challenges. Inspired by this, the digital preservation team at the British Library undertook a ‘research week’ to enable staff to focus on some digital preservation related work or training they have lacked core time to do; to create time for them to innovate. This poster aims to share our experience, share how we made it happen, and generally open up discussions on digital preservation research and development approaches within organizations.

Keywords – Research, Innovation, hack week, training

Conference Topics – Community; Exchange

I. INTRODUCTION

Technology oriented companies and departments, either formally or informally, often provide scope for employees to spend a portion (typically 10-20%) of their time away from their day-to-day activities and explore some new idea, learn, collaborate, or otherwise work on a personal project they are interested in [1, 2, 3]. A chance for staff to undertake a side project and make progress on those great ideas they just wish they had time for. The incentive for organizations is that it might just lead to the next ‘big thing’, or even just the next ‘little thing’ that brings in more customers, resources, revenue, etc. But 10 or 20% – half-day/one per week – can also be insufficient to make significant headway, especially when context switching away

from ‘normal’ work is factored in. Equally, pressure in achieving work goals and not letting down colleagues can also prevent staff from using this allocated ‘hack’ time. One solution to this is to combine all these half-days into one block of focused time, a ‘hack week’ [4].

Hack events are not unheard of in the Digital Preservation community either. The AQUA [5] and SPRUCE [6] projects both ran collaborative multi-day events designed to bring a mixed skillset of individuals to the table to solve collection-specific problems. These events fostered quick and innovative solutions that might not otherwise have been developed, especially if those organizations with the problems did not have the technical capacity to address those challenges; with these hack events, the community as a whole benefits.

The Open Preservation Foundation have also run several hack events dedicated to JHOVE and documentation [7, 8]. These were focused periods of time which brought the community of digital preservation practitioners together to move tasks forwards. By having a dedicated event that collaborators can request to their management to be part of, those participants get a dedicated time to do something they’re interested in and learn something new (albeit focused around a particular tool), as well as a sense of belonging and accomplishment. Their organizations, on the other hand, get improved tools and documentation which hopefully supports their objectives.

No doubt there are other events going on within digital preservation circles, which it would be good to hear about and share experiences on.

II. HACK WEEKS IN PRACTICE

At the British Library, the Digital Preservation Team concern themselves with a variety of tasks devoted to building and supporting digital preservation practice across the Library. We undertake a variety of analytical tasks, as well as research and development activities to develop solutions that meet real-world preservation needs faced by our colleagues. Within our team we have informal arrangements to allow colleagues to spend time on novel ideas or self-driven learning and development opportunities. But, as also faced by staff in the tech companies, actually getting to make use of that time can be challenging when there is so much else to do [9]. So, within the technical arm of our team, we decided to experiment with the equivalent of a 'hack week' to allow us dedicated time to undertake some research that each of us have been wanting to do.

During our week, a variety of work was undertaken. The main direction given to staff was that it should relate in some way to the work they were already doing. Some people used it to develop novel technical solutions they had been wanting to work on, others took time to read up about new technologies, others took it further and developed prototypes. For some it was simply a useful time to focus on self-improvement and learn an existing technology applicable to upcoming work.

Running a hack week takes a little more effort than just deciding to do one though. Yes, you need to give people the time, but there are a few other considerations to making it happen and making the week a success.

Foremost, getting buy-in from participants and management is essential. Managers need to understand the benefits of doing this and how it balances with the time spent. This can be hard when participants' interests extend beyond immediate team goals, and when individual deliverables may be unknown.

It can be especially difficult if you want to run research weeks more than once too. Our current aim is to run these events twice a year. A poorly performing first event can undermine any support gained with management though, so continuation depends on performance, which depends on planning.

A framework is needed for running the week, involving kick-off, catch-ups, and round-up meetings. Participants need supporting in the lead up to the event too. Do they know what's expected of them? Do they have something to work on? Do they have a plan?

And they need support afterwards. What happens to that new knowledge, that prototype, those new skills once the week has finished? Is it shared across the team/organization? How are ideas taken forward?

Preparation is key.

III. POSTER

We have two goals with this poster. We want to share our experience of research weeks and why we think they are beneficial, but also to engage with the broader community in a more one-to-one fashion to understand and learn from the experience of others. A poster is an ideal way to have those conversations with iPRES colleagues.

As such, this poster will provide information and talking points surrounding: our motivation for organizing research weeks; challenges we faced in getting started and during the events; the broad framework that we employed to run the week; and how we supported staff in the lead-up, during, and afterwards.

We have another research week planned between now and iPRES and so will have further experience to share.

REFERENCES

- [1] "Spotify's 2021 Hack Week Focuses on "Making Space", 19 Mar 2021, <https://newsroom.spotify.com/2021-03-19/spotify-s-2021-hack-week-focuses-on-making-space/>
- [2] "Be a force for change: Hack Week 2019", 17 Dec 2019, <https://blog.dropbox.com/topics/inside-dbx/be-a-force-for-change-hack-week-2019>
- [3] "Inside Atlassian: Building a culture of innovation", Dan Garfield, 23 Nov 2015, <https://www.atlassian.com/blog/inside-atlassian/how-atlassian-builds-innovation-culture>
- [4] "Organising a hack week", Joakim Sundén, Spotify R&D, 15 Feb 2013, <https://engineering.atspotify.com/2013/02/organizing-a-hack-week/>
- [5] <http://wiki.opf-labs.org/display/AQuA/Home>
- [6] <http://wiki.opf-labs.org/display/SPR/Home>
- [7] "JHOVE Online Hack Day Report", Becky McGuinness, 19 Oct 2016, <https://openpreservation.org/blogs/jhove-online-hack-day-report/>

- [8] "Spring Hackathon 2020",
<https://openpreservation.org/events/spring-hackathon-2020/>
- [9] "Side Project Programs Can Have Major Benefits for Employers", Tammy Xu, 6 Oct 2020,
<https://builtin.com/software-engineering-perspectives/20-percent-time>

UPSCALING THE MPT

Visualizing the Performance Impact from Application Configuration

Peter May

British Library

London, UK

Peter.May@bl.uk

[0000-0001-8625-9176](tel:0000-0001-8625-9176)

Kevin Davies

British Library

Boston Spa, Wetherby, UK

Kevin.Davies@bl.uk

[0000-0001-6522-9568](tel:0000-0001-6522-9568)

Abstract – The Minimum Preservation Tool (MPT), developed by the British Library, provides a local technical digital preservation environment to routinely fixity-check collections awaiting ingest into a long-term digital repository. Within the Library this was deployed on a standard-provision Virtual Machine using the same SHA-256 hash function as our long-term digital repository. This operates effectively for the collections currently under MPT control, but to be confident managing larger and more varied collections, we look to understand how performance can be improved, for example through use of different hash functions or through greater parallelization. This poster outlines experimental findings exploring the effect on performance of four different hash functions and four sizes of parallel processes, across three broad corpora (large, mixed, and small file-sizes).

Keywords – Minimum Preservation Tool, Performance, Checksum

Conference Topics – Innovation

I. INTRODUCTION

Trustworthy digital repository systems are a crucial component to maintaining long-term access to authentic digital content. Content obviously has to be deposited into such systems for them to perform, however the Library's experience has been that there is often a delay between acquiring content and ingesting it into a long-term repository. In this interim period, backing up content helps replicate the data, somewhat securing content at the bit-level, but does not necessarily prevent, detect, or raise awareness of bit-level corruption.

In order to safeguard this 'interim' digital material, the Digital Preservation Team at the British Library have developed the Minimum Preservation Tool (MPT) [1, 2] to provide basic integrity checking

across replicated interim data-stores. This open-source tool provides simple checksum generation, validation and cross-data-store comparison, combined with a reporting mechanism for each of these functions.

We have deployed MPT to several collections of various sizes, using a Library standard-provision virtual machine (VM) and the SHA-256 hash function to generate checksums. These choices were based on what could easily be provisioned and what hash functions are used within existing workflows.

During the time we have been using MPT under this setup it has performed effectively. We have gained experience surrounding the execution times and scheduling needs to service the collections in care. Staggering checksum validation tasks so that larger collections are validated on alternating weeks is one example of this. But as further collections are identified for MPT control, and as the amount of data to be protected increases, the question has arisen of how MPT's performance can be improved. Code analysis can be undertaken to look for efficiencies, and the VM configuration could be enhanced, but what efficiencies can be gained due to application configuration? Could we increase parallel processing of files? Or could changing the hash function used to generate checksums improve performance?

With this in mind, we initiated an internal project to look at how to upscale the MPT service. Broadly this covers two main areas: 1) investigations in a test MPT environment to understand the impact of virtual machine and application configuration on performance; and 2) to understand the deployment

and scheduling benefits afforded by containerizing the service (this latter work has yet to start).

II. INVESTIGATIONS

This poster will focus on two application configuration investigations undertaken: 1) the effects of hash function choice on MPT checksum validation performance; and 2) the effects of the number of parallel processes on processing time.

A. Environment Setup

Setup required creating a suitable test environment - a virtual machine, provisioned in alignment with standard VMs supplied by our IT department (4 x logical AMD Opteron 6276 cores, 8 GB physical memory, Windows Server 2019), with locally attached storage for test data.

In terms of test data, as the Library handles many different digital collections with varying makeup of files, we wanted to understand the overall performance across broad categories of collections. We generated three file corpuses of up to 1TB each - Large (262 files, >4GB/file avg.), Mixed (21k files, 8.5MB/file avg. (s.d. 116MB)), and Small (3.8m files, 138KB/file avg.) - each representing a broad variety of the collections held by the Library.

B. Hash Function Choices

MPT uses the standard Python hashlib library to generate checksum digests, which provides support for most commonly-used cryptographic hash functions. Of these SHA-256 was chosen as the benchmark function due to its common use in the Library already, MD5 was chosen as another frequently used algorithm, and finally BLAKE2 was selected as it showed performance improvements over SHA-256 and MD5 [3].

Non-cryptographic hash functions are typically faster [4] and, given the nature of the MPT process is to detect changes to file bit streams, such algorithms were considered acceptable. The XXHASH algorithm [5] was selected as it is considered the fastest [6], but implemented in MPT through another library [7].

Algorithms selected, checksums were generated for all files in each corpus using the test environment. Timing information was provided by the MPT reporting mechanism.

C. Number of Parallel Processes

MPT supports parallel processing using Python's multiprocessing module. Each hash function was

tested using 8, 16, 24 and 32 parallel processes in an attempt to find the optimal point where the balance of CPU usage versus disk response time provides the lowest overall processing time. Experiments were again performed using the selected algorithms on the test environment across all three corpora.

III. POSTER

The poster will present an overview of our ongoing investigations and findings to date, with particular emphasis on hash function choice and optimal number of processes. It will outline the MPT to set the scene for those unfamiliar with the tool, give details about the experimentation setup and variables under scope, as well as present results obtained.

Our aim is to share knowledge of the MPT tool and experimental evidence demonstrating how to optimize its usage. The poster therefore aligns with the call for contributions by supporting colleagues across all organizations and sharing research that influences practice.

IV. ACKNOWLEDGEMENTS

We would like to thank the other MPT team members who helped contribute to this work.

REFERENCES

- [1] K. Davies; P. May and D. Russo, Minimum Preservation Tool, 2020-07-22, Github: <https://github.com/britishlibrary/mpt>
- [2] M. Pennock, J. Beaman, P. May and K. Davies, "Back to Basics: The Minimum Preservation Tool," In Proceedings of iPRES2021, 17th International Conference on Digital Preservation, 2021.
- [3] S. Chang, R. Perlner, W. E. Burr, M. S. Turan, J. M. Kelsey, S. Paul, L. E. Bassham, "Third-Round Report of the SHA-3 Cryptographic Hash Algorithm Competition", November 2012, <https://nvlpubs.nist.gov/nistpubs/ir/2012/NIST.IR.7896.pdf>
- [4] T. Claeson, A. Sateesan, J. Vliegen, N. Mentens, "Novel Non-cryptographic Hash Functions for Networking and Security Applications on FPGA", September 2021. https://www.researchgate.net/publication/354610638_Novel_Non-cryptographic_Hash_Functions_for_Networking_and_Security_Applications_on_FPGA
- [5] Yann Collett, xxHash, 2012 - 2021, Github: <https://github.com/Cyan4973/xxHash>
- [6] B Buchanan, "When Fast Just Isn't Enough: Nobody inspects the spamish repetition", 2018-08-08. <https://medium.com/asecuritysite-when-bob-met-alice/when-fast-just-isnt-enough-nobody-inspects-the-spamish-repetition-3458a96aa04e>
- [7] Yue Du, xxhash, 2014 - 2020, Github: <https://github.com/ifduyue/python-xxhash>

BRINGING TRANSPARENCY AND PERMEABILITY TO ORGANIZATIONAL SILOS

Improving Workflow and Culture

Daniel Noonan

The Ohio State University
United State of America
noonan.37@osu.edu
[0000-0002-7021-4106](tel:0000-0002-7021-4106)

Sue Beck

The Ohio State University
United State of America
beck.697@osu.edu
[0000-0002-0895-3733](tel:0000-0002-0895-3733)

Abstract – The Ohio State University Libraries established the Digital Preservation and Access Workgroup (DP&A) [1] in early 2020 to guide the University Libraries' policies, strategies and tactics for managing, preserving and providing access to its digital collections. It brings together key individuals from across the organization to ensure that information sharing and best practices are reflected throughout the organization. The DP&A's initial charge was to identify our existing workflows that affect born digital acquisitions and processing, digitization, providing access to digital materials and the preservation thereof. This effort is to aid in answering the question, "What are the intersections, gaps, redundancies and areas for improvement?" This poster will demonstrate the progress we have made on this project, spotlighting the process analysis and improvement techniques we have brought to bear, along with our initial recommendations for workflow and organizational improvement.

Keywords – Digital Preservation, Digitization, Institutional Prioritization, Process Improvement, Workflow Analysis

Conference Topics – Innovation; Resilience.

I. INTRODUCTION

At the beginning of 2020, a group of librarians and curators proposed the creation of a workgroup to provide a cross-functional, consistent approach to managing The Ohio State University Libraries' (University Libraries) born digital acquisitions and digitized materials.

Various University Libraries' workgroups have come together over the past decade investigating issues pertinent to its digital content with success in developing guidance, while other groups' efforts have not necessarily seen the light of day. Further,

there is confusion at times as where to find definitive University Libraries' information regarding digitizing materials, accessioning born digital materials, and where it will preserve and provide access to these materials.

One of the goals of this workgroup, Digital Preservation & Access—or DP&A—is to provide a single point of access to find, discover and manage this institutional knowledge. Further, the DP&A, intends to investigate and develop the means by which it can provide transparency in decision-making for determining priorities, guidelines and standards that the Libraries adopts in these areas.

The initial charge from the sponsoring Associate Deans, meant to eventually achieve these loftier goals, is something much more basic, was to identify the University Libraries' existing workflows that affect born digital acquisitions and processing, digitization, arrangement and description, providing access to digital materials and the preservation thereof. Answering the questions:

"What are the intersections, gaps, redundancies and areas for improvement?"

"How do we approach improving workflow when under-resourced—fiscal and human?"

This poster provides a case study of the work completed thus far by the DP&A from data collection to the development of documented, visualized workflows to the initial set of recommendation for process and organizational improvement. It will not only spotlight the innovative use of business process analysis tools that were utilized, but how University Libraries, due to constraints imposed by the

pandemic, had to resiliently adapt them to conduct the work in a virtual environment. Further, it will highlight the “A-ha” moments we discovered along the way, and provide an adaptable roadmap for other institutions/organizations large and small to employee in conducting their own analysis.

II. TECHNIQUES

University Libraries, through the DP&A, has engaged in utilizing five techniques—SIPOC, RACI, Brainwriting, Workflow Visualization and Change Management—to help it analyze, visualize and understand the workflows processes they engage in. These workflows encompass the acquisition of born digital materials, the digitization of existing materials, the arrangement, description and processing of those material, along with preserving and providing access to them. These techniques come from the realm of business process improvement, with roots in Total Quality Management that continue to be used in Lean and Six Sigma programs.

A. SIPOC

The SIPOC exercise provides for a very high-level view of our workflows or processes. The steps in a workflow are aggregated up to a level of abstraction—a minimum of four and a maximum of seven process steps—that still allows us to understand suppliers (S) of inputs (I) that are transformed through the processes steps (P) into outputs (O) for customers (C). The intent is to ensure that all processes are represented.

B. RACI

Following up on the SIPOC, each group was asked to conduct a RACI to determine for each step within a process who is responsible (R), accountable (A), consulted (C) or needs to be informed (I).

C. Brainwriting

Finally, we engaged in brainwriting to further tease out the granularities of the steps identified within the SIPOCs.

D. Workflow Visualization

Utilizing the workflow visualization software, we have created graphic workflow representations based upon the detail generated during brainwriting and verified against the SIPOC and RACI.

E. Change Management

We have initially employed a change management rubric in addressing one of the five initial recommendations to help us determine how to prioritize the efforts in our workflows. The rubric challenges the group to create a Problem Statement; describe the Current State and envision a Future State; Define the desired Change; what Benefits the change will provide; what the new Process is, how it will be implemented and work; and how we will Measure Progress.

III. RESULTS THUS FAR

We believe the tools and the approaches we have brought to analyze these issues are implementable in institutions large and small, with a low technological barrier. We utilized typical office productivity software and shareware applications.

We have successfully visualized twenty-five workflows, and had considered an additional thirteen that were either out of scope, not yet developed or currently suspended. We developed an initial set of recommendations based upon five key gaps and implications of those gaps: process, prioritization, process management, resources and documentation. Each of these gaps includes several actionable factors that we have categorized into two buckets: immediate impact and long-term impact.

Our ongoing more granular analysis of the visualized workflows is looking to identify not only gaps, but intersections and commonalities and their alignment within the workflows. Further, we are attempting to apply a lens of Total Cost of Stewardship [2] as we consider capacity and equitable distribution of the workload. Our efforts are not directed at dismantling silos—they do have their functional purposes. However, we want to make those silos more transparent and permeable, exposing the work we do, creating a more inviting environment for teamwork and collaboration.

REFERENCES

- [1] Digital Preservation & Access Workgroup's Wiki, <http://go.osu.edu/libraries-dpa>
- [2] Total Cost of Stewardship: Responsible Collection Building in Archives and Special Collections, OCLC. <https://www.oclc.org/research/publications/2021/oclcresearch-total-cost-of-stewardship-tools-suite.html>

LIBNOVA CONSORTIUM

A successful community project

Teo Redondo

LIBNOVA
Spain
teo.redondo@libnova.com
[0000-0001-6465-7771](tel:0000-0001-6465-7771)

Miquel Tèrmens

University of Barcelona
Spain
termens@ub.edu

Fernando Aguilar

CSIC
Spain
f.a@csic.es

David Giarretta

Giarretta Associates
UK
david@giarretta.org

Julia Thiele

Amazon Web Services
Germany
juliatt@amazon.de

Ciprian Abaseaca

Voxility
UK
ciprian.abaseaca@voxility.com

Mikel Rufian

Bidaidea
Spain
mikel.rufian@bidaidea.com

Abstract – The aim of this poster is to show a successful example of cooperation between different public and private organizations, both within and outside the digital preservation community, working together in a consortium within the ARCHIVER project, the most innovative digital preservation research and development project in Europe.

Keywords – Digital Preservation, Community, Exchange, Research, Research Data Management.

Conference Topics – Community; Exchange.

I. INTRODUCTION

The ARCHIVER Project (Archiving and Preservation for Research Environments) is the only EOSC related H2020 project focussing on commercial long-term archiving and preservation services for petabyte-scale datasets across multiple research domains and countries [1].

On 29 January 2020, the ARCHIVER project launched its Pre-Commercial Procurement Request for Tenders with the purpose to award several Framework Agreements and work orders for the provision of R&D for hybrid end-to-end archival and preservation services that meet the innovation

challenges of European Research communities, in the context of the European Open Science Cloud.

The project team (CERN-led) encouraged companies/organisations to combine their skills and resources to form viable consortia to achieve the required results [2]. In this context, and based on this recommendation, the LIBNOVA Consortium [3] was formed, and has turned out to be one of the final contractors for the ARCHIVER project [4].

II. COLLABORATION

The consortium led by LIBNOVA has been enriched throughout the project with the incorporation of new members with expertise in the specific needs of each phase, forming a multidisciplinary cooperative and collaborative team. These are the main contributions of each member of the consortium:

- **LIBNOVA** (led): is focused on the Digital Preservation field, and provides solutions to organizations, so that big volumes of valuable data are accessible during long periods of time. The company is a leading digital preservation provider with an

international presence in several countries in the heritage, cultural and research dataset areas.

