Document Distribution Note
Release to Members: Immediate
Release to Public: 06/03/2014

iPres 2013

Lisbon, 2-6th September 2013

About the event

The annual international iPres conference was hosted in 2013 by INESC-ID in Lisbon, parallel to the Dubline Core conference. The week-long conference saw something like 150 delegates and included a significant number of ancillary workshops, tutorials and meetings. AD, SMM, SN and WK represented the DPC and numerous members were also present: Neil Grindley and Paul Stokes (JISC); Lee Hibberd (NLS); Maureen Pennock (BL); Alex Ball and Kevin Ashley (DCC); Aileen O'Carroll, Sharon Webb and Natalie Harrower (DRI); Tim Keefe (TCD); Janet Delve and David Anderson (Portsmouth University); Andrew Wilson (personal member); Tom Heritge (BBC); Anna Henry (Tate); Paul Wheatley (Leeds Univ); Simon Waddington (KCL); Catherine Jones (RCUK/STFC); …

These notes are intended to provide an informal briefing for members of the DPC not able to attend in person. They only represent the sessions that WK was able to attend. For an authoritative and comprehensive report, readers are encouraged to contact the organisers or speakers directly.

Workshop 1: From Preserving Data to Preserving Research – the Curation of Processes and Contexts, Monday 2nd September

Angela Dappert (DPC) – Introduction to Digital Preservation of Processes and Contexts

Angela introduced the workshop noting that 2 projects - Workflow4Ever and TIMBUS realised that they were building tools and approaches that are complimentary to each other, with similar motivations though to different sectors. TIMBUS starts by observing the overlap between business continuity management and digital preservation. The two are isomorphic and throw up similar problems – such as how much data must be captured before an escrow service can operate effectively. If you want to re-deploy processes at a given point in time for business continuity purposes then you face similar challenges if you wanted to capture and redeploy it into the future. Angela introduced a simple business process workflow associated with musical classification which was being used as a model to test others.

Discussion – assessing a research process for preservation is very similar to doing peer review properly – yes it's just good research and it blurs the boundary and that's a good thing; who does the work – shouldn't this become a research data management question?; there's a lot more people involved however. There's also a question of motivations and making sure that people get credit for this work. This needs to be addressed.

Rudi Mayer (SBA) – Preserving Processes

Rudi expanded the use case and described how to preserve scientific research processes. To preserve a process I firstly need to describe the context in which it is embedded – legal, technical and social aspects. Fundamentally this is an issue of representation information, though existing approaches to rep information are not really designed for this: so firstly we need a meta-model for

the description of representation information associated with a process. The example use case – music classification – is a machine learning example so it's actually quite widely applicable. It involves some regular steps – acquisition of underlying data, ground truth of genres, extract features from the data source in a numerical form, train the machine versus on the ground truth then let the machine work through a test data set. The results of the analysis can then be assessed manually: repetition with slight variations and presumably enhancements as time goes on. A few questions to ask about preservation – are there external services to preserve (and am I allowed to preserve them); what are the steps and do they all need to be preserved; what are the dependencies between processes; how are the processes configured; where did the ground truth come from; etc? Motivations for the preservation of the process are simple in research contexts and multiple layers – personal, institutional, external and so forth. The context model exists in many different layers so we need to put some kind of boundaries around it. A meta-model can guide that, for example, by capturing features that are not domain specific, described in domain independent ontologies, as well as locating features that are domain specific. Archimate was selected in TIMBUS as an enterprise architecture modelling language: archimate has three core concepts 'passive' structures, 'active' structures and 'behaviours'; and archimate is layered so allows an approach which includes technology, application and business and it has 32 fundamental concepts which are building blocks for describing domain independent features. The domain independent ontology allows us to build up a coarse description of the context of the business process. They are now building set-up extractors to help build the context description automatically. This is relatively easy for the technology configuration which is easily the most complex set of dependencies; it's less simple for the business process dependencies though these tend to be described formally in any case.