- The **Spanish National Research Council (CSIC)** is the main agent of the Spanish System for Science, Technology and Innovation. Their mission is the promotion, coordination, development and dissemination of scientific and technological multidisciplinary research, in order to contribute to the progress of knowledge and economic, social and cultural development.
- The **University of Barcelona (UB)** is the principal centre of university research in Spain and has become a European benchmark for research activity, both in terms of the number of research programmes it conducts and the excellence these have achieved.
- **David Giarretta** has led many of the most important developments in digital preservation, with EU-funding and more than 50 partner organisations. He chaired the panel which produced the OAIS Reference Model (ISO 14721), the ISO standard for audit and certification of trustworthy digital repositories (ISO 16363), and ISO 16919.
- **Voxility** provides agile Infrastructure-as-a-Service (hardware and network equipment for internet access) for hosting providers, cloud service providers or integrators and software developers, among others. They provide high-capacity and high-performance infrastructure.
- **Amazon Web Services (AWS)** is the world's most comprehensive and broadly adopted cloud platform, offering over 200 fully featured services from data centers globally. Millions of customers, from the fastest-growing startups, largest enterprises, and leading government agencies, are already using AWS.
- **Bidaidea** is a consulting company focused on Cybersecurity, with end-to-end coverage in Integral Security, Security Intelligence, Self-Protection Plans, Electronic Security and Automated Security Systems. They create

and execute integral plans for 360 Security (Physical, Electronic, Logical, Cybersecurity and Intelligence) for both SMEs and large companies and institutions.

III. POSTER INTENTION

This poster is intended to serve as a use case of successful collaboration on how a digital preservation research project involving multiple stakeholders from different fields of expertise can be articulated.

REFERENCES

- [1] Archiving and Preservation for Research Environments | ARCHIVER Project | Fact Sheet | H2020 | CORDIS | European Commission <https://cordis.europa.eu/project/id/824516>
- [2] ARCHIVER launches its Pre-Commercial Procurement Tender <https://cordis.europa.eu/article/id/413444-archiver-launches-its-pre-commercial-procurement-tender>
- [3] ARCHIVER Project | Consortium 3 <https://archiver-project.eu/consortium-3>
- [4] ARCHIVER PROJECT | PILOT PHASE AWARD - THE TWO WINNERS <https://archiver-project.eu/pilot-phase-award>

PRESERVING ELECTRONIC THESES AT THE UNIVERSITY OF ST ANDREWS LIBRARIES AND MUSEUMS

Sean Rippington

*University of St Andrews
Libraries and Museums
UK
sbr2@st-andrews.ac.uk*

Janet Aucock

*University of St Andrews
Libraries and Museums
UK
ja@st-andrews.ac.uk
[0000-0001-9616-0612](tel:0000-0001-9616-0612)*

Abstract – The University of St Andrews Libraries and Museums have been preserving and making our PhD theses accessible since they were introduced in 1917. Since 2007 theses have been required to be submitted in both print and electronic form, the electronic files being used to make full text theses digitally available on our institutional repository <https://research-repository.st-andrews.ac.uk/>. There have been aspirations for progressing to ‘electronic only’ thesis deposit since 2008¹, but this was held back by a lack of digital preservation support and infrastructure. However, during the COVID pandemic many students were only able to submit an electronic copy of their thesis and print submission was temporarily suspended – forcing a temporary move to e-only thesis deposit. This move to ‘e-only’ thesis deposit became business as usual in March 2022, pushed by various stakeholders and supported by changes to governance, training, and technical solutions.

This poster explores the emerging good practice, lessons learned, and future steps for preserving electronic theses using the University of St Andrews as a case study in ensuring that these vital scholarly outputs have integrity, are accessible, and can be used by future generations.

**Keywords – Theses, Open Access, ETD, PDF
Conference Topics – Exchange, Resilience**

I. INTRODUCTION

The University of St Andrews is moving to electronic-only thesis deposit for PhD students in March 2022 in response to emerging stakeholder needs and expectations, including the streamlining of student experience, reduction in printing and binding costs for students, and the need to save on

expensive physical library storage space. While we have considerable experience in receiving electronic (mostly PDF) copies of thesis for publication online since 2007, these have been seen as dissemination copies of paper theses, rather than authoritative ‘master’ or ‘archive’ versions in their own right. Although we do make efforts to preserve these electronic dissemination versions, including surveying deposited file formats to identify and monitor preservation risks, most of our past preservation effort and resource has been directed to paper over electronic theses.

The move to electronic only thesis deposit requires a shift in mindsets (that the electronic thesis will now be the ‘master’ or ‘archive’ copy) and requires a shift in resourcing and effort towards robust digital preservation of the electronic copy. Theses are unique scholarly outputs, and it is vital to ensure that the electronic files submitted have integrity, are accessible, and can be preserved.

II. EMERGING GOOD PRACTICE

Prior to the move to electronic only theses deposit, we researched:

- Which institutions already preserve e-only theses, and how
- Who does this well
- What concerns have been noted

A. Which Institutions Already Preserve E-only Theses, And How

We could not find a definitive list of institutions already undertaking electronic-only thesis deposit, though some research via Google, professional mail lists and projects such as the Educopia ETD+ Toolkit revealed some institutions and individuals with experience in this area. Broadly speaking they took one of four approaches:

- Accept a pdf version of the thesis only
- Accept a pdf version of the thesis only, specifying a pdf/a variant
- Accept the 'original' file(s) and a pdf version of the thesis
- Accept the 'original' file(s) and a pdf version of the thesis, specifying a pdf/a variant.

The more detailed governance, workflows, systems, training and advice were usually quite specific to the individual institution and were surprisingly heterogenous.

B. Who Does This Well

It was difficult to establish who manages preservation of e-only deposited theses well for several reasons; there are no specific external benchmarks for measuring success in this process; relatively little experience in the process is shared publicly in detail; it can be difficult to assess the success of institutions in this process without access to their systems and content; relatively few institutions (only three found at the time of writing) had undertaken and shared the results of detailed reviews of their electronic theses corpus for preservation and access issues; we could not find any published data (at the time of writing) on preferred access formats for users of electronic theses.

C. What concerns have been noted

Common concerns that arose during our research into good practice included:

- What to ask students to deposit
- How to deal with increasingly common 'non-traditional' theses

formats including video, web-based and ebook formats such as EPUB.

- The value of PDF/A as an archive format
- How to ensure and audit the integrity of formats derived from 'original' files
- Who should undertake any file format changes, and what support they need
- How to encourage the creation of 'accessible' theses
- How to audit the success of e-thesis preservation processes
- How to assess and handle any inability of students to meet technical deposit criteria (e.g. in format specification, including validity and integrity of deposited files)

III. CONCLUSIONS AND LESSONS LEARNED

Following our research and internal discussions, we concluded that a successful transition to e-only thesis deposit requires significant non-technical changes, including to governance, training, culture, and potentially staff resource, with input from a range of stakeholders including relevant committees, senior academic staff, PhD supervisors, library staff, and accessibility experts.

Our systems and processes required review and development to ensure the integrity of electronic files throughout their journey from deposit, to storage, access and reuse. This review has been supported by the publication of our digital preservation policy and submission of an outline business case for additional business analysis and funding for technical solutions for digital preservation.

In terms of what we asked of depositing students, our own process settled on:

- 1) Always ask for the source or 'original' files to act as the 'master' or 'archive' versions, as requesting or automating conversion of a thesis to a new format introduces risks around the integrity and validity of the resulting file that may be difficult to mitigate and document.

2) Ask for embedded files (video, etc.) in a separate folder so that their preservation can be managed separately if needed.

3) Ask for a PDF, which will be used as the 'access' version to be published online.

3) When converting to PDF for access provide Adobe Pro, or Nitro Pro, to produce valid PDFs.

4) Provide guidance for making minor adjustments to make the pdfs more accessible in the long term e.g. Create PDF 1.6:

- turn off additional compression of images
- Change the colour management to Gray Gamma 2.2 and Adobe RGB (1998)
- Choose 'save original JPEG images in PDF if possible'
- Also choose "embed all fonts"
- Embed hyperlinks
- Stabilise hyperlinks using a web archiving service.

Note that many of the above settings will make the PDFs bigger.

5) Provide guidance for students to look for and correct common pdf migration issues, including migration errors introduced by use of non-western characters, unusual fonts, raster images, tables, and embedded multimedia.

6) Ask students submitting their e-thesis in more unusual file formats (e.g. video, ebook formats including EPUB) to discuss their submission with library staff to identify and mitigate against preservation and access issues.

7) Allow students to submit a valid PDF-A if they wish, and provide training and guidance to support this.

8) Request that any security settings on deposited files be disabled

9) As our particular workflow will involve staff inserting a coversheet into the PDF, ensure that this process does not affect the integrity or validity of the file.

Going forward, it will be vital to secure resourcing to review and act upon changes in good practice across the sector, and to respond to our own changing capacity and user needs and expectations. This cannot be a one-off project,

and should be seen as part of an evolution in a service that has existed since 1917.

We are grateful for the support and advice we have received from various related communities of practice, including the EThOS community, DPC members, SCURL Repository Shared Services Group, and other universities in Scotland. Greater communication between institutions undertaking electronic-only theses deposit, and perhaps some funding to undertake an evidence-based review of their processes, would also be welcome. Any lessons learned across the sector are likely to be useful in other digital preservation contexts, especially those involving deposit of content by third-parties and accessibility of scholarly communications.

REFERENCES

- [1] Aucock, J. (2008). Electronic theses at the University of St Andrews: institutional infrastructure, policy and support to establish an electronic theses service. Research-Repository.st-andrews.ac.uk; University of St Andrews. <http://hdl.handle.net/10023/513>

PRESERVING PHOTOGRAMMETRY OUTPUTS

A case study at the University of St Andrews Libraries and Museums

Sean Rippington

University of St Andrews
UK
sbr2@st-andrews.ac.uk

Photogrammetry – taking overlapping photographs of an object and converting them into 3D digital models – is increasingly popular as a technique for recording, analysing, and providing digital access to heritage collections.

This poster explores steps taken at the University of St Andrews Libraries and Museums to preserve the outputs of our photogrammetry activity, including research in to emerging good practice, lessons learned, and potential future steps.

**Keywords – Photogrammetry, 3D, Heritage,
Conference Topics – Community, Exchange**

I. INTRODUCTION

Photogrammetry has become business as usual at the University of St Andrews Libraries and Museums. Driven by the push to provide more, different, and better types of online access to our collections for teaching and learning during the pandemic, we now have over 170 examples on our IIF-based Collections site. Many of the models have already been used in our Exhibit teaching and engagement tool, developed to provide new, curated, almost tactile encounters with digital objects.

Photogrammetry (and other 3D recording and visualization techniques) has developed rapidly over the past two decades as the related technology improves and other costs and barriers to entry fall. It is an increasingly mainstream technique across the heritage sector, and in other areas including architecture, engineering, surveying, medicine, entertainment and private recreation.

II. EMERGING GOOD PRACTICE

While photogrammetry techniques have existed for some time, good practice in preservation of photogrammetry outputs is an emerging topic. Authoritative guidance includes Digital Preservation Coalition event recordings¹, blog posts², and Technology Watch Reports³. There is also a professional community growing around the US-based Community Standards for 3D Data Preservation (CS3DP)⁴ project, which includes working groups, published resources and events.

III. LESSONS LEARNED

After engaging with the emerging best practice, we settled on the following principles to guide our approach to photogrammetry preservation. They are largely informed by a preprint from the forthcoming 3D Data Creation to Curation: Community Standards for 3D Data Preservation.⁵

- 1) Keep tiffs of the source object (additionally keeping raw files might be desirable but has significant storage implications). Tiffs should be 8 bit and LZW compressed to reduce storage requirements while maintaining acceptable quality.
- 2) Keep point cloud data, ideally as a text file, or in some other open format. We may need to have a text file explaining how these relate to the tiffs, as file locations in the metadata may no longer work.
- 3) Keep a record of any control points system documenting the model's relationship to real-world measurements, ideally in some open format such as a text file.
- 4) Keep the .obj file of the 3d object – other formats may be suitable, but we already produce this in our

workflow and it is listed as an 'acceptable' file format by the Library of Congress.⁶ Note that this may require preserving a separate image texture file.

5) Generate and keep a project report (in our case, a pdf generated by Metashape), documenting some of the technical settings and fully quantifying any errors.

6) Document hardware and software used, and workflow. This could be saved as a text file. Note that workflow is iterative and may change from object to object.

7) Ideally all this information will be saved with the item record in our repository.

8) .glb files will continue to be used as the access version of the object as they are small, quick to load, and work in the IIIF viewer

IV. CONCLUSIONS AND POTENTIAL FUTURE STEPS

It is anticipated that our approach to preserving photogrammetry will need periodic review as good practice in the sector changes, and our own needs and capabilities develop.

A. Changes in Good Practice

To understand and contribute to changes in good practice we will need to better engage with the relevant communities, and not just those in the heritage sector – this poster is a first step. The Digital Preservation Coalition and Community Standards for 3D Data Preservation (CS3DP) project provide a framework for ongoing engagement. We note that the best practice community for 3D preservation is largely US-based and heritage-focussed – we and others need to do more to make sure that other experiences and interests are represented in these discussions

B. Changes in Our Own Needs and Capabilities

The obvious downside to your current approach is that we are keeping rather a lot, which has consequent costs and environmental impacts. However, photogrammetry technology is constantly improving, and we're anticipating that the emerging ability to re-render 3d objects from our existing data in new and better ways will outweigh the potential costs of having to rescan the objects in the future. We will need to keep this under review and do more work to weigh the costs of preservation of photogrammetry outputs against the benefits.

We will also need to gather data over time about how our stakeholders are using our photogrammetry outputs. Ultimately these will define what we need to keep, and what we do not, as well as what data we make available and how.

REFERENCES

- [1] Building a digital future : Challenges & Solutions for preserving 3D models. Digital Preservation Coalition. (n.d.). Retrieved February 25, 2022, from <https://www.dpconline.org/events/past-events/preserving-3d-digital-engineering-models-a-briefing-day>
- [2] 3D - Digital Preservation Coalition. (2022). Retrieved 25 February 2022, from <https://www.dpconline.org/digipres/tags/3d>
- [3] Preserving 3D Data Types Series Artefactual Systems and the Digital Preservation Coalition DPC Technology Watch Guidance Note. (2021). <https://doi.org/10.7207/twgn21-14>
- [4] Community standards for 3D Data preservation. CS3DP. (n.d.). Retrieved February 25, 2022, from <https://cs3dp.org/>
- [5] Metadata Requirements for 3D Data (Blundell, Jon, Clark, Jasmine L., DeVet, Katherine E., and Hardesty, Juliet L. 2020) – a preprint from the forthcoming 3D Data Creation to Curation: Community Standards for 3D Data Preservation (Moore, Jennifer, Adam Rountrey, and Hannah Scates Kettler. 2022. Association of College & Research Libraries.
- [6] Recommended formats statement. Recommended Formats Statement - Design and 3D | Resources (Preservation, Library of Congress). (n.d.). Retrieved February 25, 2022, from <https://www.loc.gov/preservation/resources/rfs/design3D.html>

QUALITY ASSURANCE FOR BORN-DIGITAL INTERACTIVE NARRATIVES

The New Media Writing Prize Collection as a case study

Giulia Carla Rossi

The British Library

UK

giulia.rossi@bl.uk

[0000-0002-6645-9987](tel:0000-0002-6645-9987)

Tegan Pyke

Cardiff Metropolitan University

UK

tpyke@cardiffmet.ac.uk

[0000-0002-1276-5217](tel:0000-0002-1276-5217)

Abstract – The UK Legal Deposit Libraries have been researching and building experimental collections of emerging formats for the past five years, including curated collections of web-based interactive narratives in the UK Web Archive. The New Media Writing Prize Collection is one of such collections, created using web archiving tools to capture instances of the online interactive works that were shortlisted or won the Prize since its launch. This poster briefly outlines the collection, and focuses on the quality assurance criteria adopted to assess the quality of the captures. These were the result of a short PhD Placement at the British Library and they expand on technical criteria to include considerations on the narrative and literary quality of digital interactive publications.

Keywords – Quality assurance, emerging formats, digital interactive narratives, New Media Writing Prize, web archiving

Conference Topics – Resilience; Innovation.

I. INTRODUCTION

The British Library, together with the other five UK Legal Deposit Libraries, has been collecting born-digital formats under Non-Print Legal Deposit Regulations since these came into force in 2013 [1]. Publications in scope include e-books and e-journals, as well as websites identified as published or hosted in the UK [2]. Some of these publications are also what the Legal Deposit Libraries have named 'emerging formats': born-digital formats with complex structure and technical dependencies that are usually not addressed by standard collection management methodologies. These formats have no print counterpart and strong software and hardware dependencies that make them especially vulnerable to the risks of rapid obsolescence [3].

One of the emerging formats that the Legal Deposit Libraries chose to prioritize in their research is web-based interactive narratives. These are non-linear interactive stories, delivered via a browser, that require some form of reader's input in order to determine how the narrative will unfold. Research conducted by Lynda Clark [4] during her Innovation Placement at the British Library identified different tools, platforms and interaction patterns for the creation of these publications, as well as a variety of content and genres. Focusing on web-based formats meant that the Libraries could rely on the already established workflows and tools of the UK Web Archive to experiment with capturing examples of interactive narratives.

A first collection [5] was launched in 2019 using a combination of different web archiving tools, followed two years later by the New Media Writing Prize Collection.

II. THE NEW MEDIA WRITING PRIZE COLLECTION

The New Media Writing Prize Collection (NMWP Collection) [6] is the result of a collaborative project with Bournemouth University to collect copies of all shortlisted and winning entries to the New Media Writing Prize since its launch in 2010. The Prize is an annual celebration of innovative and interactive literary works created using digital tools, with no limits on interactive format types or nationality of participating authors [7]. The variety of formats represented, as well as the diverse background of the shortlisted authors and winners made for a significant collection of contemporary digital literature. While only web-based works could be collected using web archiving tools, the collection

proved to be the perfect case study to research quality assurance measures for digital interactive narratives.

A. Quality Assurance Considerations

As part of a 2021 PhD Placement with the British Library, researcher Tegan Pyke carried out quality assurance of the current NMWP Collection [8]. During an initial review of the Collection, it became evident that 'complete' was not a possibility for many of the pieces, as common digital-born writing practices like external hosting and interconnectivity create issues in isolated archival environments. Amira Hanafi's shortlisted 2014 entry to the New Media Writing Prize, *What I'm Wearing* [9], for example, is created by weaving contradictory quotes on women's clothing together, with each quote hyperlinked to its source article. When placed in the isolation of the UK Web Archive, the piece suffers a total loss of interactivity with connections to external addresses no longer possible [10].

These instances are further exacerbated by new media's multimodal nature, where multiple forms of communication are used in conjunction [11]. This means retention of only the primary, linguistic mode of communication leads to incomplete and, often, unfinishable work.

B. New Criteria

Taking these issues into consideration, a set of quality assurance criteria were established based around the literary elements identified in all New Media Writing Prize entries, regardless of genre, format, or platform. These were:

1. Narrative

The narratives of new media span many communicative modes, all of which must be retained for clarity. A work was counted as narratively complete when the central storyline could be followed from start to finish.

2. Theme

Spanning multiple modes as narrative does, the theme had to remain understandable and present throughout a capture for a work to pass thematically.

3. Atmosphere

Digital writing follows the concept of Text as Game [12], where atmosphere is established via incorporation of assets, page arrangement, and relational networks. For a work to pass, the intended atmosphere had to be retained.

C. Results

Out of the 76 works captured by the UK Web Archive for the New Media Writing Prize collection, 33 had passing instances according to the new criteria. Out of the 43 works that failed, 20 were due to narrative incompleteness and one due to loss of atmosphere. Out of the others, nine were affected by media player obsolescence and 13 by playback errors caused by the UK Web Archive's software platform, the World Wide Web Annotation and Curation Tool.

A total of 16 of the failing captures can be improved by recrawls targeting assets and deeplinks by curators of the NMWP Collection.

III. POSTER LAYOUT

The graphical poster covering this work will be in an A2 print-out, as well as an interactive digital format. It will illustrate the quality assurance workflow based on the above criteria in the form of a nonlinear hypertext narrative, mirroring the structure of the very same publications it seeks to preserve. This poster aligns with the call for papers by looking at preservation approaches for new and emerging media and by adopting holistic quality assurance methods that include literary elements alongside technical considerations.