Discussion: If I am a 3rd party and want to re-use your data I need only to refer to a portion. After a while this breaks down. Where are the boundaries? There's a paradox that preserved data is widely used, but not necessarily high value: high value data needs to be preserved to ensure it's used more often. What does the thing that preserves all this stuff look like? A single repository or a federated set of them.

Kevin Page (Oxford e-Research Centre, University of Oxford) – Research Objects

Research objects are essentially aggregations of resources that bundle together the contents of research work – data, experiments, examples, bibliography, workflows and so forth. Defining the boundaries of a research object involves looking at some of the existing places and services which are already in use – like packs in MyExperiment…

Raul Palma (Workflow4ever) – Demo of tools to produce a RO from Workflow4ever

Demo …

Stefan Pröll (SBA) Data Citation

Publication and citation for traditional scholarship is well understood – data citation is encouraged but not necessarily always well done. Data is an essential part of research and there are many reasons for calling for data as a form of publication itself. Citation of data encourages reproducibility transparency, documentation, context, identification, impact and reuse. But citing data needs

unique identification for subsets and complete data sets. We need machine-readable and human-readable metadata and we need clear incentives (and mandates) for making data citable. There are standard and widely used identifiers like ISBN or ISSN, but mostly data is shared by URI / URL. More complex systems like DOI supplement this by providing a resolver service. DataCite and ARK exist to provide persistent identifiers for data, though they provide different types of service: for example ARK allows you to generate subsets of data and is free, while DOI has a large commercial base. There are 3 related initiatives in this space that are worth looking at: Codata committee on data citation standards; Force 11 manifesto of recommendations on scholarship; and the RDA working group on data citation. So far data has to be static to be cited: but the question arises how to reference live (transactional) data. This could be done by storing not the data but by assigning a persistent identifier to the query that generated the data. Storing the SQL query would allow a much lighter preservation layer in terms of data though it's also much more complex and assumes that the data doesn't change (or that a strong change history is maintained – a bit like Memento?)

Catherine Jones (STFC) Research Context Preservation in Scape

STFC is a research council with large scale scientific facilities and has significant data lifecycle management challenges. STFC has control of researchers' data for extended parts of the lifecycle. Currently it maintains a very substantial data store and a metadata catalogue. Current plans are to build a 'Investigation research object' from this catalogue which in turn creates the potential for a data journal with a variety of linking and validating tools. This data journal is assisted by the fact that STFC is a funder with unique access to resources – which means significant leverage in terms of data sharing.

Session 1: Preserving Digital Objects, Tuesday 3rd

Andrew Lindley (Austrian Institute of Technology) Database Preservation Evaluation Report - SIARD vs. CHRONOS

Database preservation is surprisingly under-represented in digital preservation tools and literature and it's not even very prominent in the research. In 2012 AIT had was invited to evaluate 2 different approaches to preserving database – SIARD and CHRONOS. The goal of this paper is to broaden discussion on database preservation by comparing one of the most popular tools within memory institutions (SIARD) and within industry (CHRONOS). SIARD is a format and a tool for software independent archiving of relational databases from the Swiss national archives; CHRONOS is a commercial tool owned and developed by an SME called CSP which is used in automotive and financial institutions. Typically tools to archive databases are not included in commercial (and open source) relational databases, even though typically large quantities of data (perhaps 85%) stored in rdms are no longer needed. So there is a strong rationale for thinking about preserving databases, even just to optimise current systems. Side by side comparison of functionality suggested that CHRONOS had appropriate functionality for database retirement, continuous and partial archiving and application retirement; SIARD had functionality for database retirement. Side by side comparison of performance in export suggests that CHRONOS has a focus on exporting primary data and datatypes so is able to export a wider range of outputs than SIARD; SIARD focuses on preserving primary data and exclusively supports core SQL 1999 elements and only tabular content is restored.