REFERENCES

- [1] legislation.gov.uk, "Legal Deposit Libraries (Non-Print Works) Regulations 2013. S.I.2013 No.777", legislation.gov.uk <https://www.legislation.gov.uk/uksi/2013/777/made> (accessed Mar. 01, 2022).
- [2] The British Library, "Legal deposit and web archiving," The British Library - Legal Deposit, <https://www.bl.uk/legal-deposit/web-archiving> (accessed Mar. 01, 2022).
- [3] C. Smith and I. Cooke, "Emerging Formats: Complex Digital Media and Its Impact on the UK Legal Deposit Libraries", *Alexandria*, vol. 27, no. 3, pp. 175-187, Dec. 2017. doi: [10.1177/0955749018775878](https://doi.org/10.1177/0955749018775878).
- [4] L. Clark, G.C. Rossi, S. Wisdom, "Archiving Interactive Narratives at the British Library", in *Interactive Storytelling. ICIDS 2020* (Lecture Notes in Computer Science, vol 12497), AG. Bosser, D.E. Millard, C. Hargood, Eds., Springer, Cham, 2020. doi: [10.1007/978-3-030-62516-0_27](https://doi.org/10.1007/978-3-030-62516-0_27)
- [5] UK Web Archive, "Interactive Narratives", UK Web Archive, <https://www.webarchive.org.uk/en/ukwa/collection/1836> (accessed Mar. 01, 2022).
- [6] UK Web Archive, "New Media Writing Prize", UK Web Archive, <https://www.webarchive.org.uk/en/ukwa/collection/2912> (accessed Mar. 01, 2022).
- [7] New Media Writing Prize, "FAQs", New Media Writing Prize, <https://newmediawritingprize.co.uk/faqs/> (accessed Mar. 01, 2022).

- [8] T. Pyke, "Quality assurance in the new media writing prize collection", British Library, London, UK. 2021. DOI: <https://doi.org/10.23636/1y1j-by18>
- [9] A. Hanafi, "What I'm wearing", amira hanafi, <http://whatimwearing.amiraha.com/> (accessed Mar. 02, 2022).
- [10] UK Web Archive, "What I'm wearing", UK Web Archive, <https://www.webarchive.org.uk/wayback/archive/20200515090252/http://whatimwearing.amiraha.com/> (accessed Mar. 02, 2022)
- [11] K. L. Arola, J. Sheppard, and C. E. Ball, *Writer/Designer: A Guide to Making Multimodal Projects*, New York, USA: Bedford/St. Martin's, 2010, pp. 3-13.
- [12] M. L. Ryan, *Narrative as Virtual Reality*, Baltimore, USA: John Hopkins University Press, 2001, pp. 191-199.

DIGITAL PRESERVATION CAPABILITIES OF THE BURSA ULUDAG UNIVERSITY

Survey in the Light of Digital Preservation Coalition Rapid Assessment Model

Dr. Özhan Sağlık

Bursa Uludag University
Türkiye
ozhansaglik@uludag.edu.tr /
ozhan.saglik@gmail.com
[0000-0002-1436-7431](tel:0000-0002-1436-7431)

Abstract - While performing their primary functions as research and education, universities also carry out in the ordinary course of activities such as naturally occurring personnel employment, procurement and promotion. As a result of these activities, materials with different types are created. These materials, which are created electronically or digitized, are the memory of the universities as well as evidence of their activities. Therefore, these materials need to be preserved for a long time. To predict the success of these materials in terms of long-term preservation, it is necessary to examine the capabilities of universities in this field. While doing this review, is thought to benefit from the Rapid Assessment Model (RAM) created by Digital Preservation Coalition (DPC). Because the RAM is designed to evaluate the digital preservation capability of organizations at a basic level. As a result of this evaluation, it will be possible for universities to improve their digital preservation capabilities and monitor their progress. In this study, in which quantitative research methods will be used, Bursa Uludag University, one of the universities with the highest number of students in Turkey, is the sample. The study aims to contribute to raising awareness in universities about how DPC RAM can be used in the evaluation of digital preservation practices in universities.

Keywords - Digital preservation in universities, Bursa Uludag University, DPC RAM
Conference Topics - Resilience

Universities, while performing their primary functions as research and education, also carry out in the ordinary course of activities such as naturally occurring personnel employment, procurement and promotion. As a result of these activities, materials with different types are created. These materials, which are created electronically or digitized, are the memory of the universities as well as evidence of their activities. Therefore, these materials need to be preserved for a long time. In the circumstances, questions such as how these materials will be preserved and how preservation practices will be evaluated come to mind. To analyze these questions, a tool is needed to examine the capabilities of universities in this field [1]. Thus, it was thought to benefit from the RAM created by DPC [2].

DPC RAM is defined as a maturity modelling tool that has been designed to enable a rapid benchmarking of an organization's digital preservation capability [3, 4]. "The model provides a set of organizational and service level capabilities that are rated on a simple and consistent set of maturity levels". Thereby, "it will enable organizations to monitor their progress as they develop and improve their preservation capability and infrastructure and to set future maturity goals" [2].

The problem of the study is determined as "there is a lack of a model in monitoring the long-term preservation practices of digital materials created in

I. INTRODUCTION

universities". The question of the study is, "in the light of the analysis criteria in DPC RAM, at what level are the digital preservation capabilities of the universities". The hypothesis can be stated that "when the digital preservation capabilities of universities cannot be monitored with a consistent model, sufficient success in preservation may not be achieved". In this study, in which quantitative research methods will be used, Bursa Uludag University, one of the universities with the highest number of students in Turkey, is the sample. The dependent variable is the digital preservation capabilities of the universities, and the independent variable is the analysis criteria of DPC RAM in the study. Survey design is adopted in the research, and the attitudes of Bursa Uludag University on digital preservation capabilities will be examined. The cross-sectional research type is used as the data will be collected once. To analyze the results obtained through face-to-face interviews, the visualization tool in DPC RAM will be used. The researcher will not be in a guiding position when participants answer the questions. Since DPC formed the questions, no additional evaluation regarding validity and reliability will be done in the study. The study aims to contribute to raising awareness in universities about how DPC RAM can be used in the evaluation of digital preservation practices in Turkish universities.

II. SURVEY

Bursa Uludag University is one of the universities with the highest number of students in Turkey [5]. Due to this feature, there are many materials to be preserved for a long time. The materials which have archival value are kept in the Head of Library and Documentation. Therefore, DPC RAM will be analyzed by the manager of this unit. The situation resulting from the university's answers can be stated as follows in Fig. 1.

There are many different types of digital materials at Bursa Uludag University. Those with the archival value among these materials are transferred to the Head of Library and Documentation. After that, the preservation of materials is the responsibility of this unit. Although the Head of Library and Documentation has established systems and developed procedures to protect these materials, an assessment of its capabilities has not yet been made. Making this assessment will also help the institution perform successful digital

preservation practices. It has been seen that DPC RAM can be used for this purpose.

III. CONCLUSION

As a result of the study, the hypothesis "when the digital preservation capabilities of universities cannot be monitored with a consistent model, sufficient success in preservation may not be achieved" has been confirmed. The RAM introduced by DPC is useful in determining the current state; but some improvements are needed on the target level. Because when a target is set, how it is achieved should be revealed with concrete criteria. It may be possible to encounter subjective evaluations of the person or unit performing the DPC RAM analysis. As a result, it may be possible that the digital preservation capabilities are not adequately reflected. As a solution to this, the issues determined at the levels in DPC RAM can be made a criterion. It is thought that new research to be carried out in this direction will be useful. Nevertheless, DPC RAM provides a successful assessment of the current digital preservation practices of organizations.

REFERENCES

- [1] Maemura, E., Moles, N. and Becker, C. (2017), Organizational assessment frameworks for digital preservation: A literature review and mapping. *Journal of the Association for Information Science and Technology*, 68: 1619-1637. <https://doi.org/10.1002/asi.23807>
- [2] Digital Preservation Coalition, Rapid Assessment Model. <https://www.dpconline.org/docs/miscellaneous/our-work/dpc-ram/2433-digital-preservation-coalition-rapid-assessment-model-v2/file>
- [3] Dafter, H. (2021). "The Postal Museum's Case Study of the DPC Rapid Assessment Model". <https://www.dpconline.org/blog/postal-museum-ram-case-study>
- [4] Barticioti, F. (2021). "Assessing where we are with Digital Preservation". <https://www.dpconline.org/blog/wdpd/assessing-where-we-are>
- [5] Yüksek Öğretim Kurulu, İstatistikler. <https://istatistik.yok.gov.tr>

Digital Preservation Coalition Rapid Assessment Model (DPC RAM): Bursa Uludag University

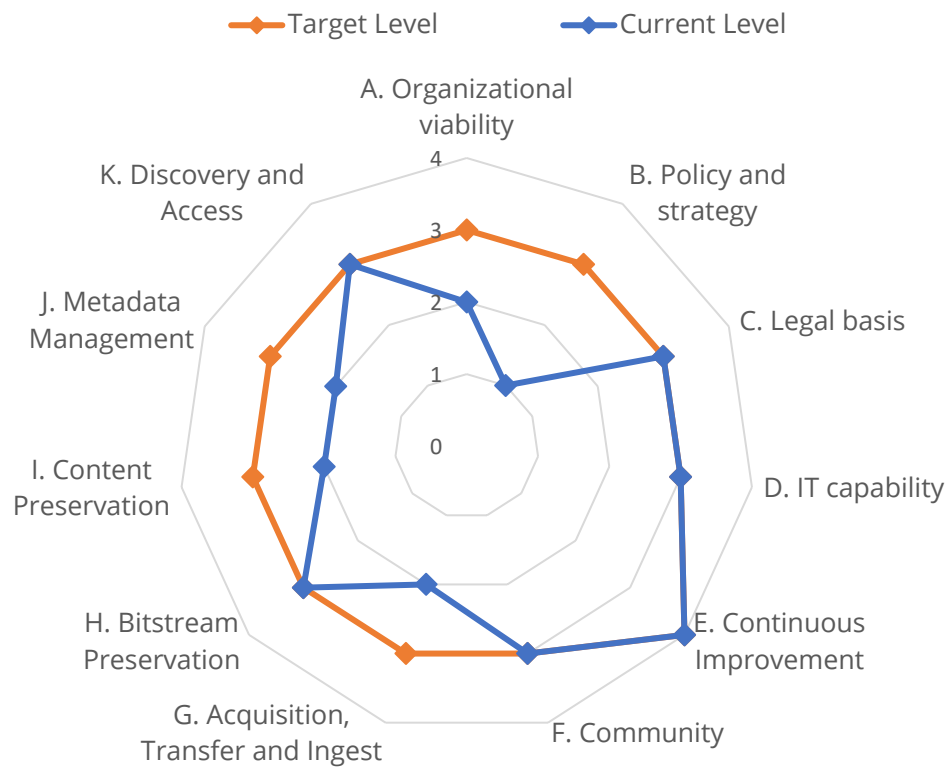


Figure 1 Bursa Uludag University DPC RAM Analysis

CREATING WORKFLOWS TO SCALE OUT OPEN ACCESS E-BOOK ACQUISITIONS AT THE LIBRARY OF CONGRESS

**Andrew Cassidy-
Amstutz**

*Library of Congress
United States of America
acas@loc.gov*

Kristy Darby

*Library of Congress
United States of America
kdar@loc.gov*

Elizabeth Holdzkom

*Library of Congress
United States of America
eholdzkom@loc.gov*

Camille Salas

*Library of Congress
United States of America
csala@loc.gov*

Lauren Seroka

*Library of Congress
United States of America
lseroka@loc.gov*

Abstract - This poster illustrates the origins, exploration, and routinization of Open Access (OA) e-book acquisition, processing, and preservation at the Library of Congress (LC) led by the Digital Content Management section (DCM). This includes discussion of technical workflows, special issues and opportunities presented by expanded telework during the COVID-19 pandemic, ongoing challenges in expanding and routinizing OA e-book collecting, and iterative process development.

Keywords - Open Access, E-books, Metadata, Acquisitions

Conference Topics - Resilience

In January 2017, the Library of Congress adopted a set of strategic steps related to its future acquisition of digital content entitled Collecting Digital Content at the Library of Congress. The first strategic objective in this plan is "expand and routinize acquisition and access of openly licensed and openly available digital works," identifying OA e-books as the target for year one. The Digital Content Management Section (DCM) engaged fully in this effort, collaborating with divisions throughout the Library and working with multiple acquisition streams to review and analyze the large amount of e-book content coming into the Library.

DCM led several pilot projects to test technical methods for obtaining e-books from various sources, transforming descriptive metadata, and processing the content for presentation on the Library's website, loc.gov. The first pilot explored the level of effort required to process ten OA e-books identified in the

Directory of OA Books (DOAB) for which the Library already had print holdings. DOAB is a community-driven platform that provides data and e-book files, when available, for nearly 60,000 peer-reviewed OA e-books from more than 600 publishers. This resource offered a unique opportunity to systematically identify OA monographs that can be added to the Library's collection. Acquiring the e-book files, putting them in managed, long-term preservation storage, and providing access on loc.gov rather than linking to the files on the open web required experimentation and iteration.

DCM staff took the lessons and workflows from this pilot of ten books and embarked upon another pilot to gauge how well the work would scale, this time working 100 titles through the process. Staff then began regularly processing DOAB e-books for which the Library already had print holdings.

This work required the creation, refinement, and evolution of workflows dealing with both metadata and digital content. DCM staff collaborated with staff from the Library of Congress Integrated Library Systems Program Office (ILSPO) to transform existing MARC bibliographic records for print books to correctly-formed bibliographic records for the corresponding e-book in batch using the MarcEdit software suite. Each e-book MARC record included the Creative Commons license information applied to the title in the 540 field, which was a new practice at the Library. Over time, DCM staff have developed and continue to develop Python scripts to pull and

analyze MARC data as well as to create and enhance MARC bibliographic records in bulk. Processing the e-book content for online presentation required new workflows as well. The infrastructure of loc.gov requires thumbnail images be created and preserved alongside the e-book file for presentation. DCM staff established workflows to manage these files and generate derivatives using Bash and Python.

The COVID-19 pandemic prompted the Library to pivot to telework in March 2020. DCM and the Collection Development Office (CDO) developed a pilot project to enable Recommending Officers, the Library's subject matter experts, to select electronic books from DOAB remotely for inclusion in the Library's permanent collection. The project resulted in the development of a functional end-to-end process allowing titles from DOAB to be identified according to LC subject areas, selected, described, preserved, and made available on the Library's public website. From the beginning of the pandemic in March 2020 until September 2021, the Open Access Books collection on loc.gov grew from approximately 300 titles to over 3500 titles. The Library's shift to telework and the resilience and flexibility of staff across the organization created a unique opportunity to grow the Library's collection of OA e-books in spite of the challenges posed by the pandemic. This project helped raise awareness of the ongoing engagement with OA monographs at the Library as well as the possibilities of expanding and routinizing the work.

While DCM staff worked on developing and applying the workflows to process and make the e-books available to users of loc.gov, staff from many service units at the Library came together to contribute to the ongoing success of this program. The success of the telework project was only possible because of the strength of collaboration. The OA e-books endeavor thrives because of the partnerships between and among DCM, the CDO, ILSPO, General and International Collections Directorate, Special Collections Directorate, Office of General Counsel, Acquisitions and Bibliographic Access Directorate, and the Office of the Chief Information Officer. All of these entities continue to dedicate resources, time, and expertise to build and support the Library's OA e-book program.

The work of routinizing, acquiring, processing, and making available OA e-books at the Library of Congress is not without challenges. These challenges

arise in the technical work required to manage e-books and make them available for use as well as analyzing and repurposing descriptive metadata, created both at the Library as well as supplied by aggregators and publishers. The quality of files and metadata varies based on supplier; identifying, isolating, and determining the best ways to work with the corrupt, incorrect, or incomplete data is often time- and resource-consuming. As each OA e-book has an OA license which is reflected in the MARC bibliographic record, ensuring that licenses match the appropriate e-book is critical and sometimes difficult. Finally, a large collaborative project involving dozens of engaged staff members from many divisions and directorates requires transparency, flexibility, extensive documentation, and careful and clear communication at every point.

DCM launched a phased initiative in October 2021 to review the pilots' processes with the goal of creating mature and routinized workflows. The result was increased automation and scalability of the workflows through the development of various Python scripts. With improved efficiency and increased output, the Open Access Books Collection nearly doubled to 5600 titles as of August 2022. OA e-book processing is now routine at LC and we will repeat the workflow in the winter of 2022 and expect to continue to grow the collection on an annual basis.

BEHIND THE SCENES

3 decades of digital preservation

Barbara Sierman

DigitalPreservation.nl

Netherlands

bsierman@DigitalPreservation.nl

[0000-0002-8190-3409](tel:0000-0002-8190-3409)

Abstract – Digital preservation is often described as being not just a technical topic, but as something developed by people. But do we keep record of who contributed to what in the past decades? Not knowing “where we came from” introduces the risk of reinventing wheels, making mistakes and ignoring important results. Now is the right time to involve the community and to prepare a publication of 3 decades of digital preservation.

**Keywords – Digital preservation history
Conference Topics – Community**

I. INTRODUCTION

Digital preservation is often described as being not just a technical topic. People are at the heart of digital preservation and people play an important role in meeting the demands of ever-changing technologies. In the past 30 years, individuals used their networks, organizations and intellectual capital to contribute to the development of digital preservation. They worked behind the scenes and contributed to improvements that are still in use today.

With support from the community, I would like to bring those individuals into the spotlights with a book about the history of digital preservation.

II. PIONEERS IN DIGITAL PRESERVATION

Several digital preservation pioneers of the first hour are currently retiring. Some of them have already passed away. Now that I’m retired myself and looking back at my career, I realize that I was part of an important era in the development of digital preservation.

In my opinion, it is important to know “where we came from” and to investigate and document the history of digital preservation. Which initiatives were started, which ones led to success and which ones failed? How did people and organizations cooperate? Who were the frontrunners? How did people communicate? Which ideas changed over time?

Currently a large number of people are involved in digital preservation and the group of practitioners is certainly larger than 25 years ago. But what do they know about their predecessors?

Luckily there are still many people around who have memories they want to share. I would like to contact them to create with their help an overview of these crucial decades in digital preservation. It is my intention to pay credit to the influential people behind the scenes and to make them and their contributions more visible.

During this process I also hope to get more insight into the networks that initiated important collaborations. Who were involved in starting Initiatives like the DPC, OPF, NCDD/NDE and nestor? In developing products like PRONOM or audit standards like CTS and ISO 16363? Which role played the people of the European Commission?

III. RE-USE OF IDEAS AND FINDINGS

In my farewell speech, when I retired from the KB, I mentioned that digital preservationists should be more aware of insights and products that were already developed.[1] “Re-use” them for further development. Either by refreshing old ideas which may only now have become relevant or by understanding why certain decisions were made or why and how practices have been established. To

know about the people that created the tools and organizations that are available today.

To be aware that in adjacent domains similar insights were developed. To know the outputs of crucial European projects and to build upon them and make better use of insights of the past. Reflections on the past might help shaping the future.

IV. BOUNDARIES OF THE SUBJECT

As digital preservation is an international domain, it will not be possible for me to sketch all the initiatives and developments around the globe. Given my experience in the library and archival world, those domains will be my main focus. As a European, the developments in Europe and especially the contributions of the European Commission will be another focus area. Around 50 projects were co-financed by them between 1995-2020, with hundreds of participating organizations and even more participants, (net)working together.

When relevant, I will include activities in other continents like Australia, Asia, Canada, and the US.

V. SOURCES OF INFORMATION

Existing literature and the Internet will of course be my first resource of information. But not everything can be found on the Internet. Especially in the early days many important documents were just printed and sent around by (international) post services. So, I'm very interested in printed material as well. Please, don't throw those documents away when a digital preservation colleague retires: I might be interested! And of course, I would like to involve the digital preservation community in various ways. By distributing a survey every now and then or by having interviews with key players. And by sharing progress updates and learnings, starting with regular blog posts on <https://digitalpreservation.nl> where I invite you herewith to comment and contribute.

I intend to have my findings published in 2026, which seems to be a nice target, 30 years after the appearance of the report Preserving Digital Information of the Taskforce on Archiving of Digital Information.[2]

REFERENCES

[1] <https://digitalpreservation.nl/seeds/recordings-seminar-re-use/>

[2] <https://www.clir.org/pubs/reports/pub63/>

CONCEPT MODEL FOR DEVELOPMENT OF PRESERVATION PLANS

Asbjørn Skødt

Danish National Archives

Denmark

assk@sa.dk

Abstract – The Concept Model enables your archive to develop preservation plans for content types (a grouping of data created for the same purposes) and related file formats in a documented approach by using a set of methods.