In terms of pre- and post-processing via DB scripts suggests that CHRONOS interacts with the database via shell commands and scripts whereas SIARD does not allow you to write to the database at all therefore does not allow scripting and messaging within the database. Similarly SIARD has no functionality to deliver data retention controls while CHRONOS delivers this either directly with its own tools or by integrating with other tools: so CHRONOS enables data security policies to be implemented more easily. In terms of user roles and management, CHRONOS has a mature rights and access management layer which includes GUI interfaces for operators whereas SIARD relies on the user and rights management tools that may be (or may not be) contained within the archived system. Access and performance are maximised for use of the underlying data in CHRONOS via a middleware layer that provides views onto an imported version of the archived data whereas SIARD depends on the access and performance of the underlying data and only limited tools for access. CHRONOS enables tools to track syntactic and semantic changes in its middleware layer which is not possible in SIARD. It's important to remember that SIARD's functionality is more narrowly defined than CHRONOS so it's not a direct comparison, but the conclusion should be obvious….

Discussion – what about price? CHRONOS requires a one-off fee that is affordable …

Tom Heritage (BBC) File Based Preservation of the BBC's Videotape Archive

More recent generations of BBC's archives – the D3 video tape for example – are among the more 'at risk' of the BBC's collections. D3 tapes are being digitised into uncompressed MXF formats on TTO tapes meaning that the BBC's archival outputs will be digital. The process has taken 6 years. Ingex software is the core of the process controlling a lot of the workflow. It's essentially quite a simple workflow though the scale of the process is enormous so the workflow has to be really robust to ensure that all the possible errors are detected. The individual files that are produced can be huge so need to have a really strong QA – a single file error could create significant loss. Simply reading and moving files can be inordinately time consuming so it's necessary to check frame by frame within the file for errors. Since 2007 the BBC has improved massively its capacity to be compliant with OAIS.

Klaus Rechert et al (BWFLA – University of Freiburg) Large Scale Curation and Presentation of CD-ROM Art

Digital Art provides an attractive entry point for digital artists and for access to art, but digital culture is practice rather than a series of artefacts: 'there is no form outside of practice'. So a different approach is needed – a reduced amount of rich simulated environments that enable interaction of and with many artefacts. BWFLA is working on the idea of emulation as a service and this forms the basis of the approach. The ingest and access workflow was presented. https//demo.bw-fla.uni-freiburg.de/ (username 'bwfla', password 'demo')

Session 2: Digital Repositories

Eld Zierau (Royal Library of Denmark) Applying OAIS to Distributed Digital Preservation

OAIS assume that work in preservation should be distributed in a variety of ways but it doesn't really enable or encourage distributed preservation actions. As a community we have become quite good

in talking about collaboration but tend to talk about individual repositories and work in small groups. A number of case studies have been developed in distributed digital preservation – Internet Archive, Chronopolis, DuraCloud, MetaArchive, Archivematica and BitRepository which all have elements of distributed digital preservation.

Keynote: Digital Information Storage in DNA, Wednesday 4th

Paul Bertone (European Biomatics Institute)

This keynote started with a relatively minor but interesting piece of research at EBI. EBI is like CERN except for biology, and it has a mission to undertake research and to provide and encourage access to information. EBI manages lots of different types of information and lots of different routes and uses for that data. Biology is changing, represented in part with a huge deluge of data: a typical experiment today can produce 25 times as much data as the entire Human Genome Project did. EBI can no longer really do this as a single institute so a new infrastructure – Elixir – is being developed for data management and sharing with data management delegated. Storing data on media is notoriously transient so any large scale storage initiative – even one for fixed archival data - carries a massive cost in terms of refreshment, and the issue becomes more and more expensive the more transient the data. Now DNA can encode information and it is incredibly long-lived. The information stored is fundamentally biological but it can be manipulated to store and encode other types if information. More recent experiments show that this is not an experimental feature: within 10 years it would be possible to store some very large datasets in some (very very) small strings of DNA: the whole of Google in a cardboard box. The costs are high for the synthesis and you really don't want to have to read the data too often – but that means it's like tape rather than flash. Under proper conditions DNA has an indefinite lifespan: small short strings of DNA are incredibly robust even if they break down at chromosomal level. It's also incredibly small. Imagine the National Archives was more like a seed bank.