Keywords – File formats, analysis, methodology, stakeholders, preservation planning

Conference Topics – Innovation.

I. OVERVIEW OF THE METHODS

The Concept Model provides several steps to complete. They are:

1) Screening and prioritization of content type: Screening a content type serves the purpose of determining what should happen when your archive, through various sources, come across problematic data, where it is presumed the data cannot be immediately converted to one or more of your existing preservation formats.

2) Pre-analysis of content type: This step deals with the collation of knowledge concerning the content type under investigation. The collated information are used in any following steps.

3) Migration Assessment: It is the results of this assessment, which determines if the investigated content type can be migrated to one or more of the existing preservation formats with an acceptable loss of quality measured in loss of significant properties. The assessment involves mapping of significant properties and interview of stakeholders.

4) Format Assessment: The step applies a set of criteria in a matrix to enable your archive to score relevant file formats in a quantified manner and finally create a sum to compare the suitability of select file formats. The results of the method allows you to narrow down the number of relevant file

formats. This should be summarized in a recommendation.

5) Testing of data and software: The purpose of this step is to test existing software and potentially prototype new software for identification, characterization, conversion and validation of file formats based on custom-made or real-world data samples.

6) Consequence Assessment: The method measures and compares factors such as time, money and quality on a number of different parameters highlighted in previous steps in order to quantify the pros and cons of the most suitable preservation formats from the Format Assessment.

7) and finally the drafting of a Preservation Plan for your archive's management: The last step in the concept model is the drafting of a preservation plan for the investigated content type and specific preservation plans for the file format. The purpose is to create a documented foundation for your archive's preservation of digitally created data. Preservation plans document metadata, identifiers, risks, validation requirements, preservation actions/solutions such as migration paths, preservation level, monitoring and collate the assessments from the previous steps.

II. HOW TO USE

To support use, we provide guides and templates for each step. We present the Concept Model in an English translation and we kindly ask you to bear this in mind when reading. We created the English translation to support international dissemination of the model [1].

The purpose of presenting the Concept Model is to publicize our internal methods and contribute to the field of digital preservation. Furthermore, we seek to enable discussions on the methods presented in this poster abstract, and encourage you to raise any issues for us to consider.

REFERENCES

- [1] Concept Model for Development of Preservation Plans, GitHub. <https://github.com/the-danish-national-archives/concept-model>

WHAT DOES DATA LOSS REALLY COST?

A lot more than you think

Paul Stokes

Jisc
UK

Paul.stokes@jisc.ac.uk
[0000-0002-7333-4998](tel:0000-0002-7333-4998)

Tamsin Burland

Jisc
UK

Tamsin.burland@jisc.ac.uk
[0000-0002-5129-979X](tel:0000-0002-5129-979X)

Sarah Middleton

Digital Preservation Coalition
UK

sarah.middleton@dpconline.org
[0000-0002-7671-403X](tel:0000-0002-7671-403X)

Abstract – Jisc and the Digital Preservation Coalition (DPC) have undertaken an anonymous surveying exercise in order to unearth the true cost of catastrophic data loss—not only in terms of the value of the data, but also the cost of the knock-on effects that may only become apparent some considerable time after the event. This poster is intended to present the findings from that survey and introduce a final report which will help organisations make a stronger case for robust and effective digital preservation practice.

Keywords – Sustainability, Cost, Value, Risk, Data loss.

Conference Topics – Resilience, Exchange.

I. INTRODUCTION

Digital Preservation is about mitigating risk. Mitigations cost money. It is hard to justify spending that money without a firm grasp of the magnitude of the sums of money involved (the value of what's at risk) and the likelihood of loss.

There is already a growing body of work relating to quantifying the likelihood of loss occurring (for example, The Digital Archiving Graphical Risk Assessment Model [DiAGRAM] from the National Archives [1]). Unfortunately, it's not so easy to value the data at risk. The knock-on effects of data loss (reputation loss for instance) are even more challenging to quantify in monetary terms. Often the sums involved only become apparent a long time *after* a disaster has happened.

We know that destructive data disasters have already happened (inevitably one might argue). Some recent headlines illustrate this:

- *Server crash takes out rich digital archive at Memorial University* [2]

- *Victoria University of Wellington accidentally nukes files on all desktop PCs* [3]
- *PASIG 2017: "Sharing my loss to protect your data" University of the Balearic Islands* [4]
- *University loses 77TB of research data due to backup error* [5]

This means that there is (potentially) data extant that would give an insight into the problem. Such data, if suitably anonymised/redacted, could form the foundation of a "Cost of failure" publication showing how devastating the impact and cost of real-world data loss can be. Regrettably (and quite understandably), those who have suffered this type of loss are rarely willing to acknowledge the fact, let alone talk about the numbers involved. So we have little insight into the true extent of their losses.

With this in mind, the Digital Preservation Coalition (DPC) and Jisc set out to provide a means whereby individuals and organisations could with confidence and anonymously provide information about the extent and cost of any significant data loss events to a partnership of two trusted organisations (namely Jisc and the DPC). The intention is also to provide a mechanism to collect lessons learned and mitigation strategies.

The aim is to collect examples, from both the UK and overseas, from a range of sectors to represent the customer / membership bases of both Jisc and DPC—for example Higher Education, Research, Public Sector, GLAM, private sector—in order to highlight why organisations should invest in digital preservation. This will help them make a sustainable business case with credible exemplar data.

To achieve this end, an anonymous survey was created using the Jisc On-Line Survey Tool [6] and published in February 2022. Both Jisc and the DPC

publicised the survey in the following weeks/months. At the time of writing the survey is still open and collecting data. It is intended that it will close at the end of April. The survey results are to be used to create a publication for launch at iPres 2022.

II. THE POSTER

This poster is intended present the methodology used, the anonymous aggregated findings and to highlight key headline findings from the survey. It is also intended to introduce the final report.

REFERENCES

- [1] The Digital Archiving Graphical Risk Assessment Model (DiAGRAM) from the National Archives—
<https://nationalarchives.shinyapps.io/DiAGRAM-dev/>
- [2] Server crash takes out rich digital archive at Memorial University—
<https://www.cbc.ca/news/canada/newfoundland-labrador/mun-digital-archives-wiped-out-1.4787960>
- [3] Victoria University of Wellington accidentally nukes files on all desktop PCs—
<https://arstechnica.com/gadgets/2021/03/university-of-wellington-accidentally-deletes-files-on-all-desktop-pcs/>
- [4] PASIG 2017: “Sharing my loss to protect your data” University of the Balearic Islands—
<https://blogs.bodleian.ox.ac.uk/archivesandmanuscripts/2017/09/27/pasig-2017-sharing-my-loss-to-protect-your-data-eduardo-del-valle-university-of-the-balearic-islands/>
- [5] University loses 77TB of research data due to backup error—
<https://www.bleepingcomputer.com/news/security/university-loses-77tb-of-research-data-due-to-backup-error/>
- [6] Cost of Data Loss Survey—
<https://jisc.onlinesurveys.ac.uk/cost-of-data-loss>

THE CO₂ EMISSIONS OF STORAGE AND USE OF DIGITAL OBJECTS AND DATA

Impact measures to consider

Lotte Wijsman

National Archives of the
Netherlands,
The Netherlands
lotte.wijsman@nationaalarchief.nl

Arie Groen

National Library of the Netherlands,
The Netherlands
arie.groen@kb.nl

Tamara van Zwol

Dutch Digital Heritage Network,
The Netherlands
tvzwol@beeldengeluid.nl

Robert Gillesse

International Institute of Social History,
The Netherlands
robert.gillesse@iisg.nl

Abstract – The storage and use of digital heritage objects produce carbon dioxide (CO₂) emissions. Cultural heritage organizations can take several measures into consideration in order to diminish these CO₂ emissions. However, how much CO₂ do storage and use produce and what measures could have (the most) effect? We examined the CO₂ impact and possible measures on the basis of a case study. We have focused our investigation on the impact of servers, infrastructure, cloud storage and use.

Keywords – carbon footprint, sustainability, storage, users, carbon dioxide emissions

Conference Topics – Environment

I. INTRODUCTION

Preserving digital objects for the public contributes, like many human activities, to carbon dioxide (CO₂) emissions and consequently has an impact on the environment. The Dutch digital heritage community is (becoming) conscious of the subject and wishes to examine the facts. What is the environmental impact of the storage and use of collections? And what measures can be taken to lessen the CO₂ impact?

This poster presentation provides insight into certain measures that can be taken, based on a CO₂ impact case study of the Delpher platform [1]. In

Delpher you can search and find millions of digitized text from Dutch newspapers, books, and. These documents come from the collections of various Dutch scientific institutions, libraries, and heritage institutions. Delpher is developed and managed by the National Library of The Netherlands (KB). The case study was executed by the company PHI Factory and the Green IT expert group within the Dutch Digital Heritage Network¹. We have examined storage and data use in this case study, focusing on the CO₂ impact of servers, the server environment/infrastructure, cloud storage, and the end use: searching through the files on the platform and downloading files. The poster presents our findings in those four areas [2].

II. SERVERS

Servers provide the computing power and storage required to store and make digital collections available for users. These servers are the main cause of CO₂ emissions. This is due to both the electricity consumption and the indirect CO₂ emissions from the production of the servers.

Creating digital compartments in the servers, like the KB has done for the data on Delpher, ensures that the capacity of these servers can be used more

¹ PHI Factory uses the guidelines from 'The Green House Gas Protocol' to measure the CO₂ footprint.

efficiently. This can be done by means of virtual machines or containers. The KB's servers consume now 242 MWh (or: 242,000 kWh) annually, which is equivalent to the electricity consumption of 98 average Dutch households in a year.

III. SERVER ENVIRONMENT

The location/environment of the servers has a major influence on the total of CO₂ emissions. If data is stored locally, on the level of one institution, there is a good chance that actions facilitating the servers, such as cooling them, consumes as much or even more energy than the servers themselves. To reduce the CO₂ impact of the infrastructure around the servers you can think about sharing servers with multiple organizations to use them most effectively. By moving the servers from the KB local location to a more efficient, external colocation data center, as in the case of Delpher considerable savings can be made on electricity costs: saving annually the amount of 151 MWh. Because many servers are located here, facility systems such as cooling can do their job much more effectively. Therefore, this method is not only more sustainable, but also more economic.

You can also opt for more green energy, like the KB has done. Green energy is any energy type that is generated from natural resources, such as sunlight, wind or water. Because green energy is generated from a renewable source, the CO₂ emissions are a whole lot lower than in the case of energy from fossil sources. The annual carbon footprint of Delpher's servers is less than 4 tons of CO₂ equivalents per year, which equals 4 hot air balloons of 200 m² (the size of a soccer field) filled with CO₂.

IV. CLOUD STORAGE

With cloud storage, the data and computing power of many companies is divided over servers. This makes for very efficient use of (the capacity of) the servers since every available space is being occupied. The advantage of storage in a cloud environment is that the type of providers behind it (e.g. Microsoft and Amazon) are at the forefront of the development of facility systems and the use of containers to make the capacity of their servers as efficiently as possible. Naturally, cultural heritage organizations have to consider if they are willing to store their data in a large datacenter under the control of such a provider in perhaps a different country, under different rules and regulations.

Because Delpher concerns itself with national Dutch cultural heritage, it has been decided to store the data at a Dutch colocation and not via an international cloud provider.

V. DATA USE

Retrieving files from a digital collection, loading webpages and using the search index causes CO₂ emissions. In the case of Delpher a large part of the digital collection will not be downloaded by a user, but searched, which has only a limited impact. Still, there are ways to even diminish this impact. This could be done by e.g. offering lower resolution versions of the digital object files. In addition, to make it even more effective, you can also limit the user features on the website so that fewer files have to be searched in the data store. For example if you do not offer 'search all' as a standard option, but let users indicate which specific material (newspapers, books or magazines) should be searched.

VI. CONCLUSIONS

With the findings from the case study and the aforementioned recommendations, cultural heritage institutions can start to examine the CO₂ impact of their own digital collections and make choices for a climate-resilient future. Also, the case study does ideally stimulate further discussion about selection and deduplications of collections in and between cultural heritage institutes.

REFERENCES

- [1] Delpher platform <https://www.delpher.nl/>
- [2] The Greenhouse Gas Protocol <https://ghgprotocol.org/>

COMMUNITY ARCHIVES AND DIGITAL SUSTAINABILITY

Challenges and Opportunities

John Pelan

Scottish Council on Archives
Scotland, UK
J.pelan@scottisharchives.org.uk
[0000-0003-1617-717X](tel:0000-0003-1617-717X)

Audrey Wilson

Scottish Council on Archives
Scotland, UK
a.wilson@scottisharchives.org.uk
[0000-0002-4205-3835](tel:0000-0002-4205-3835)

Sean Rippington

University of St Andrews
Scotland, UK
sbr2@st-andrews.ac.uk

Abstract – A short poster presentation on the issues facing smaller community organisations and volunteer-led groups related to digitisation, digital preservation, and digital sustainability.

Keywords – Community, Archives, Digitisation, Digital Preservation

Conference Topics – Community

I. INTRODUCTION

Audrey Wilson, Partnerships and Engagement Manager, Scottish Council on Archives (SCA) and Sean Rippington, Digital Archives and Copyright Manager, University of St Andrews, will summarise some of the challenges facing community groups dealing with digitisation and digital preservation issues.

II. THEMES

The poster presentation will set out some of the issues covered in the Developing Your Digital Skills: Digitisation Webinar Series coordinated by SCA and the Community Archives and Heritage Group Scotland network in 2021. The Developing Your Digital Skills series was created as a result of a survey which SCA sent out to community groups in early 2021. The presentation will also be informed by feedback and conversations which have emerged from SCA's ongoing work of supporting community groups, including voluntary heritage organisations, who maintain or would like to maintain a small archives.

We know that community archives are often run by volunteers in all parts of Scotland, from urban towns and cities to remote parts of Scotland. Before the pandemic, many people found it too costly and time-consuming to travel to training events. Covid19

made everyone, young and old, appreciate the importance of communicating on a digital platform and how it leads to more opportunities and the ability to engage with the world. The webinar series allowed everyone from anywhere to take part and learn how to Develop their Digital Skills, able to ask questions of the presenter and post messages in the chat room. With just over 800 people registering for the webinars, it was a huge success.

III. MAIN ISSUES

The main challenges facing community groups, including voluntary heritage organisations in terms of maintaining a digital archives, include:

- Lack of expertise/knowledge
- Unfamiliarity with terms such as digital preservation
- Sustainability of digital archive material on social media sites such as Facebook|
- Ageing demographic
- Vulnerability of data on personal computers and hard drives

IV. DIGITAL PRESERVATION ISSUES

The digital archives of community groups are listed on the Digital Preservation Coalition 'Bit List' as 'Critically Endangered'.

The Bit List mentions: poor documentation; lack of replication; lack of continuity funding; lack of residual mechanism; dependence on small number of volunteers, lack of preservation mandate; lack of preservation thinking at the outset; conflation of

backup with preservation; conflation of access and preservation; inaccessible to web archiving; dependence on social media providers; distrust of 'official' agencies as some of the key factors behind the extreme vulnerability of digital community archives.

The Bit List also states that "Typically born digital material is more at risk - community groups may not know about the risk of loss. Many are unaware of digital preservation terminology. It is the ad-hoc nature of these groups and projects which is of great concern."

This poster event will set out the key issues and invite contributions from digital preservation professionals to explore ways of helping community groups with their digital requirements.

REFERENCES

- [1] Scottish Council on Archives, Developing your Digital Skills for Community Archives course 2021, <https://www.scottisharchives.org.uk/latest/developing-your-digital-skills-for-community-archives-webinar-series/>
- [2] Digital Preservation Coalition Bit List, <https://www.dpconline.org/digipres/champion-digital-preservation/bit-list/critically-endangered/bitlist2021-community-groups>

LESSONS FROM THE NATIONAL ARCHIVES OF SINGAPORE'S JOURNEY DEVELOPING A DIGITAL PRESERVATION SYSTEM FOR PUBLIC RECORDS

Kevin Wong

National Archives of Singapore

Singapore

Kevin_wong@nlb.gov.sg

Abstract – This poster presents a case study of the ongoing development of a digital preservation system for public records at the National Archives of Singapore (NAS). It describes some challenges faced and lessons learnt, applying the conceit, “If I could travel back in time and speak to myself shortly after I joined this project, what would I say?”

Keywords – digital preservation system

Conference Topics – Resilience.

I. INTRODUCTION

The National Archives of Singapore (NAS) began planning for a digital preservation system for public records in 2016. This project has taken significantly longer than planned, and is still ongoing at the time of this presentation. Delays have stemmed partly from security considerations and difficulties engaging qualified vendors locally, and thus may not have been easily avoided. Nevertheless, the delay has resulted in some notable challenges.

II. THE NEED FOR A DIGITAL PRESERVATION SYSTEM

The NAS is given a mandate under the National Library Board (NLB) Act of Singapore to implement a records management regime across the whole of government and to preserve records of archival value for future access.

The NAS has so far mainly taken custody of public records in paper, preserving them on microfilm and, where necessary, in the original. However, as public records in Singapore are increasingly born-digital because of a nationwide push towards digitalisation, and while physical space remains a premium in our island city state, the NAS anticipates the need to preserve most of its collections digitally in the future.

Thus, the need to develop a digital preservation system for public records arises.

III. A BRIEF HISTORY OF THE PROJECT

Planning for the preservation system began with the project in 2016, with security being a key consideration from the start. Relevant security-related stakeholders were consulted. This turned out to be a time-consuming process, taking up nearly a year, not least because it was also the team's first time working on such a system, and some learning and experimentation had to take place along the way. The outcomes of this consultation shaped the design of the system significantly.

An initial tender to develop the system was published in 2018. However, this tender was unsuccessful, in part because of the limited number of vendors operating in Singapore with the experience needed to address both the security and digital preservation requirements.

In 2019, the NAS embarked on a Proof-of-Concept side project, working with a vendor to identify pain points obstructing the implementation of a preservation system complying with the NAS's specific requirements. Specifications for a second tender were drafted to broaden the digital preservation requirements and to provide greater clarity on how security requirements could be addressed. A second tender was called in late 2021 and awarded in February 2022.

IV. CHALLENGES FACED

The project is now entering its 6th year. This longer-than-expected project time frame has led to several challenges, which the project team has had to address:

1) Transfers put on hold: Transfers of digital records to the NAS were put on hold awaiting the implementation of the system, since the system is needed to perform checks on records before ingestion. This has led to increased pressure from government agencies and risk that records would be left unmanaged and lost.

2) Spill-over effects: The digital preservation system is designed as part of a suite of interconnected systems and local standards, whose development has continued even as that of the preservation system has lagged behind. Uncertainty about what the preservation system will finally look like spills over into these related projects, as additional care must be taken to ensure they all work together once completed.

3) Scope creep: As the project has gotten older, it has become easier for staff to lose sight of what a preservation system is actually intended to do, so that it is sometimes assumed that the system will solve any preservation-related problem, when in fact better solutions may lie elsewhere.

4) Knowledge transition: The project is old enough now that ordinary staff turnover has led to none of the archivists involved in the project today having been around at its start. New staff have had to learn digital preservation very quickly, relying on email records to understand decisions made by staff who have since left the organisation.

V. LESSONS LEARNT

If the author had access to a time machine, he would have the following advice for his younger self:

1) Jumpstart your digital preservation education. The best way to learn is by doing, so start doing things as soon as you can.

a) Understand that a preservation system is only part of the puzzle. Understand what activities are involved in digital preservation, and where the system fits in these activities. Start doing the work that you can do, and start planning for the work that needs to be done once the system is ready.

b) An end-to-end workflow really does matter. You will come across this concept very quickly in your digital preservation research. It is very easy to brush off as common sense, but don't. The sooner you realise how important this idea is, the better.

2) Expect delays and surprises, and factor them into your plans. If you are delaying accepting digital transfers, don't take for granted that you are going to be able restart them in X years. Think about what you can do for your stakeholders in the meantime, and what they can do for you. You may even want to include plan for interim transfers.

3) Manage expectations and lead by example.

a) Whatever digital preservation work you do aside from work on the system, keep management constantly appraised, so that they understand as well as you do that the system is not the whole solution, and an ongoing investment of resources is needed for the project to be sustainable.

b) No matter how much you tell them otherwise, people are going to keep acting as if preservation is only about storage. You have to be prepared to keep showing them that it is not through your actions, for example, by involving them in preservation planning discussions.

4) Don't over-specify your system. Build in flexibility. Make up for your own, developing expertise by asking for a vendor who will not only take direction from you, but will partner with you to figure out and develop the system together. Ideally, you would get a vendor with subject knowledge, but you can just as well do with a vendor who will ask you the right questions during requirements gathering.