Discussion –

- where do errors occur: they are more likely to occur at synthesis stage rather than at the reading stage;
- is it better to store in the 'junk DNA' of a living organism or stable biological material: better in stable biological material because gene mutation happens in a random manner

Session 1: Co-operation in digital preservation

Yvonne Friese (Goportis) - Benefits of geographical, organizational and collection factors in digital preservation cooperations: The experience of the Goportis consortium

GOPORTIS is a network of four agencies in Germany which share their preservation architecture and work together for a range of themes. This is now working well they there are a clear series of barriers and risks to collaboration like this – capacity, compatability, priorities and practical commitment. Collaboration on a geographical basis – like Florida. Being based in the same country simplifies that referring to legislation issues; and where there is a union catalogue (or some other existing shared infrastructure) then workflows can be simplified. Where there is an organisational

symmetry – like MetaArchive – then you can widen your own horizons and capacities by swapping good practice more directly and you use similar types of vocabulary making this process relatively simple. But different organisational cultures can sometimes be hidden and quite surprising when they are discovered. Finally collaboration based on collection types – like the PrestoCentre – can create a truly global community. This reduces the overall cost for collaboration; and the fact that material is similar but not identical which improves the collaboration. Goportis is based on all three of these factors – geography, institution and collection types.

Maite Braud (Tessella) ENSURE: Long term digital preservation of Health Care, Clinical Trial and Financial Data

ENSURE project is looking for economical solutions to long-term preservation, with a significant interest outside the normal memory institutions that we normally meet. The drivers for preservation are different. In health care there are specific issues of privacy (data protection), large scale of data, regulation, legislation and obsolescence. Many of these are familiar even if the data protection, regulation and legislative issues are distinctive: so there is a lot of potential to share our knowledge with them. Financial services have a greater dependencies on in-house applications than health care, strict retention (and deletion) schedules and large amounts of data in continuous streams. Ensure has established a high level architecture which allows for evaluation and configuration of preservation actions and it integrates (it assumes) a cloud based architecture. Data on the cloud can be vulnerable and the obvious solution – encryption – is problematic for preservation as encryption can be brittle over time. We need to ensure the security of data if we want to acquire the benefits of cloud storage.

APARSEN Panel Session – Peter Doorn, Rene van Horik and others

Rene described the outline of the APARSEN NoE; Barbara Sierman outlined APARSEN's three test audits using the three step European framework; Simon Lambert introduced some of the potential infrastructure services that could be developed – such as planning, representation information and such; Sharon McMeekin presented the outcomes of recent research on training and plans for the near future. David Giaretta talked about how the EC views preservation and asked how DP could add value to users, arguing that access to research data would bring the biggest return on investment. Carlos Morais-Pires reflected on the EC's view of the future of digital preservation research and services in Europe.

Keynote: Digital Information Storage in DNA, Thursday 5th

Carlos Morais Pires (European Commission E-infrastructure DC CNECT C1)

Carlos is in charge of the data-infrastructure element of the Horizon 2020 Programme (H2020). H2020 is a major programme with many millions of euros and it is outward looking with a lot of engagement with the US and Australia.