VI. MOVING FORWARD

While frustrating for those involved, the delays in the project timeline may be seen as a blessing in disguise, as they have forced the team to grapple with problems that some might argue are unavoidable, and better confronted earlier than later. The team has learnt a lot in this time, and with the second tender successfully awarded, NAS looks forward to implementing its new system by 2023.

BIT PRESERVATION USING THE OPEN SOURCE BITREPOSITORY.ORG FRAMEWORK

*Use, benefits, robust design principles and
enhancements during the past ten years*

Eld Zierau

Royal Danish Library
Denmark
elzi@kb.dk
[0000-0003-3406-3555](tel:0000-0003-3406-3555)

Mathias S. Jensen

Royal Danish Library
Denmark
masj@kb.dk

**Rasmus B.
Kristensen**

Royal Danish Library
Denmark
rbkr@kb.dk

Abstract – The Royal Danish Library has used the open source BitRepository.org framework as basis for bit preservation of Danish cultural heritage for the past ten years. This poster will present the capabilities of the BitRepository.org framework with respect to how it can support advanced bit preservation on changing software and media technologies. The BitRepository.org framework enables use of storage of copies on all types of current and future media, it supports daily bit preservation operations, it enables setup with high access possibilities as well as providing a basis for high operation security at all levels. The poster will also present experience with the use of the BitRepository.org as well as how the Royal Danish Library uses it for different levels of bit safety, confidentiality, access and costs.

Keywords – bit preservation, open source, information security, independence, future proof.

Conference Topics – Resilience; Community.

I. EXTENDED ABSTRACT

This poster will present the Danish bit preservation solution with focus on the underlying open source software framework BitRepository.org, which the Royal Danish Library has used for ten years for bit preservation of Danish cultural heritage. Furthermore, the background for the development of BitRepository.org and its actual use for securing Danish cultural heritage bits will be presented.

As for all bit preservation solutions, the framework can be seen as an implementation

supporting the three main principles of bit preservation:

- A number of copies of data
- Independency between copies of data with respect to technology, organization and placement
- Frequent Integrity checks of copies both locally on copies and between copies

The terminology used corresponds to the following general view of a bit repository with bit preservation.

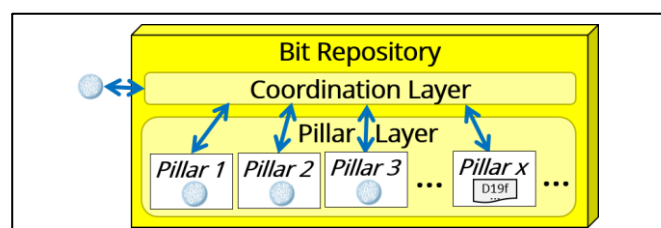


Figure 1 A general view of a bit repository with bit preservation

The Coordination Layer includes services like integrity checks between different copies, and each Pillar represents the organization and technology in serving the storage and safety of an individual copy.

The poster will present the capabilities of the BitRepository.org framework as a basis to support advanced bit preservation including various requirements to:

- **bit safety**, by allowing pillar services for the different copies of data to be instantiated on different technologies and media, in different organizational environments at different places, and with high independency between the different pillars.
- **confidentiality**, by supporting encrypted communication between components and allowing offline pillars as well as security around the individual copies. Furthermore, in 2022 encrypted copies will be supported. This is needed when copies are placed at another organization.
- **sustainability**, by being independent of the implementation of the different pillar services with respect to changing software & media technologies and geographical and organizational location, e.g. the Royal Danish Library is currently replacing one of the pillar services in order to obtain independence between copies.
- **access**, by making it possible to have pillars that are particularly well suited for access purposes, e.g. the Royal Danish Library has recently transferred the Danish web archive (Netarkivet) to the BitRepository.org framework, where one of the pillar services is designed to support access via Wayback applications¹.
- **costs**, by making it possible to have pillars with cheap storage facilities such as tapes, and to some extent by being an open source framework.

The poster will include a description of the robust design principles which enable fulfillment of these requirements. One of the main principles is that components of the system must have **no** direct knowledge of each other's implementation. This principle is ensured by design of a common message protocol, which is the only common knowledge between the components of the system. The poster will therefore also include an illustration of how this protocol is implemented.

The poster will also contain a description and illustration of the services that support execution and monitoring of bit preservation actions in the daily operation, e.g. actions like checks of missing files, consistency checks of checksums across all involved copies, surveillance of recalculation time for checksums for individual copies, the possibility of replacing faulty copies, and various monitoring operations.

The poster will also describe why the Royal Danish Library joined forces with the Danish National Archives to develop Bitrepository.org in the first place, and why we are still convinced that this is the best solution for our bit preservation. This description will be accompanied by a presentation of our current implementation, and the recent additions to support further independence and support of placing pillars with encrypted data in e.g. an organization under foreign jurisdiction.

If animated posters are possible, we will provide an animation presenting a demo of parts of the system.

We will be happy to provide a supplementary short paper, if wanted.

REFERENCES

The references provide some literature about the practices of bit preservation [1,2,3], which also includes some description of Bitrepository.org [2,3], as well as references to the criteria and usage guide which is helpful in evaluating bit preservation solutions.

- [1] D. S. H. Rosenthal, "Bit Preservation: A Solved Problem?", Proceedings of the 5th International Conference on Preservation of Digital Objects, London, Great Britain, 2008, pp. 274-280.
- [2] E. Zierau, "The Rescue of the Danish Bits". Proceedings of the 15th International Conference on Preservation of Digital Objects, 2018, DOI: 10.17605/OSF.IO/U5W3Q.
- [3] E. Zierau, "Comparing How To Take Care of Humans' and Bitstreams' Lives", Proceedings of the 18th International Conference on Preservation of Digital Objects, 2021.

¹ One Wayback application is e.g. described here: https://en.wikipedia.org/wiki/Wayback_Machine

PRESERVING COMPLEX DIGITAL OBJECTS REVISITED

Patricia Falcao

Tate

UK

Patricia.Falcao@tate.org.uk

[0000-0003-2798-5631](tel:0000-0003-2798-5631)

Caylin Smith

Cambridge University

Library

UK

caylin.smith@lib.cam.ac.uk

[0000-0001-6340-5708](tel:0000-0001-6340-5708)

Sara Day Thomson

Edinburgh University

Library

UK

Sara.Thomson@ed.ac.uk

[0000-0002-3896-3414](tel:0000-0002-3896-3414)

Abstract - This workshop revisits the iPres 2019 workshop on *complex digital objects* to address the opportunities and challenges generated by works created using novel or non-standard technologies. Collection management solutions for such objects are becoming an increasing need for museums, libraries, and archives. At the 2019 workshop, participants indicated they currently have complex objects in their care without a solution for preserving them. This updated workshop will draw on sector-wide progress as well as innovations catalyzed by rapid collecting initiatives to document the COVID pandemic. Preserving Complex Digital Objects - Revisited will again create an opportunity for digital preservation professionals to share insights and experiences and forge new paths forward together.

Keywords - file formats, time-based media, technology watch, collaboration, capacity building

Conference Topics - Community; Exchange; Innovation

I. LEARNING GOALS

- Participants will collaborate and exchange knowledge and practical experiences with other group members to enhance community understanding of approaching objects for which few or no collection management solutions exist.
- Using the concept of *minimum viable preservation (MVP)* [1], participants will gain practical know-how to get started in planning for the preservation of complex digital objects at their home institutions.

II. DESCRIPTION

Although progress has been made within preservation communities (e.g., digital preservation, time-based media conservation, web archiving) since the 2019 workshop, the challenges that works created using novels or non-standard technologies

pose to collecting institutions persist and grow as technology evolves. These works cannot be resolved by any single sequence of preservation actions, reference model, tool, or service. Collecting institutions must react to the growing remit of their collections, the ways creators realize their works, and adapt to these contexts.

This workshop addresses the practical challenges of 'complex digital objects', as defined in the 2019 workshop. No matter how up-to-date, responsive, and well-resourced an institution's response to digital preservation might be, the knowledge needed to manage and preserve these objects will always lag behind the growth of the technology used in their creation.

The organizers will apply the research they have undertaken in this area to small group activities. This approach will help to engage members of the digital preservation community to cultivate shared knowledge and to anticipate similar challenges that their institutions will encounter.

III. BACKGROUND

A. Tate's Time-based Media Conservation

Tate's Time-based Media (TiBM) conservation team is responsible for the preservation of Collection artworks using performance, film, slides, video, audio, and software. More recently, web-based artworks have also made their way into the collection, which has resulted in new research into preservation strategies and implementing new processes within the pre-existing framework. In some cases, the object of preservation is not necessarily the software or data but the experience of the artwork. The TiBM team has developed risk assessment and analysis processes to evaluate the vulnerability of individual artworks and technologies

and identify the diverse options for preservation (from storage to migration and emulation). Documentation of the artwork and its technical history, while making the work more sustainable, pre-empt future issues, and guide any intervention to maintain the artwork's functions in the present.

B. Cambridge University Library

Cambridge University Library is increasingly collecting born-digital works that can be considered *complex digital objects*. These works exist within the Library's archives as well as deposited to the University's institutional repository. This challenge is not unique to CUL but experienced by libraries and other collecting institutions worldwide as digital works are increasingly made in a diverse range of formats, many of which share characteristics with more complex digital works found in time-based media collections.

CUL Digital Preservation takes a lifecycle approach to digital preservation, embedding activities that help ensure ongoing and faithful access when and where necessary. In addition, the Digital Preservation team is engaged in a wider community of practitioners researching complex digital objects, including the UK Legal Deposit Libraries' Emerging Formats work about the collection management needs of complex born-digital published works in scope to collect under the UK's legal deposit regulations.

C. Edinburgh University Library

Edinburgh University Library (EUL) takes a converged approach to digital preservation, collaborating across teams and working closely with academic partners. Across the Archives, Art Collections and Museums, Digital Library, Research Data Management, Learning Teaching Web, and beyond, professionals with different backgrounds collaborate and share best practice. As the formats and media used to disseminate information evolve, so too does the Library's commitment to support their preservation and use by an international community of researchers.

This integrated approach supports the increasing need to find inter-disciplinary solutions for maintaining access to complex digital objects. From database-driven services to world-changing research data in obscure formats, these complex objects require a range of professional skill sets to understand and maintain. As EUL continues to

explore experimental and practical solutions, they seek to share lessons learned and discover how others approach these challenges.

IV. CONTENT

This workshop will discuss definitions for complex digital objects and provide an overview of the known challenges to preserving them. The first section of the workshop will focus on three predominant challenges:

- 1) Defining the complex digital object and its significant properties and using this information to decide what to preserve.
- 2) Problem-solving technical dependencies, including software and hardware environments.
- 3) Strategizing for digital rights management and intellectual property rights.

The organizers will present three case studies that exemplify these challenges. Participants will then break out into small groups for an activity designed to analyze and problem-solve the challenges of preserving complex digital objects. The activity will lead participants through a practical, 'less is more' approach, like the MVP approach described by Matthew Addis and 'parsimonious preservation' described by Tim Gollins [2]. Each small group will focus on identifying preservation needs (based on end user requirements) and then on formulating targeted solutions. Though the definition of 'minimum viable' will vary from institution to institution, the practical constraints of maintaining such complex digital resources (especially if at scale) are almost universal.

In the final 30 minutes of the workshop, participants will feed back the results of their small group activities and discuss common trends as well as divergent approaches. Feedback will be collected and recorded to document ideas and analysis generated by participants. The workshop aims to identify opportunities for collaboration in the development of new approaches.

REFERENCES

- [1] M. Addis, "Minimum Viable Preservation", DPC Blog, 12 November 2018, <https://www.dpconline.org/blog/minimum-viable-preservation>.
- [2] T. Gollins, "Parsimonious preservation: preventing pointless processes!", in Online Information 2009 Proceedings, <https://cdn.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation.pdf>

WORKSHOP: CHANGING CURRICULUMS FOR A CHANGING WORLD?

Living in Interesting Times: Digital Preservation Education, Pedagogy and Skills

Ann Gow

University of Glasgow
United Kingdom
Ann.gow@glasgow.ac.uk
[0000-0002-0172-6036](tel:0000-0002-0172-6036)

Paul Gooding

University of Glasgow
United Kingdom
paul.gooding@glasgow.ac.uk
[0000-0003-1044-509X](tel:0000-0003-1044-509X)

Zoe Bartliff

University of Glasgow
United Kingdom
Zoe.bartliff@glasgow.ac.uk
[0000-0002-1644-3151](tel:0000-0002-1644-3151)

Yunhyong Kim

University of Glasgow
United Kingdom
Yunhyong.kim@glasgow.ac.uk
[0000-0001-5400-0389](tel:0000-0001-5400-0389)

Kathryn Simpson

University of Glasgow
United Kingdom
Kathryn.simpson@glasgow.ac.uk
[0000-0002-6883-4146](tel:0000-0002-6883-4146)

Abstract – As we live through the significant challenges of pandemic, information wars, climate change, and war in Europe, what skills do current students and future practitioners need to cope with managing digital materials in our historical moment? What might a holistic syllabus encompassing DP skills, knowledge, and personal qualities alongside awareness of social and political trends, and an understanding of the interplay between the two, look like in a Higher Education context? This workshop aims to discuss and disrupt the ideas around Digital Curation (DC) and Digital Preservation (DP) education for future professionals and those managing digital collections. It will bring together stakeholders drawn from the those delivering DP education, those learning, employers, and practitioners, to interrogate and reflect on the suitability of existing curricula in Higher Education (HE) for a changing world.

Keywords – Pedagogy, Skills Training, Education.
Conference Topics – Community, Resilience.

I. INTRODUCTION

Information Studies at the University of Glasgow runs three programmes at undergraduate and Master's level that incorporate Digital Curation (DC) and Digital Preservation (DP) across the curriculum. All are accredited by CILIP and ARA. While a good

number of our students do move to employment in the cultural and heritage sector, many students take roles across different sectors that are not directly engaged in digital preservation. Our experience is that digital preservation is, like information literacy, a key skill for the world and our students in whatever roles they shape and inhabit in the years to come.

Increasingly, however, those roles will be shaped by global events that present significant challenges to existing DP practices and knowledge. The climate emergency requires us to interrogate the environmental impact and sustainability of DP processes; the pandemic requires us to reassess how digital materials are preserved and made accessible in a time of crisis; and war in Eastern Europe necessitates a deep understanding of misinformation and securing digital culture in crisis. iPRES 2022 therefore presents an ideal opportunity for us to reassess skills and knowledge, and the mechanisms by which they are taught in HE, to ensure students and future practitioners are equipped for these emerging challenges.

A. Aims

The overall workshop aim is to assess whether DC and DP education in Higher Education continues to meet community needs. Existing frameworks detail the knowledge, skills and experience required for DP and DC practitioners. DigCurv [1], and subsequent frameworks [2], encompass personal and professional qualities alongside domain expertise and intellectual abilities. However, considering the significant challenges posed by climate change, conflict, and pandemic, this workshop provides a timely moment to interrogate our existing syllabi and pedagogies, to equip practitioners for a changing and demanding world.

The workshop will therefore address three overarching questions that will structure discussions and inform a subsequent academic paper:

- 1.) What is the current status of DC and DP education in HE institutions?
- 2.) What emerging skills and knowledge will DP practitioners need in the coming years?
- 3.) How might existing frameworks require adapting to meet these emerging needs?

B. Workshop Structure

This in-person workshop will consist of two 90-minute blocks. These workshops will be structured around the following topics:

Block A:

- Plenary Presentation: The State of Play in DP/DC Education (30 minutes)
- Presentation and Discussion: DP Education and Global Crisis (30 minutes)
- Breakout 1: What skills and knowledge will DP practitioners need in the coming years?

Block B:

- Breakout 2: How might existing frameworks require adapting to meet these emerging needs? (30 minutes)
- Feedback session and synthesis (45 minutes)
- Summary and next steps (15 minutes).

C. Participants

We aim to attract a wide range of participants, drawing on the experience of educators, practitioners, students, and others. This is likely to include the following groups:

- Educators in HE and elsewhere;
- Practitioners interested in skills development;

- Former students in taught programmes with a DP/DC component;
- Current Postgraduate Research students;
- DP practitioners with expertise in skills and training, for CPD or within their organisation;
- Digital archivists from UofG Library and Archive Services;
- Researchers/practitioners with interests in social, environmental, and political contexts within which DP/DC activities operate.

The workshop is built around breakout discussions to ensure attendees can actively contribute, with speakers from various backgrounds.

D. Workshop Outcomes

The main outcome will be a paper submitted to a relevant journal, proposing how DP pedagogy and syllabi might react to the new challenges of preserving digital content in a world faced with pandemic, war and climate emergency. The organisers have a longstanding interest in pedagogies for digital curation [1] and digital humanities [3], and seek to work with the broader community to address these vital topics.

The workshop discussion will inform this paper, and participants will be invited to contribute directly as co-authors – although this will not be a requirement. More broadly, we hope this workshop will lead to an inclusive community of practice with shared interest in the changing contexts for DP education and training, and which will be able to contribute to further research in this area.

REFERENCES

- [1] L. Molloy, A. Gow, and L. Konstantelos, "The DigCurv curriculum framework for digital curation in the cultural heritage sector. International Journal of Digital Curation, vol. 9(1), pp. 23-241. June 2014.
- [2] S. Mason. Digital Preservation at Oxford and Cambridge Training Needs Assessment Toolkit. University of Oxford. 2018.
- [3] F. Benatti, P. Gooding, and M. Sillence, "Learning digital humanities in a community of practice: the DEAR model of postgraduate research training. Digital Humanities Quarterly, vol. 15(3). 2021.

WELCOME TO FEDORA 6.0

Features, Migration Support & Integrations For Community Use Cases

Arran Griffith

LYRASIS

Canada

arran.griffith@lyrasis.org

Daniel Bernstein

LYRASIS

USA

danie.bernstein@lyrasis.org

In July, 2021, the long-awaited Fedora 6.0 was released. This workshop will provide an overview of the software itself, a look at our roadmap and path to release, as well as dive into some important new features that helped return Fedora to its digital preservation roots. We will showcase and demonstrate the much-anticipated migration tooling and documentation as we work through a hands-on migration. Lastly we will demonstrate how to integrate Fedora with your ecosystem via the Camel Toolbox.

This is a technical workshop pitched at an introductory level so no prior Fedora experience is required. General knowledge of the role and functionalities of repositories would be beneficial. Attendees who wish to participate in the optional hands-on sections will need to access an online sandbox via a URL which will be provided ahead of the workshop.

Keywords: Fedora, Repository, Open Source, Migrations

Conference Topics – Innovation; Resilience

I. INTRODUCTION

In July, 2021, the long-awaited Fedora 6.0 was released. This workshop will provide an overview of the software itself, a look at our roadmap and path to release, as well as dive into some important new features that helped return Fedora to its digital preservation roots. We will showcase and demonstrate the much-anticipated migration tooling and documentation as we work through a hands-on migration. Lastly we will demonstrate how to integrate Fedora with your ecosystem via the Camel Toolbox.

The workshop will include several hands-on portions that will allow attendees to exercise Fedora features, while learning about their purpose and function. These features are accessible via a built-in web interface, so no command line experience is required.

II. HANDS-ON BREAKDOWN

We propose to break the hands-on portion down in to the following segments for easier comprehension:

Section 1: Fedora 6 Technical Overview & Resources Management

- Highlight and test new features of Fedora 6.0 and understanding how to work with resources within the Fedora platform.

Section 2: Migration

- Participants will engage in a migration of a small data set from Fedora 3.x to Fedora 6 using the migration tooling.

Section 3: Fedora and the Camel Toolbox

- Understanding the Camel Toolbox and demonstrating how to integrate Fedora with your ecosystem using it.

This is a technical workshop pitched at an introductory level so no prior Fedora experience is required. General knowledge of the role and functionalities of repositories would be beneficial. Attendees who wish to participate in the optional hands-on sections will need to access an online sandbox via a URL which will be provided ahead of the workshop.

III. LEARNING OUTCOMES

By the end of this workshop, attendees will be able to:

1. Be familiar with core and extended Fedora features and its integration capabilities
2. Create and manage content in Fedora
3. Understand how to use the migration tooling to ensure a successful migration
4. Understand how Fedora supports digital preservation

REFERENCES

- [1] Manuscript Templates for Conference Proceedings, IEEE.
http://www.ieee.org/conferences_events/conferences/publishing/templates.html

THE CLIMATE CRISIS AND NEW PARADIGMS FOR DIGITAL ACCESS

James Baker

*University of Southampton
UK*

*j.w.baker@soton.ac.uk
[0000-0002-2682-6922](tel:0000-0002-2682-6922)*

Rachel MacGregor

*University of Warwick
UK*

*rachel.macgregor@warwick.ac.uk
[0000-0002-4296-6159](tel:0000-0002-4296-6159)*

Anna McNally

*University of Westminster
UK*

*a.mcnelly@westminster.ac.uk
[0000-0002-4023-2207](tel:0000-0002-4023-2207)*

Abstract – The need for action on the climate crisis is more urgent than ever, and our demand for energy to power both storage and access to data is merely fueling the emergency. This workshop will look at the different pressures on born-digital archives, digitized records and their respective uses by researchers - of all disciplines - but with a particular focus on digital humanities. Based on Climate Crisis initiatives across both the digital humanities and digital preservation sectors, participants will consider the extent to which the expectations of access to digital materials are derived from older paradigms, and how we as a community can plan and advocate for alternatives.

Keywords – Sustainability, access, community, humanities

Conference Topics – Environment; Resilience

I. INTENT AND BACKGROUND

The need for action on the climate crisis is more urgent than ever, and our demand for energy to power storage and access to data is merely fueling the emergency. Preservationists [1] and researchers [2] alike have joined in calls to enact change within their communities. In a recent reflection on the Greening the Digital Humanities Workshop a participant wrote: “The middle scale, the often distinctly unpoetic activity of organizing with a few others to influence an organization, a sector, a community of practice, a regulation or practice, is often what goes missing.” [2]. To this end we encourage members of the digital preservation community of practice to come together to establish new paradigms for a more sustainable future.

Digital archivists and curators are under increasing pressure to provide everything as digital, perfectly cataloged, available instantly. Metrics of digital material added to an online portal are often a

key way that archivists and curators report on and justify their work to their institutions. Yet as the sector comes to increasingly recognise the carbon cost of their work, and the threat it imposes to the very material we are trying to preserve, the sector needs to ask some difficult questions.

Do we really need to digitize everything? Does everything need to be on instant access storage? Do we need to create access copies of those files on ingest or can we wait until someone requests to see them? Does the climate crisis empower us to be more active in appraising new acquisitions and deaccessioning already-preserved resources? And through these discussions about data how can we use our expertise to contribute to climate justice?

These same discussions are taking place within the digital humanities community [4], with calls to reconsider both the structures currently underpinning digital-based research and methods of its dissemination. The experience of a global pandemic has shown that new ways of working are possible and iPres 2022 seems like an ideal opportunity for the custodians and users of digital records to consider alternative future practices.

In order to propose and argue for alternatives, we need the tools that enable us to make good decisions. At a very simplistic level, we need to know whether it is more harmful for someone to fly from New York to London to look at a resource, or for the archivist in London to make high resolution digital images of them available online 24/7 so that someone in America can view them at 3am GMT? To gather those tools, we need as a community, to engage with like-minded communities in

comparable fields, and - ideally - we need to proceed with a sense of urgency.

II. OUTCOMES

Participants will:

- leave with knowledge that will help them begin planning for action.
- gain insight into where their expertise is most needed.
- learn of comparable initiatives, including the Digital Humanities Climate Coalition [3] and the Digital Humanities and Climate Crisis Manifesto [4].
- consider how to adapt existing risk assessment processes (such as The National Archives' DiAGRAM project [5]) for climate justice.
- form agendas and priorities for community action.

III. AUDIENCE

The workshop is aimed at those engaged with digital preservation (technical, archival, curatorial) and researchers (all disciplines but with a particular focus on digital humanities).

IV. STRUCTURE

The workshop will be led by practitioners and researchers from across the archives, digital preservation and digital humanities sectors to encourage and facilitate lively discourse and debate.

1) *Introduction*: The workshop will open with an introduction to outline some of the issues and how they relate to digital preservation.

2) *Brainstorm*: A brainstorming activity utilizing an interactive whiteboard to capture initial thoughts.

3) *Breakout Sessions*: Separate facilitated sessions to take a closer look at the issues raised.

4) *Plenary Session*: Draw the session together and agree agendas and priorities for community action.

The breakout sessions would include:

- Developing an advocacy strategy aimed at raising awareness of the issues with practitioners, researchers and organizations.
- Analyzing the risks of challenging preservation practices.

- Considering the research questions, methods and approaches which are likely to have the greatest impact.
- Identifying who we need to influence, how, and in what timescales.

V. CONCLUSION

We anticipate raising as many questions as answering them, but we aim to identify changes that can be made at a personal and organizational level and push for an agenda which drives sectoral change, causes least environmental harm, and supports a just transition.

REFERENCES

- [1] Toward Environmentally Sustainable Digital Preservation. <https://dash.harvard.edu/handle/1/40741399>
- [2] Reflections on Greening the Digital Humanities, and What Comes Next? <https://www.cdcs.ed.ac.uk/news/reflections-greening-digital-humanities-what-comes-next>
- [3] Digital Humanities Climate Coalition. <https://www.cdcs.ed.ac.uk/digital-humanities-climate-coalition>
- [4] Digital Humanities and the Climate Crisis: a manifesto. <https://dhc-barnard.github.io/dhclimate/>
- [5] Safeguarding the nation's digital memory. <https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/research-collaboration/safeguarding-the-nations-digital-memory/>

THE BITS IN THE BYTES

Understanding File Format Identification

Francesca Mackenzie

*The National Archives, UK
United Kingdom
francesca.mackenzie@nationalarchives.gov.uk*

Andrea Hricikova

*The National Archives, UK
United Kingdom
andrea.hricikova@nationalarchives.gov.uk*

Andrey Kotov

*The National Archives, UK
United Kingdom
andrey.kotov@nationalarchives.gov.uk*

Kathryn Phelps

*The National Archives, UK
United Kingdom
Kathryn.phelps@nationalarchives.gov.uk*

Abstract – This workshop will provide practical experience of analyzing digital files to create signatures for format identification. It will build confidence in file-format analysis and develop participants' understanding of a range of methods that can be applied to different files and content types. We will explore the approaches adopted by the major file format identification tools used by the digital preservation community. During this workshop, attendees will gain hands-on experience in the tools needed to contribute file format research to the open-source registry PRONOM; some participants will be analyzing the digits, or hex, of their files for potentially the first time. As well as being educational, file format identification is a lot of fun!

PRONOM as a tool ties in well with the themes of the conference. PRONOM is open source and used across the globe in the information management and digital preservation sectors, and beyond. It embodies the value of data for all and encourages understanding of file formats for future preservation needs.

PRONOM particularly embodies the key conference themes of community and exchange. We rely on so many talented file-format researchers around the world to analyze digital collections, flag issues and contribute to our shared knowledge of file formats. We want to continue the conversation with the digital preservation community and enable more people to participate in this collective endeavor.

Keywords – Collaboration, Hex, Community, Formats, Conversation.

Conference Topics – Exchange; Community

I. OVERVIEW

The workshop will be organized and run by the PRONOM file format team at The National Archives, UK. By the end of this workshop participants will be equipped with the knowledge and skills required to conduct file-format research and submit new entries to PRONOM. Attendees will gain an understanding of the theory behind creating file format signatures and will be able to apply this to their own digital collections. An understanding of what file formats you have in your repository is an essential part of digital preservation. Why? For the same reason it is essential for a conservator to know if a physical record is vellum or paper, digital preservation experts have to know what file formats are in their collection in order to effectively preserve them for future generations.

The workshop will be a mix of demonstration, lecture and hands-on activities. We would aim to make the workshop hybrids so that participants who cannot attend the conference in person would also be able to attend. Sample files will be provided but participants will be encouraged to bring their own unidentified file formats too.

We would additionally wish to elicit and capture feedback from participants on how they would prefer to continue engaging with and contributing to PRONOM in the future. How best can PRONOM facilitate file format research and collaborate with the community?

II. KEY AIMS

The aim and benefits of the workshop is that participants should come away with:

1) An understanding about file formats, signatures and the approaches available for characterizing files for digital preservation.

2) Learn how file format signatures are created, tested and submitted to PRONOM.

3) Have an increased knowledge of the range of (free) resources available to file format analysts and gain some practical experience using these tools.

4) An understanding of their own digital collections at a more binary level.

5) To feel confident enough to join the PRONOM community! Everyone can do file format research, and if they so wish, everyone should be able to contribute to research and help us improve.

III. TARGET AUDIENCE

Everyone is welcome who is willing to learn more about file formats and their identification. This workshop would be particularly useful for anyone who currently works or plans on working with the PRONOM technical registry, or wishes to develop a deeper understanding of format-based tools or workflows.

IV. WORKSHOP STRUCTURE

Introduction to file format research (10 minutes)

File format signatures within PRONOM (20 minutes)

Activity: File format signatures using hex editors and file format specifications to create a signature. This will include reading byte streams, file format specifications and creating your own signature. (40 minutes)

Activity: Signature testing in which we will cover use of the skeleton suite, file format utility tool and testing your signature using DROID. (30 minutes)

BREAK (20 minutes)

Advanced signature development. This will cover container signature development and finding file samples. (40 minutes)

Activity: Open file format research and feedback (50 minutes)

This is a time for participants who want to stay to work on their own unidentified formats to stay and ask questions. It would also be valuable to ask the community how PRONOM can better support them.

V. OTHER INFORMATION

4.1 Please Bring

Own laptop. It is also recommended to have an installation of DROID; installation of a text editor (e.g. notepad++ or sublime); installation of a hex editing software (e.g. HxD). Unidentified file formats also welcome.

4.2 Workshop Length

The workshop will be 3.5 hours and include a 20 minute break.

REFERENCES

- [1] Manuscript Templates for Conference Proceedings, IEEE. http://www.ieee.org/conferences_events/conferences/publishing/templates.html

ETERNALIZE DBs WORKSHOP

Exchange on sustainability and re-use of database content

Kevin A. McMahon

*New Mexico, United States
kamcmah@comcast.net
+1 505 944 6511*

Dr. Kai Naumann

*Landesarchiv Baden-
Württemberg
Germany
kai.naumann@la-bw.de
0000-0002-2799-1030*

This is a short report on the workshop as it happened on 12th September 2022 on iPRES 2022. It was conceived as a follow-up to the virtual DBs for 2080 workshop at Landesarchiv Baden-Württemberg in October 2021 that addressed questions of how to enable memory institutions of all kind to store the very diverse and voluminous database content the world creates in a trustworthy, secure, safe, and efficient way [1].

Keywords – database content, format migration, emulation, GeoPackage, SIARD
Conference Topics – Sustainability

I. INTRODUCTION

The workshop proceedings of October 2021 have been released in August 2022 [1], the slides and videos are also available [2] and a short report has been published [3]. Lead questions resulting out of the workshop and its proceedings were discussed and new connections were established.

II. GENERAL LEAD QUESTIONS

There were lead questions resulting out of the 2021 proceedings. During the workshop, no really new answers arose, but the tasks were further explained:

- How to raise risk awareness about the lack of standard procedures for the revival of “cold” database content?
- How to join inventive spirit and standardization potential worldwide and between communities?

- Who are our allies? Can commercial software producers or large organizations in specific sectors contribute generic software services for interoperability? Examples are e-discovery tools in law, self-explaining bootstrap routines for decoding DNA-coded data in biotechnology, long-term requirements about nuclear waste disposal in nuclear science.

III. OUTLOOKS

Participants were encouraged to join the DILCIS Relational DataBase Archiving Interest Group (RDB-AIG) mailing list [4] and spread the word about it. Interest group meetings will be announced on this medium.

The database engineering community has been addressed with an article on SIGMOD Record [5].

The Swiss Federal Archives told us that SIARD Suite, first released in 2007, has been continued and happily announced completion of SIARD Suite version 2.2, an open source product [6]. The Swiss federal administration has achieved an IT security rating for the tool. SIARD Suite could thus be added to the standard software directory of federal agencies.

SIARD users again reminded people of using the excellent case studies [7] on the standard.

REFERENCES

- [1] Naumann, Kai (ed.). Databases for 2080. Workshop Proceedings, Stuttgart 2022, <https://nbn-resolving.org/urn:nbn:de:101:1-2022071903>.

- [2] Databases for 2080 workshop at Landesarchiv Baden-Württemberg, October 5-6 2021, <https://www.landesarchiv-bw.de/de/aktuelles/termine/72973>.
- [3] Naumann, Kai. "Databases for 2080 – Preserving database content for the long term", ABI Technik, vol. 42, no. 1, 2022, pp. 78-80. <https://doi.org/10.1515/abitech-2022-0009>.
- [4] URL <https://listserv.dilcis.eu/info/rdb-aig>
- [5] Appuswamy, Raja. "Towards Passive, Migration-Free, Standardized, Long-Term Database Archival", ACM SIGMOD Record, Volume 51, Issue 2, June 2022. <https://dl.acm.org/doi/10.1145/3552490.3552506>
- [6] URL <https://www.bar.admin.ch/bar/en/home/archiving/tools/siardi-suite.html>
- [7] URL <https://dilcis.eu/content-types/siardi>

THE VALUE OF CATASTROPHIC DATA LOSS

Data loss as a community benefit.

Paul Stokes

Jisc
UK

Paul.stokes@jisc.ac.uk
[0000-0002-7333-4998](tel:0000-0002-7333-4998)

Tamsin Burland

Jisc
UK

Tamsin.burland@jisc.ac.uk
[0000-0002-5129-979X](tel:0000-0002-5129-979X)

Sarah Middleton

Digital Preservation Coalition
UK

sarah.middleton@dpconline.org
[0000-0002-7671-403X](tel:0000-0002-7671-403X)

Abstract – Jisc and the Digital Preservation Coalition (DPC) have undertaken an anonymous surveying exercise in order to unearth the true cost of catastrophic data loss—not only in terms of the value of the data, but also the cost of the knock-on effects that may only become apparent some considerable time after the event. This information can be used to help organisations make a stronger case for robust and effective digital preservation practice and to inform those trying to take steps to avoid their own data loss disaster. This workshop, intended to bring together those who have lost data with those who wish to avoid losses, explores the usefulness of such data loss events to the community.

Keywords – Sustainability, Cost, Value, Risk, Data loss.

Conference Topics – Resilience, Exchange.

I. INTRODUCTION

Digital Preservation is about mitigating risk. Mitigations cost money. It is hard to justify spending that money without a firm grasp of the magnitude of the sums of money involved (the value of what's at risk) and the likelihood of loss.

There is already a growing body of work relating to quantifying the likelihood of loss occurring (for example, The Digital Archiving Graphical Risk Assessment Model [DiAGRAM] from the National Archives [1]). Unfortunately, it's not so easy to value the data at risk. The knock-on effects of data loss (reputation loss for instance) are even more challenging to quantify in monetary terms. Often the sums involved only become apparent a long time *after* a disaster has happened.

We know that destructive data disasters have already happened to others (inevitably one might argue). Some recent headlines illustrate this:

- *Server crash takes out rich digital archive at Memorial University* [2]
- *Victoria University of Wellington accidentally nukes files on all desktop PCs* [3]
- *PASIG 2017: "Sharing my loss to protect your data" University of the Balearic Islands* [4]
- *University loses 77TB of research data due to backup error* [5]

This means that there is (potentially) data extant that would give an insight into the problem. Such data, if suitably anonymised/redacted, could form the foundation of a "Cost of failure" publication showing how devastating the impact and cost of real-world data loss can be. Regrettably (and quite understandably), those who have suffered this type of loss are rarely willing to acknowledge the fact let alone talk about the numbers involved so we have little insight into the true extent of their losses.

With this in mind, The Digital Preservation Coalition (DPC) and Jisc set out to provide a means whereby individuals and organisations could with confidence and anonymously provide information about the extent and cost of any significant data loss events to a partnership of two trusted organisations (namely Jisc and the DPC). The intention is also to provide a mechanism to collect lessons learned and mitigation strategies.

The aim is to collect examples, from both the UK and overseas, from a range of sectors to represent the customer / membership bases of both Jisc and DPC—for example Higher Education, Research, Public Sector, GLAM, private sector—in order to

highlight why organisations should invest in digital preservation. To help them make a sustainable business case with credible exemplar data.

To achieve this end an anonymous survey was created using the Jisc On-Line Survey Tool [6] and published in February 2022. Both Jisc and the DPC publicised the survey in the following weeks/months. At the time of writing the survey is still open and collecting data. It is intended that the first iteration will close at the end of April. The survey results are to be used to create a publication for launch at iPres 2022.

The survey results and the subsequent publication are, however, only part of the story. Knowing how and why disasters happened and the magnitude of the problem in monetary terms doesn't necessarily mean that the information is useful. It needs to be coupled with strategies to use that information, strategies to mitigate, and strategies to prevent. Above all, the individuals that make up the community need to know about these disasters and how to bring them to the attention of the appropriate people in their organisations in such a way as to ensure that they are acted upon.

And that's where the proposed workshop comes in.

II. THE WORKSHOP

This half day workshop is intended to be a forum where those who have lost data can exchange information with those who would very much like to avoid having their own data loss disaster. The workshop will be run under Chatham House Rules allowing participants to share information freely.

There will be three strands of discussion:

- Past events with invited speakers to set the scene. Up to four speakers each offering a short insight into:
 - The cause of their disaster
 - The magnitude of their disaster—how much data was lost, how much it cost them.
 - Mitigations—what they wish they'd had in place, what they've put in place since.
- There will also be an opportunity for ad-hoc contributions from the floor (attendees will be encouraged to come prepared to share their insights)
- An introduction to the Jisc/DPC survey
 - the preliminary results from the survey

- how to submit information to the survey
- using information from the survey report
- Discussion exploring the usefulness of such data loss events to the community. The direction of the dialogue will be dictated by those in attendance, but possible areas for discussion include:
 - Are they useful? If not, why not?
 - What could the community do to maximize their benefit / avoid losing the insights
 - Strategies for using disaster case studies to drive policy and business cases

The discussions from the session will be written up (in a suitably anonymised form) and fed back into the Jisc/DPC cost of data loss outputs.

REFERENCES

- [1] The Digital Archiving Graphical Risk Assessment Model (DiAGRAM) from the National Archives—
<https://nationalarchives.shinyapps.io/DiAGRAM-dev/>
- [2] Server crash takes out rich digital archive at Memorial University—
<https://www.cbc.ca/news/canada/newfoundland-labrador/mun-digital-archives-wiped-out-1.4787960>
- [3] Victoria University of Wellington accidentally nukes files on all desktop PCs—
<https://arstechnica.com/gadgets/2021/03/university-of-wellington-accidentally-deletes-files-on-all-desktop-pcs/>
- [4] PASIG 2017: “Sharing my loss to protect your data” University of the Balearic Islands—
<https://blogs.bodleian.ox.ac.uk/archivesandmanuscripts/2017/09/27/pasig-2017-sharing-my-loss-to-protect-your-data-eduardo-del-valle-university-of-the-balearic-islands/>
- [5] University loses 77TB of research data due to backup error—
<https://www.bleepingcomputer.com/news/security/university-loses-77tb-of-research-data-due-to-backup-error/>
- [6] Cost of Data Loss Survey—
<https://jisc.onlinesurveys.ac.uk/cost-of-data-loss>

REGISTERING OUR PRESERVATION INTENTIONS

A collaborative workshop on digital preservation registries

Ross Spencer

Ravensburger AG
Germany
ross.spencer@ravensburger
.ag
[0000-0002-5144-9794](tel:0000-0002-5144-9794)

Paul Wheatley

Digital Preservation
Coalition
UK
paul@dpconline.org
[0000-0002-3839-3298](tel:0000-0002-3839-3298)

Euan Cochrane

Yale University Library
United States
euan.cochrane@yale.edu
[0000-0001-9772-9743](tel:0000-0001-9772-9743)

Kate Murray

Library of Congress
United States
kmur@loc.gov
[0000-0003-1325-0829](tel:0000-0003-1325-0829)

Andrew N. Jackson

The British Library
UK
Andrew.Jackson@bl.uk
[0000-0001-8168-0797](tel:0000-0001-8168-0797)

Francesca Mackenzie

The National Archives
UK
francesca.mackenzie@nationalarchives.gov.uk
[0000-0002-1863-7662](tel:0000-0002-1863-7662)

Registries that describe the technical context within which our digital data resides make up a critical part of our digital preservation infrastructure. This workshop seeks to bring together those involved in developing, supporting and utilizing preservation registries along with the wider community of users and contributors. It aims to provide a space for discussion on the future of the preservation registries landscape, identifying gaps in provision, understanding changing user needs, and exploring opportunities for collaboration.