It used to be the case that computing resources and data were scarce and one had to manage transactions to minimise impact. Although this has changed tremendously it's important to remember that fundamentally e-science is still science: we must not lose sight of the purpose and

nature of the scientific processes. But at the same in the 21st century, science harbours emerging features that challenge the way that scientists work. Educators, students and even curious citizens can be involved. For that to occur, data is a keyword. By data the EC understands recorded factual material commonly accepted in the scientific community as necessary to validate research findings. To support this we need data infrastructures – services, applications tools, knowledge and policies for research data to be discoverable, understandable, accessible, curated and preserved. In a sense the physical and technical infrastructure become less visible and data becomes the infrastructure for science. These technical developments are supplemented by policies that reinforce the European research area, that support access to scientific information and policies that specifically talk about preservation – cf commission Recommendations on Access and Preservation of Scientific Information (2012). Something like 1bn euros in the next 7 years for ICT aspects of research infrastructures. The size of the funding resources available to inform the development of data-infrastructures seems large but it is relatively small in comparison to the size of the sector and the total investment from national and local funders. The e-infrastructure has three basic layers: linking, sharing and collaborating. There are barriers to achieving this: for instance to reconcile funding cycles with the variable timescales of technological development; how to achieve interoperability when ICI are composed of distributed parts requiring local developments and optimisation; top-down planning versus more chaotic community based change and so forth. Ultimately the development of infrastructure is influenced by social, economic and technological dimensions: it requires interaction and engagement of different stakeholders to take their share of responsibility and contribute to lead the way. The Research Data Alliance is a good way to mix the strategic planning (top down) while engaging a diverse community. Homeless data quickly becomes no data.

H2020 is about to launch a new cycle of R&D&I. The following areas are in scope for 2014: connectivity, data, computing, core services, skills, virtual environments and so forth. E-infrastructure should be included in call 3 with 9 specific elements for things like

- managing preserving and computing with big research data – development and deployment of integrated secure on-demand service-driven and sustainable infrastructure
- e-infrastructure for open access – robust e-infrastructure supporting open access policies in Europe by providing reliable and permanent access to scientific records
- Common infrastructure policy and practice – a small amount of resource specifically designed to support the research data alliance and its community
- Pan-european high-performance computing infrastructure and services – proviging access to the best
- Centres of excellence for computing applications – establishing a limited number (8-1) centres of excellence for the application of HPC in scientific and industrial domains focussing on scientific industrial and social challenges
- Network of HPC centres
- Providing core services across e-infrastructure – support to harmonise and/or deploy core infrastructure (e.g. EduRoam) which support interoperation of research communities across Europe (imagine a single library account across Europe)
- Virtual research environments

- New professional skills for e-infrastructures – enabling the development of curricula for information literacy in science to use and deploy the e-infrastructure

Discussion –

- Calls are not necessarily ordered numerically – so it could happen that Call 5 could come before call 3. Call 3 should open towards the end of this year and likely close in april. There will be a big bank of calls in the first year so need to manage the capacity.
- March 26-28 2014 in Dublin is the next plenary of the RDA
- There will be some focus for knowledge transfer into small companies
- Science is described quite narrowly – shouldn't it also include social sciences and the humanities (Yes!)

Workshop: Preservation At Scale, Thursday 5th

(no power for the start of the workshop)

Discussion: the introductory presentations were mostly about optimisation not scale, so what is the nature of scalability and how do we achieve it.

2 parts to the problem – academic libraries sometimes behave as memory institutions to ensure long term access, but that's not always their job: the motivation to support students and researchers is much stronger but that's really about continuity of access to data sets / publications which might be better looked after elsewhere. Need to understand the drivers and make clear decisions with an understanding of the organisational strategy.

What's the problem? What is the scale issue? Various reasons:

- We're not preserving anything like enough stuff. There's probably about three or four times as much e-journal content that we need to protect.
- Scalable solutions – organisations are large and have complicated management structures that need to be scalable
- We keep hearing about big data
- Users have growing expectations –
- Heterodoxy of material
- Granularity of content – from holding statements down to specific commas in publications.
-

About this document

Version 1	Written at conference	8/11/2012	WK
Version 2	Distributed	8/11/2012	DPC members

Document Distribution Note
Release to Members: Immediate
Release to Public: 06/03/2014