Keywords - File formats, Preservation Registries, Preservation Tools
Conference Topics - Innovation, Community

I. THE PRESERVATION REGISTRY SPACE

The digital preservation community has long sought to capture and describe information about file formats, software, environments and the wider technical infrastructure it operates within. It has worked to develop registries in which to record this information and provide access to it in ways that might support the understanding, rendering, use and ultimately the preservation of digital data. This journey has been long and at times a somewhat rocky road. Some repositories have fallen by the wayside. But the community has learned a great deal

from “doing” over the long term. The current landscape of preservation registries has benefitted from this learning and shows much promise for the future. The following list includes some of the most well known preservation registries but is certainly not exhaustive:

- PRONOM [1]
- WikiData [2]
- COPTR [3]
- LoC Sustainability of Digital Formats [4]

II. A COLLABORATIVE REGISTRY WORKSHOP

This workshop will seek to bring together a number of key groups and individuals involved in the development, maintenance, feature enhancement and everyday use of our preservation registries. It will seek to foster communication and collaboration between these groups in order to support and advance our digital preservation capability. Specifically it will aim to:

- Provide an update on the state of the art of current preservation registries
- Explore potential gaps in the preservation registry landscape (and how they might be filled)

- Foster collaboration between our registries and utilization of them, in particular looking at how registries might better cater for new and innovative uses

The workshop will aim to support those playing a range of roles relating to preservation registries, including:

- Registry developers and maintainers
- Vendors and other software developers whose applications utilize registries in the delivery of preservation functions
- Other registry users, including those active in connecting registries and their data in novel and helpful ways

III. WORKSHOP STRUCTURE

The workshop will use a lightweight structure with the aim of being adaptable to the conversations and needs of the attendees. It will begin with an opportunity for those involved to update the group on the current status of their registry work, whether as part of maintaining a registry, contributing data to a registry or utilizing the data in registries to deliver preservation value. It will then apply an unconference approach to shape the remaining agenda, most likely with the use of break out groups to allow a number of different challenges to be explored through discussion, collaboration and possibly even hackathon type experimentation on the day.

IV. WORKSHOP OUTPUTS

There will be three key outputs from the workshop:

1. The results of interactions and idea generation which will aim to spark innovative and beneficial registry development and exploitation post-iPres.
2. Follow up meetings/workshops/forum to coordinate further collaboration and discussion as appropriate (with support offered from the DPC).
3. A write up of the current state of the art of preservation registries with a focus on supporting practitioners in seeking information from preservation registries, which will take the form of a DPC Technology Watch Guidance Note.

REFERENCES

- [1] PRONOM, [Online], <https://www.nationalarchives.gov.uk/pronom/>
- [2] Wikidata for digital preservation, [Online], <https://wikidp.org/>
- [3] Community Owned digital Preservation Tool Registry, [Online], <https://coptr.digipres.org/>
- [4] Sustainability of digital formats, [Online], <https://www.loc.gov/preservation/digital/formats/>

UNDERSTANDING AND IMPLEMENTING PREMIS

A tutorial

Karin Bredenberg

Kommunalförbundet
Sydarkivera
Sweden
karin.bredenberg@sydarkiv
era.se
[0000-0003-1627-2361](tel:0000-0003-1627-2361)

Eld Zierau

Royal Danish Library
Denmark
elzi@kb.dk
[0000-0003-3406-3555](tel:0000-0003-3406-3555)

Michelle Lindlar

TIB Leibniz Information
Centre for Science and
Technology
Germany
michelle.lindlar@tib.eu
[0000-0003-3709-5608](tel:0000-0003-3709-5608)

Abstract – This half day tutorial will provide participants with a short introduction to the PREMIS Data Dictionary [1] as well as to basic methods of implementations. The goal is for attendees to have a basic understanding of what PREMIS is, how digital preservation metadata can be used in processes and how the data dictionary can be implemented in workflows and systems.

Keywords – Preservation metadata, Preservation strategies and workflows; systems, and tools; Case studies, best practices and novel challenges; Training and education

Conference Topics – Community, Resilience.

I. INTRODUCTION

The PREMIS Data Dictionary for Preservation Metadata is the international standard that provides a key piece for digital preservation activities, playing a vital role in enabling the effective management, discovery, and re-usability of digital information. Preservation metadata provides provenance information, documents preservation activity, identifies technical features, and aids in verifying the authenticity of digital objects. PREMIS is a core set of metadata elements (called “semantic units”) recommended for use in all preservation repositories regardless of the type of materials archived, the type of institution, and the preservation strategies employed.

The PREMIS Data Dictionary was originally developed by the Preservation Metadata: Implementation Strategies (PREMIS) Working Group in 2005 and revised in 2008 and 2015. It is maintained by the PREMIS Editorial Committee and the PREMIS Maintenance Activity is managed by the Library of Congress [2].

PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into several commercial and open-source digital preservation tools and systems.

We have seen a constant call for PREMIS to undertake tutorials, such as this, as more and more organizations come to grips with digital preservation.

The tutorial aims at developing and spreading awareness and knowledge about metadata to support the long-term preservation of digital objects.

II. FORM OF THE TUTORIAL

We prefer an on-premise tutorial, since the interaction with the audience can provide a much better result.

However, although it is recommended to avoid a hybrid tutorial, we will gladly present it as such. It will be with focus on the participants in the room, but where a video transmission can allow virtual participants to follow the tutorial and ask questions e.g. in a google doc document or via a chat. We encourage collaborative note taking by the participants, which is a task easily shared by on-site as well as virtual participants.

III. SUMMARY OF THE TUTORIAL

This tutorial provides in its first part an introduction to PREMIS and its data model and an examination of the semantic units in the Data Dictionary organized by the entities in the PREMIS data model, objects, events, agents and rights.

The second part of the tutorial presents how the preservation community can use PREMIS metadata

support tools for the implementation of software, repository systems and data management practices.

The third part of the tutorial presents high level examples and case studies of PREMIS implementation, using PREMIS in XML and PREMIS in RDF, in relation to the PREMIS Ontology.

Throughout the tutorial we will include examples of implementation experiences which are built upon the institutional experiences of the tutors.

The tutorial includes an exercise section. If time permits, these will be run and discussed during the tutorial session. However, priority is given to questions asked regarding the data model and basic PREMIS functionality. If high audience interaction throughout the main part of the tutorial leaves no room for exercises to be run during the allocated time slot, only a brief introduction to the exercises is given and they will be given as take-home exercises. In either scenario, answers to the exercises are included in the materials.

IV. CONTENT OUTLINE

The draft outline for the tutorial is outlined below.

Introduction to PREMIS community and support

- 1) Background (brief history and rationale of PREMIS)
- 2) Benefits of implementing PREMIS
- 3) Website, PIG, id.loc.gov

Implementation

- 1) Outline of main Entities
- 2) Data Dictionary
- 3) Ontology

Implementation case studies

- 1) PREMIS in METS
- 2) PREMIS in XML
- 3) PREMIS Conformance and repository interoperability

Wrap up and exercises

Time for questions and comments is planned between the different sections of the outline.

V. EXPECTED LEARNING OUTCOMES

Participants will understand:

- 1) What PREMIS is and why it exists;
- 2) The benefits of implementing PREMIS;
- 3) The nature of the existing PREMIS community;
- 4) The critical role PREMIS plays in the digital preservation community.

In addition, participants will get insight into:

- 1) How PREMIS may be used in conjunction with METS;
- 2) How different organizations implement PREMIS within their own repositories;
- 3) How PREMIS, deals with Semantic Web Technology

VI. INTENDED AUDIENCE

This tutorial is aimed at those who want to learn about PREMIS as the tool for selecting, designing, planning, managing, or as a part of a preservation project or repository using preservation metadata. This includes digital preservation practitioners (digital librarians and archivists, digital curators, repository managers and those with a responsibility for or an interest in preservation workflows and systems) and experts of digital preservation metadata and preservation risk assessment.

REFERENCES

- [1] PREMIS Editorial Committee. 2015. PREMIS Data Dictionary for Preservation Metadata. Accessed 2021 located at <http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf>. Web archived: archive.org,, archive time: 2017-02-10 06:23:29 UTC archived URL: <http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf>.
- [2] PREMIS website. Accessed 2022. Located at <http://www.loc.gov/standards/premis/index.html>.
- [3] METS website. Accessed 2022. Located at <http://www.loc.gov/standards/mets/index.html>

CYBER RESILIENCE

Protecting your most valuable asset!

Greg Hewitson

Dell

United Kingdom

Greg.Hewitson@Dell.com

I. INTRODUCTION

Data that is at the heart of ransomware attack. Thanks to the rapid digital transformation happening today, data has become an organization's most critical asset.

Attackers know this very well and target their attack to gain access to your critical data.

When they have access to your data, attackers can do multiple things. They can remove access to your own data by encrypting the contents with a key. They also attack data protection techniques to make sure that all restore capabilities are deleted. This way they increase the probability that you'll pay for the for the key to unlock your data again. This is known as an encrypting ransomware attack. If the value of the content is high, they can also ask to pay a ransom to prevent this content from being shared. OR they can use this information for espionage and try to remain in stealth modus as long as possible (like SolarWinds attack).

Conference Topics – resilience

BUT NOT ALL DATA IS EQUAL!

II. DESCRIPTION

What is Cyber Resilience and what's best practice. Here we'll take a deeper dive at the bigger picture.

Cyber Resiliency = The capacity for an organisation to protect, detect, respond to and recover from a cyber-attack with minimal impact. It's clear that every organization will have a different cyber resilience strategy, which makes creating a cyber resiliency strategy not an easy task. Different frameworks have been created to guide organizations in building their own cyber resiliency strategy.

We like to utilize the Cyber Security Framework (CSF) from the National Institute of Standards and Technology (NIST). This framework can help to create a common language between the board of directors, your IT administrators, your partners, and your customers during the complex task of creating and maintaining a cyber resiliency strategy. The NIST CSF consists of 5 pillars for 5 core functions of the framework:

Identify: understand what you have in your organization and assess the risk. Define a target security profile of your organization.

Protect: how are you going to defend what you have against the known bad

Detect: detect suspicious behavior, infiltrations, and breaches

Respond: how to respond when a detection is made?

Recovery: how to get your business back up and running.

Every organization should identify its target profile by going through the list of security controls and identifying

We'll then take the audience through how to understand their current security profile and what their target security profile steps are to mitigate Cyber threats to their critical assets.

III. TARGET AUDIENCE

This topic is best suited to C-Suite level, IT Storage managers and data preservation officers.

REFERENCES

- [1] University of Texas at San Antonio
Written story: <https://www.delltechnologies.com/asset/en-us/products/storage/customer-stories-case-studies/delltechnologies-customer-story-the-university-of-texas-san-antonio.pdf>
- [2] Video Testimonial: <https://www.delltechnologies.com/en-us/video-collateral/h18929-cv-powerscale-u-of-texas-san-antonio.htm>
- [3] Short Social Video: <https://www.delltechnologies.com/en-us/video-collateral/h18930-cv-powerscale-u-of-texas-san-antonio-quotes.htm>
- [4] Perspectives Customer Story (Kendra Ketchum):
<https://www.delltechnologies.com/en-us/perspectives/how-wartime-medic-kendra-ketchum-became-a-tech-trailblazer/>
- [5] PowerProtect Cyber Recovery for Sheltered Harbor, written story: <https://www.dell.com/en-uk/dt/data-protection/cyber-recovery-solution.htm#tab0=0&pdf-overlay=https://www.delltechnologies.com/asset/en-uk/products/data-protection/briefs-summaries/powerprotect-cyber-recovery-for-sheltered-harbor-solution-brief-h18199.pdf>

WRITING BINARY BY HAND

An introduction to binary file formats

Martin Hoppenheit

Landesarchiv Nordrhein-Westfalen

Germany

martin.hoppenheit@lav.nrw.de

Abstract – In this tutorial participants will learn how to approach (i.e., read) a file format specification and based on this specification create binary files by hand. This is useful when learning about, researching and experimenting with file formats as well as to create test or example files.

Keywords – file format, binary

Conference Topics – innovation; resilience

I. INTRODUCTION

Binary file formats are often believed to be complicated, inapproachable, and somewhat arcane. This is because, other than text-based formats, they require specialized, sometimes proprietary viewers and editors to interpret their structure and content.

While it is possible to open any binary file in a hex editor, the vast sequence of hex characters that it throws at its users does not always help to establish understanding – hex-encoded gibberish is still gibberish.

This tutorial follows another approach: Instead of looking at existing binary files, participants will create them from scratch. Starting from a file format specification, they will write small example files by hand, thus gaining a practical understanding of the file format and a basis for further exploration and experiments.

A blog post based on the tutorial is available [1].

II. TUTORIAL OVERVIEW

After a short introduction explaining the basics of binary files, hexadecimal notation and the tools used in the tutorial, participants will be pointed at the essential parts of the TIFF file format specification [2] and create a minimal TIFF file along the way. After that, they will be able to adapt and extend their example files diving deeper into the TIFF

specification. Alternatively, they may move on to other format specifications and create other example files themselves.

Although binary files will be created writing hex code (like in a hex editor), their readability will be greatly improved using a notation and tool called Literate Binary [3]. This notation allows combining binary (hex) and textual content in a single text file (Markdown, to be precise) from which both a binary file and corresponding documentation in formats like HTML or PDF can be generated. This is particularly useful when documenting example files.

III. GOALS

Participants will gain a general idea of the mechanics of binary file formats. They will see that a binary file is not just a random sequence of bytes but structured data.

Participants will realize that file format specifications can be surprisingly readable and there is no need to shy away from them. Reading a format specification is useful not only to create example files but also to understand error messages from file format validation.

Participants will learn to create binary files by hand. This helps when learning about a file format, but it also allows reproducing problems in the form of minimal examples (without additional content blurring the crucial aspects or violating copyright) or providing test files for file format identification and validation tools.

IV. INTENDED AUDIENCE

Everybody can participate; there are no strict prerequisites. Programming skills are not required. It helps though if participants have a basic

understanding of hexadecimal notation (e.g., as described in [4] or from working with PRONOM signatures) and are not afraid of occasional command line use. To get most out of the tutorial participants should bring a laptop with their favorite text editor and the Literate Binary tool installed.

REFERENCES

- [1] Writing binary by hand, Martin Hoppenheit. <https://martin.hoppenheit.info/blog/2022/writing-binary-by-hand/>
- [2] TIFF 6.0 Specification, Aldus Developers Desk. <https://archive.org/details/TIFF6>
- [3] Literate Binary. <https://github.com/marhop/literate-binary>
- [4] Representing binary data, Martin Hoppenheit. <https://martin.hoppenheit.info/blog/2017/representing-binary-data/>

SCALABLE CURATION OF EMAIL WITH OPEN-SOURCE TOOLS:

Review, Appraisal, and Triage of Mail (RATOM)

Christopher Lee

*University of North
Carolina
USA*

Kam Wood

*University of North
Carolina
USA*

Email has served, and continues to serve, as the communication and recordkeeping substrate of numerous contemporary phenomena across all sectors of society. A small but essential portion of the email has continuing value and warrants long-term preservation. Many collecting institutions have acquired email accounts and collections. Unfortunately, the pipeline from email acquisition into digital preservation environments is still relatively immature. One of the fundamental challenges has been that formats for packaging email accounts (particularly PST) can be brittle, complex and dependent on proprietary software. Another fundamental challenge is that information within email can be subject to various sensitivities (e.g., protected health information, financial information) that collecting institutions must address. Without reliable and scalable tools to address these two challenges, email risks being locked into relatively hidden collections. Quite the opposite of date for all!

The creation of Archival Information Packages (AIP) depends on a variety of functions including, but not limited to: extracting email content from proprietary packages; locating messages with retention value; identifying instances of sensitive and personally identifying information; and accurately tagging features of email messages, including those corresponding to real-world entities such as persons, places, organizations, and events.

The Review, Appraisal, and Triage of Mail (RATOM) project presented in this tutorial began as a two-year effort to develop and test software and workflows to support the review and processing of email in collecting institutions. The RATOM project developed software to scan email archive files (including PST, OST, and mbox) and record and export content, metadata, and derived features such as entities identified using natural language processing (NLP) into a simple SQLite database that can be queried as part of a larger set of digital curation workflows.

The RATOM tools are designed to minimize the effort required to run computationally complex email analysis tasks. For example, RATOM provides a single-command tool that makes it simple to replace the default NLP model used to identify entities with a model for a different language, a custom model, or a multi-language model. While some existing LAM access systems incorporate NLP to describe the contents of collections, this technology is often tightly coupled to the platform being used or is applied strictly to file types that tend to share common structures and metadata.

Email contains encoded text, markup, and attachments, but importantly also structured metadata in the header that can be used to cue identification of persons and organizations and describe their relationships. Identifying entities, relationships, and other features of interest by processing open text from heterogeneous

collections of files (such as those extracted from a disk image) is inherently “noisier,” as the extracted text will often contain patterns of features (such as persons, places, and organizations) common to a wide range of devices and production environments (e.g. documentation of system files). By exposing header metadata features in database entries where they are explicitly linked to entities identified in open text, the RATOM tools provide a mechanism by which cross-format search procedures can be easily implemented.

In this tutorial, participants will learn to work with the core RATOM tools, including the email processing library and associated command-line utilities. Using publicly available corpora including PST files from the Enron collection, participants will explore the different options provided by libratom and its utilities to extract content and metadata from email backup files, scan content for entities of interest, and query the SQLite database it produces as output. The tutorial will also include an introduction to selecting and working with pre-trained spaCy language models and provide participants with a clear understanding of which models are appropriate for which tasks and use cases.

No programming experience or prior experience with command-line tools is required. Participants may bring a laptop with Windows 10 or 11, macOS 11 or 12, or modern Linux distribution (Ubuntu 22.04LTS or later) to fully participate. Instructions for installing the software in advance may be found at <https://github.com/libratom/libratom/blob/master/README.md>. However, it is not mandatory for participants to run the software on their own machines. The tutorial will be conducted in part using remotely hosted interactive notebooks and a web-based SQLite database browsing utility.

LABDRIVE TUTORIAL

A Research Data Management and Digital Preservation Platform

Teo Redondo

LIBNOVA
Spain
teo.redondo@libnova.com
0000-0001-6465-7771

Antonio G Martinez

LIBNOVA
Spain
a.guillermo@libnova.com

Maria Fuertes

LIBNOVA
Spain
mfuertes@libnova.com

Abstract – LABDRIVE is a Research Data Management and Digital Preservation platform resulting from the ARCHIVER Project. It allows organizations to capture the research data they produce, helping them to properly manage, preserve and allow access to it, during the whole research data lifecycle. The purpose of this tutorial is to introduce the main features of LABDRIVE as well as explain how it works through a tutorial (a guided demonstration).

Keywords – Research Data Management, Digital Preservation

Conference Topics – Innovation; Exchange.

as well as associated tools (datasets, software tools, etc.).

With LABDRIVE, R&D organizations can keep the research data they produce for the long term, in a single platform. Researchers can manage their research datasets with the best tools, adopting good practices for digital preservation and also keeping code and data together in one single platform during the lifecycle, independently of functionality, protocols and featured needs.

I. INTRODUCTION

LABDRIVE [1] is a Research Data Management and Digital Preservation platform powered by LIBNOVA that focuses on scientific datasets. LABDRIVE allows organizations to transition from a siloed approach in which each series of datasets, departments or units is using multiple, disaggregated systems to keep content to a single repository that can adapt to the particularities of each dataset, unifying all content into a single platform.

The platform works for organizations both with a few gigabytes of data, to organizations managing several petabytes. Digital preservation principles are always present, so Data protection comes first. The platform is fully aligned with OAIS, ISO16363, and presents a variety of redundant checks and processes for safeguarding valuable research data.

LABDRIVE is primarily oriented towards research-intensive scientific and academic institutions that need to preserve research projects, working objects

II. MAIN CHARACTERISTICS OF LABDRIVE

A. Metadata-driven, virtualized scalable storage

- Digital Preservation Administrators can assign a specific Storage Policy to each Data Container in the platform (storage types, replicas, technologies, providers and integrity policies) to use at data container level.
- A single repository supports multiple storage providers and types (for very high volumes of content).
- Transition from one storage policy to another (even from a storage provider to another), fully managed by the platform.
- Virtualized storage so file paths remain unchanged when the underlying storage technology is changed.
- Extensible storage architecture (cloud object storage, CEPH, tapes, etc.).

B. Code-driven, advanced content management

- LABDRIVE lambda functions can be defined by the organizations (or integrators) so the platform automatically processes the content using the logic defined.
- C. *Easy to use and powerful*
- Equally capable web interface and API, so users can easily manage the platform while power users can automate every process.
- D. *Strong digital preservation technology*
- Digital preservation principles always present: Data protection comes first.
 - Fully aligned with OAIS, ISO16363, redundant checks and safe processes.

III. TUTORIAL CONTENT

The contents would be divided into 3 blocks and would be roughly as follows:

- A. *LABDRIVE Introduction*
- Architecture and overview
 - How research content is to be organized
- B. *LABDRIVE Configuration*
- Users and permissions
 - Archival organization
 - Container – concept and usage
 - Metadata configuration
- C. *LABDRIVE Operations*
- Create a data container
 - Upload content
 - Download content
 - Introduction to metadata – concept and usage
 - Searching
 - File versioning and recovery
 - Working with data containers
 - LABDRIVE functions
 - Storage mode transitions
 - Advanced operations – Jupyter Notebooks & API usage

REFERENCES

- [1] LIBNOVA LABDRIVE Public Documentation
<https://docs.libnova.com/labdrive>

CONTINUOUS IMPROVEMENT TOOLS FOR DEVELOPING CAPACITY AND SKILLS

A Tutorial

Sharon McMeekin

*Digital Preservation Coalition
Scotland*

sharon.mcmeekin@dpconline.org
0000-0002-1842-611X

Jenny Mitcham

*Digital Preservation Coalition
United Kingdom*

jenny.mitcham@dpconline.org
0000-0003-2884-542X

Amy Currie

*Digital Preservation Coalition
Scotland*

amy.currie@dpconline.org
0000-0001-9099-8457

Abstract - The ability to apply a carefully considered and well implemented approach to continuous improvement of digital preservation capabilities can greatly benefit practitioners when looking to set and achieve objectives. This tutorial aims to provide attendees with the skills and tools to develop and implement a methodology for continuous improvement at their organization using resources developed by the Digital Preservation Coalition.

Keywords - Maturity modelling, skills, good practice, continuous development, benchmarking
Conference Topics - Resilience, Community.

current capabilities, set future targets, and plan for developments to meet those targets.

As part of their member support activities, the Digital Preservation Coalition (DPC) has created a number of resources to facilitate the continued development of digital preservation capabilities within an organization. These include the DPC Rapid Assessment Model¹ (DPC RAM) and the forthcoming DPC Skills Framework (to be published Spring 2022). These two resources are the focus of the proposed tutorial.

I. INTRODUCTION

Digital preservation cannot be a static activity. Ensuring the longevity of digital content requires proactive management and maintenance of the organizational and technological infrastructures we deploy. But how best to structure this management and maintenance to ensure its success?

Since the early days of digital preservation, the community of practice has sought ways to benchmark an organization's capabilities. An audit and certification approach was the original method championed, however, in recent years the more flexible approach of maturity modelling has started to gain popularity. A maturity model provides a framework for assessing the level of capability of an organization across defined areas relating to policy, processes, procedures, and infrastructure. Maturity models allow an organization to understand their

II. SUMMARY OF THE TUTORIAL

This aim of this tutorial is to empower practitioners by providing them with the tools and skills required to plan, advocate for, and assess their progress with developing digital preservation capabilities within their organization.

It will begin by providing them with a solid understanding of the importance and benefits of a continuous improvement approach to benchmarking their digital preservation capabilities. Following this, attendees will be introduced to and led through two practical exercises:

1. Using DPC RAM to assess an organization's capabilities with reference to policy, processes, procedures, and infrastructure.

¹<https://www.dpconline.org/digipres/implement-digipres/dpc-ram>

2. Carrying out either an individual or organizational skills audit using the DPC's Skills Framework and Audit Toolkit.

As well as practical exercises, attendees will be encouraged to engage with live polling to allow benchmarking of digital preservation maturity of the organizations represented within the tutorial cohort.

The tutorial will finish with an overview of other DPC resources that can help practitioners with planning and advocating for their digital preservation activities.

III. CONTENT OUTLINE

The following is a draft outline of the tutorial content, including proposed timings:

1. Intro. to Continuous Improvement (c. 30mins)
 - a. What is continuous improvement?
 - b. Benefits of a continuous improvement
 - c. Introduction to continuous improvement tools from the DPC
2. Focus on DPC RAM (c. 60mins)
 - a. Introduction to DPC RAM
 - b. Exercise: completing a DPC RAM assessment
3. Break
4. Focus on the DPC Skills Framework (c. 60mins)
 - a. Introduction to the DPC Skills Framework and Audit Toolkit
 - b. Exercise: completing a personal or organizational skills audit
5. Feedback and Wrap-Up (c. 30mins)
 - a. Overview of DPC resources to support continuous improvement
 - b. Tutorial feedback

IV. INTENDED AUDIENCE

This tutorial will benefit individuals and organizations from across many sectors who wish to assess their current digital preservation capabilities and plan for future developments. It will also benefit researchers wishing to incorporate an understanding of these processes into their work, and educators who hope to expand or enhance their curricula on the topics covered.

V. LEARNING OUTCOMES

Tutorial attendees will be able to:

3. Explain the importance of continuous improvement
4. Plan their approach to continuous improvement
5. Complete a DPC RAM assessment for their organization
6. Describe the skills required for digital preservation
7. Undertake a skills audit of digital preservation staff at their organization

VI. SHORT BIOGRAPHIES OF ORGANIZERS

Sharon McMeekin is Head of Workforce Development at the DPC, and her role includes leading training and skills projects, and acting as managing editor of the 'Digital Preservation Handbook'. Sharon is an archivist and experienced practitioner and has contributed to a number of international training and development projects in digital preservation. She is a frequent guest lecturer for information management courses, and is a trustee of the Scottish Council on Archives.

Jenny Mitcham is Head of Good Practice and Standards at the DPC where she engages in a range of projects to develop good practice resources for digital preservation. This has included a project working closely with the UK Nuclear Decommissioning Authority, during which she co-created DPC RAM. Jenny has worked in digital preservation for nearly two decades, having previously held roles at the Archaeology Data Service and the University of York.

Amy Currie is Training and Grants Manager at the DPC, where she works on the development of digital preservation training and skills projects and manages the Career Development Fund. She completed her PhD at the University of Glasgow in 2021, where she previously worked as a teaching assistant and co-convenor in the Information Studies department.

USING ePADD FOR EMAIL PRESERVATION

Implementing the ePADD+ Project Enhancements

Ian Gifford

University of Manchester
United Kingdom
ian.gifford@manchester.ac.uk

Sally DeBauche

Stanford University
United States
debauche@stanford.edu

Tricia Patterson

Harvard University
United States
tricia_patterson@harvard.edu

Abstract – For decades, email has been a ubiquitous and critical record of humanity. The ePADD+ project enhanced the open source email archiving tool, ePADD, with new features supporting effective long-term preservation, including additional format eligibility, retention of original order, format normalization, preservation and provenance metadata, and export for archival repositories. In this tutorial, attendees will learn about the new functions offered in ePADD and how they are implemented in the workflows of the ePADD+ partner institutions. Authors will do a live demonstration of the tool, utilizing the new functions and guiding attendees through optional hands-on participation. The authors will engage attendees in a broader discussion around the future development road map for ePADD and strengthening the sustainability of the community supported tool.

Keywords – email, preservation, open source, collaboration, sustainability

Conference Topics – Innovation, Community

I. INTRODUCTION

Email has persevered as a cornerstone of communication for decades and, consequently, a critical individual, institutional, and cultural record of our time. As collection policy and capacity evolve, email records are proliferating in archival collections. Originally developed by Stanford University in 2015, ePADD is a popular open-source tool in the digital archives community supporting appraisal, processing, discovery and delivery of email collections. In 2021, Stanford, Harvard University, and the University of Manchester began collaborating on a grant-funded project to embed preservation functionality into ePADD in order to consolidate the number of tools needed to steward email records. The resulting enhancements enable the tool to package collections for proactive long-term archival care. The enhanced ePADD tool has a more robust set of preservation metadata users can

track and record, as well as the optional inclusion of a complete set of original email, preservation copies, and dissemination copies. As part of the Email Archives: Building Capacity and Community regrant program, the ePADD+ project additionally sought to strengthen the sustainability of the open source, community supported project.

In this tutorial, Harvard, Manchester, and Stanford will demonstrate ePADD's new features and discuss how these are utilized in their own institutional workflows for preservation of email records. The workshop will walk attendees through a hands-on exercise of generating a preservation-ready export from ePADD, with optional participation from attendees that have an instance of ePADD installed. It will conclude with a facilitated conversation around sustaining ePADD as a community-supported tool, prioritizing enhancements, and expanding the community of code contributors.

II. TUTORIAL AGENDA

A. Overview

The authors will open the tutorial with an overview of the ePADD+ project, introducing the project's objectives, partners, and outlining the resulting preservation enhancements with which ePADD is equipped.

B. Incorporating ePADD into Preservation Workflows

Each partner institution (Harvard, Manchester, and Stanford) will introduce their email preservation workflows. This will provide attendees with a model for how to incorporate ePADD into different

institutional contexts, including the customization of preservation packages in accordance with the institutions' distinct preservation repositories.

C. *Demonstration and Hands-On Exercise*

Once the attendees have an understanding of the new tool functions and have synthesized the implementation of these functions, the authors will do a live demonstration of ingest, appraisal, and export of a preservation-friendly package from the ePADD Appraisal module. Participants that have a recent download of ePADD on their laptops can optionally follow along with the demonstration to gain hands-on experience utilizing the preservation functions.

D. *Discussion*

The tutorial will conclude with a facilitated discussion about the future sustainability and development of the ePADD tool.

III. ACKNOWLEDGEMENTS

The ePADD+ project is supported by the University of Illinois (#098468-18280) as part of its Email Archives: Building Capacity and Community (EA:BCC) initiative, funded by the Andrew W. Mellon Foundation (#1905-0768).

REFERENCES

- [1] "EPADD," *Stanford Libraries*. [Online]. Available: <https://library.stanford.edu/projects/epadd>.
- [2] "RFC 8493 - The Bagit File Packaging Format (v1.0)," *Document search and retrieval page*. [Online]. Available: <https://datatracker.ietf.org/doc/html/rfc8493>.

AUTOMATED TOPIC MODELLING IN ARCHIVES PORTAL EUROPE

Kerstin Arnold

*Archives Portal Europe
Germany
kerstin.arnold@archivespor
taleurope.net*

Marta Musso

*Archives Portal Europe
United Kingdom
marta.musso@archivespor
taleurope.net
0000-0002-3728-3548*

Kostantinos

Stamatis

*Archives Portal Europe
Greece
kostas.stamatis@archivespor
taleurope.net*

I. INTRODUCTION

Archives Portal Europe (APE, www.archivesportaleurope.net) is the portal of European archives, an aggregator that connects on a single research point the catalogues and digitised archival material of all archives in and about Europe. It currently hosts material from more than 30 countries and in 24 languages, and from a variety of archival institutions (such as State and city archives, university and parish archives, private institutions, etc). In order to navigate this extremely heterogeneous material, one of the research tools made available by Archives Portal Europe is by “topics”, curated collections in which each archival institution participating into the APE project can tag its documents according to a specific topic. Because topics are maintained manually by the archivists, and because of the vast amount of archival material ingested in the portal, it is impossible to have a comprehensive body of topics that describe the whole of the APE repository.

In this scenario, automated topic detection can be a fundamental tool to guide archival research and to allow archives to be accessible to potentially world-wide users, in a situation where national and linguistics barriers blur, or are re-defined.

II. SUMMARY OF THE TUTORIAL

This workshop presents the creation of an AI tool for automated topic detection in the APE corpus, a vast, inhomogeneous, and multi-lingual collection of historical archival catalogues – the first such project

to be designed for archival descriptions rather than corpora of specific documents.

The development is based on supervised machine learning, with a combination of human inputs in different languages (collectively-created taxonomies for each topic), and the usage of Wikipedia pages to model the relevant vocabulary and entities. The first iteration of the algorithm was tested on a sample of 9 topics in 5 languages, and the second iteration enlarged the sample to 13 topics and 12 languages, for a total of more than 500,000 descriptive units, and it also introduced Boolean operators and wildcards.

The workshop will explain how the tool was built, and will allow users to test it live, gathering feedback on its usability and possible future implementations outside of the specific corpus of Archives Portal Europe.

Index

Abaseaca, Ciprian	460	Burland, Tamsin	479, 501
Addis, Matthew	397	Byrne, Helena	433
Aguilar, Fernando	460	Caldrone, Sandi	202
Aït El Mekki, Touria	374	Caron, Bertrand	28,329
Alföldi, István	184	Cassidy-Amstutz, Andrew	473
Anderson, Bethany	202	Cazeaux, Hugues	220
Anderson, David	184	Chadash, Daniel	225
Anderson, Janet	184	Chartier, Madison	162
Anderson, Seth	40	Cherrington, Corey	392
Andriamahady, Dina	220	Cochrane, Euan	40, 225, 503
Arnold, Kerstin	519	Collins, Claire	40,231
Arp, Laurie	208	Conroy, Sarah	53
Aucock, Janet	462	Cothey, Viv	231
Bähr, Thomas	293	Crabtree, Jonathan	406
Bailey, Jefferson	211	Currie, Amy	99, 515
Baker, James	495	Cushing, Amber	53
Barnes, Miranda	399	Darby, Kristy	473
Bartliff, Zoe	491	Davies, Kevin	456
Beck, Sue	458	Day, Michael	436
Beking, Angela	429	Day Thomson, Sara	345,349,489
Bell, Alistair	433	de La Houssaye, Jordan	28
Bell, John	404	de Vries, Denise	240
Beneito Arias, Paloma	447	DeBauche, Sally	517
Benoit III, Edward	431	Descheemaeker, Pauline	173
Bernstein, Daniel	493	Dillo, Ingrid	406
Bredenberg, Karin	215, 244, 329, 505	Disher, Tara	392
Brown, Michael	300	Dongrong, Zhang	193
Bruys, Alix	28	Doyle, Phoebe	53
Burchmore, Tess	53	Dumont, Jean-Noël	271
Burgi, Pierre-Yves	220	Elkiss, Aaron	244

Index

England, Elizabeth	424	Hale, Meredith	162
Espenschied, Dragan	250	Halvarsson, Edith	311
Falcao, Patricia	489	Hanson, Karen	81
Faria, Luís	489	Harsanyi, Regina	404
Fernandes, João	397	Hart, Michaela	378
Ferreira, Miguel	61	Haunton, Melinda	441
Ferriter, Meghan	145	Hawes, Anisa	349
Forbes, Megan	208	Healy, Sharon	433
Fortin, Émilie	438	Heesakkers, Driek	413
Fox, Claire	40	Hegarty, Niamh	53
Fraimow, Rebecca	402	Henry, Joshua	202
FRANÇOIS, Robin	255	Hewitson, Greg	507
Friha, Lamia	220	Hibberd, Lee	443
Fuertes, Maria	513	Higgins, Christopher	260
Gates, Ethan	40	Higgins, Sarah	260
Gates, Syreeta	409	Hiromatsu, Takeshi	445
Genest, Elisabeth	440	Hobbs, Mark	413
Gentry, Steven	162	Hofsink, Inge	244
Giaretta, David	460	Holdzkom, Elizabeth	473
Gieschke, Rafael	70	Hoppenheit, Martin	509
Gifford, Ian	517	Hricikova, Andrea	497
Gillesse, Robert	371,481	Huay Ho, Mui	413
Gooding, Paul	491	Hui, Liu	193
Gow, Ann	491	Ikeuchi, Ui	445
Dumont, Jean-Noël	374	Imker, Heidi	202
Greenberg, Jonathan	81	Inácio de Oliveira, Vagner	235
Griffith, Arran	493	Ippolito, Jon	404
Groen, Arie	371,481	Ito, Shinsuke	445
Guimarães, Miguel	61	Jackson, Andrew	266, 503
Ha Lee, Jin	392	Jakeway Manchester, Eileen	145

Index

Jensen, Mathias	487	Luong, Hoa	121, 202
Joguin, Vincent	271	Ma, Xiao	250
Johnston, Leslie	329	MacGregor, Rachel	451, 495
Jones, Shawn	277	Mackenzie, Francesca	497, 503
Kaminski, Jaime	184	Maeda, Yukio	445
Kelly, Rebecca	53	Makhlouf Shabou, Basma	220
Kim, Yunhyong	491	Markus, Katharina	359
Kinnaman, Alex	89	Martell, Allan	431
Klein, Martin	277	Martinez, Antonio	513
Knight, Jessica	413	Martinez, Ruby	410
Konstantelos, Leo	282	Martins, Dalton	235
Kotov, Andrey	497	May, Peter	453, 456
Kramer-Smyth, Jeanne	447	McGann, Morgan	53
Krause, Jan	220	McMahon, Kevin	499
Kristensen, Rasmus	487	McMeekin, Sharon	99, 515
Kufeldt, Payton	53	McNally, Anna	495
L'Hours, Hervé	406	Mears, Jaime	145
Laakso, Mikael	388	Meiser, Linda	220
Lake, Sarah	287	Middleton, Sarah	479, 501
Lange, Sebastian	311	Milic-Frayling, Natasa	108
LaPlant, Lisa	356	Mitcham, Jenny	305, 315, 515
Ledoux, Thomas	28	Mooney, James	311
Lee, Christopher	511	Morris, Lidia	392
Lee, Jamie	409	Moulds, Lyndsey Jane	250
Lehtonen, Juha	244	Munshower, Alan	89
Leigh, Alexandra	173	Murray, Kate	503
Lin, Yi-Ting	449	Musso, Marta	519
Lindlar, Michelle	293, 329, 505	Nagasaki, Kiyonori	445
Love, Valerie	300	Nappert, Mireille	438
Lowry, James	409	Narlock, Mikala	121

Index

Naumann, Kai	499	Romkey, Sarah	329
Nef, Andreas	244	Rossi, Giulia Carla	467
Neugebauer, Tomasz	287	Rufian, Mikel	460
Noonan, Daniel	458	Russell, Alan	443
O'Sullivan, Jack	329	Sağlık, Özhan	470
Oberhauser, Jurek	40	Salas, Camille	473
Ormond, Cian	53	Sanchez, Crystal	319
Paro Costa, Paula	235	Schaefer, Sibyl	422
Paterson, Clare	282	Schlarb, Sven	215
Patterson, Tricia	517	Schmalz, Marc	388, 392
Peurt, Laura	451	Schmid, Katharina	433
Pelan, John	483	Seals-Nutt, Kenneth	153
Peltzman, Shira	424	Seroka, Lauren	473
Pennock, Maureen	436	Sierman, Barbara	475
Petters, Jon	497	Simpson, Kathryn	491
Phelps, Kathryn	497	Skødt, Asbjørn	477
Pohlkamp, Svenia	293	Smith, Caylin	489
Popham, Michael	315	Smyth, Tom	324
Potter, Abigail	145	Snijder, Ronald	388
Proença, Diogo	184	Snyder, Kylie	392
Prom, Christopher	410	Spencer, Ross	503
Pyke, Tegan	467	Stamatis, Kostantinos	519
Quine, Gerard	53	Steeman, Marjolein	329
Ramírez-López, Lorena	402	Steinke, Tobias	244
Ranade, Sonia	418	Steinmeier, Daniel	366
Reba, Martina	53	Stewart, Garth	334
Rechert, Klaus	40,70	Stokes, Paul	479, 501
Redondo, Teo	460, 513	Strathmann, Stefan	293
Rippington, Sean	462,465,483	Suárez, Juana	402
Rochat, Rebecca	255	Takeo Magruder, Michael	108

Index

Talboom, Leontien	418	Wong, Kevin	485
Tallman, Nathan	340	Woods, Kam	511
Tarver, Hannah	162	Woods, Ronan	53
Tèrmens, Miquel	460	Work, Lauren	424
Thiele, Julia	460	Yan, Emma	282
Thornton, Katherine	153	Zarnitz, Monika	293
Thorsted, Tyler	353	Zierau, Eld	487, 505
Tieman, Jessica	356		
Tindall, Alexis	378		
Trei, Kelli	202		
Tunnat, Yvonne	359		
Underdown, David	173		
van den Eijkel, Susanne	366		
Van Den Hurk -Van't Klooster, Eva	371,383		
van Zwol, Tamara	371,383,481		
Vasseur, Édouard	374		
Verhoff, Deb	81		
Vízner, Pamela	402		
Walls, David	356		
Wang, Hannah	340		
Weatherburn, Jaye	378		
Wendler, Robin	244		
Wheatley, Paul	305, 503		
White, Rachel	162		
Wijsman, Lotte	371.383,481		
Williams, Sarah	202		
Wilson, Audrey	483		
Wilson, Carl	184,215		
Wise, Alicia	388		
Wittmann, Rachel	162		