



Preserving the past: the challenge of digital archiving within a Scottish Local Authority

Lorraine Murray

0909613m

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in
Information Management and Preservation in the Humanities Advanced Technology and
Information Institute, University of Glasgow, September 2017

Contents

1	Abstract	1
2	Methodology	4
3	Introduction	6
4	The changing role of the Archivist	11
5	Inverclyde Archives	15
6	Challenges of born digital accessions within Local Authority	21
6.1	NRS digital preservation Skills for the Future projects	23
7	Setting the scene: portable document format [PDF]	29
8	Solving the PDF Preservation Problem: DPC Student Project	31
8.1	Trialing the veraPDF toolset	36
8.2	Project conclusions and recommendations	42
9	Conclusion	45
10	Bibliography	
11	Appendices	

Acknowledgements

I would like to thank Yunhyong Kim, Adele Redhead and Ann Gow from the Humanities Advanced Technology and Information Institute at The University of Glasgow for all their support and patience throughout this project.

I would also like to thank Sara Day Thomson, Sharon McMeekin, William Kilbride and Paul Wheatley from the Digital Preservation Coalition for help and advice received during the DPC Student Project.

Thanks also to Susan Corrigan from the National Records of Scotland for her kind permission to reproduce unpublished primary sources relating to the work of Penny Wright and Ruth Marr in connection with the NRS digital preservation Skills for the Future projects.

Additional thanks goes to Alana Ward, Libraries, Archives and Museum Manager at Inverclyde Council for allowing for the use of primary source material during the veraPDF trial and to Janice Miller, Chair of ASLAWG for her insights and recommended sources.

Finally, thanks also to Lucy Stock, Sean MacMillan and Vincent Gillen for proof reading, comments and suggestions, and to my family for their patience during my research.

1. Abstract

Local Authority [LA] Archives reside within an already challenging environment of legislative obligations, corporate compliance, resource limitations and budgetary cuts. Due to advances in technology and increasing user expectations of requiring information in a more immediate and accessible way, this has exacerbated the already difficult task of providing an Archive service within a local authority framework. LA Archives have a duty of care to their collections regardless of the format of the material contained therein. The statutory requirements of preserving a digital collection in perpetuity is a much more challenging prospect than preserving a physical collection; digital objects are far more fragile and often come with greater and more complex preservation risks when compared to their analogue counterparts.

This dissertation discusses the evolving role of Archivists and Information Professionals with reference to the fairly recent inclusion of born digital objects being collected by Archival repositories. It will identify some of the challenges faced by LA Archivists who must address the prospect of having to increasingly accession these digital assets. It includes a brief discussion about the National Records of Scotland [NRS] Skills for the Future digital preservation projects; “Counting the Bits – Local Authority Capacity Planning” and “Good Foundations – Local Authority Digital Preservation Guidance” which highlight the challenges of digital preservation within a local authority environment.

Drawing from my experience as a local authority Archivist for Inverclyde Council; Inverclyde Archives has been chosen as an example of one such Scottish local authority

Archive service faced with the challenges that accompany the onset of caring for digital assets in their custody.

In addition to identifying some of the issues of being responsible for a digital archive, the next logical step is to investigate whether any of these issues are likely to pose a risk to the sustainability and accessibility of the digital collection in future. Given the large scope of digital file formats which may reside within an Archival collection, one commonly used file format has been chosen as a starting point to discuss possible preservation risks. Within the context of Inverclyde Archives, the three most common digital file formats within the collection are PDF (portable document format), Microsoft Word documents and JPEG files.¹ As PDF files are considered one of the most robust and prevalent digital file formats both on the web and within many organisations,² the focus will be on identifying any potential preservation risks present in PDFs. Not all PDF files are created equally; the PDF/A version is deemed appropriate for Archival storage³ and is the preferred version accepted by many large Archival institutions worldwide such as the Library of Congress in the USA.⁴ Less so in the UK; in the case of the Archaeology Data Service [ADS], the preferred textual document format accepted for digital ingest is whatever format was used to create the digital file as “PDF content is often downsampled during the PDF process, leading to loss in the original data

¹ Information supplied by Inverclyde Archives, August 2017

² Duff Johnson, The 8 most popular document formats on the web <http://duff-johnson.com/2014/02/17/the-8-most-popular-document-formats-on-the-web/>

³ <https://www.loc.gov/preservation/digital/formats/fdd/fdd000318.shtml>

⁴ <https://tools.ietf.org/html/rfc3778>

streams.”⁵ In the instance where the original created file is not available, PDF/A is preferred by ADS.⁶

However, many document creators, users and those who are responsible for these types of files are not aware of the different varieties and flavours of PDFs available, nor that some could pose more of a preservation risk than others.

As part of this dissertation research, I took part in a student project run by the Digital Preservation Coalition [DPC] in partnership with the veraPDF Project⁴ to trial a set of new PDF validation tools.⁷ Running the veraPDF validation software allows the user to investigate whether their PDF files conform to ISO 19005 standards⁸ (i.e. if the files are PDF/A compliant) and can help identify any files showing evidence of possible preservation risks. This knowledge allows the user to make decisions on how best to mitigate these issues in an attempt to preserve the PDF files for future use.

⁵ <http://archaeologydataservice.ac.uk/advice/FilelevelMetadata.xhtml#Documents>

⁶ <http://archaeologydataservice.ac.uk/advice/FilelevelMetadata.xhtml#Documents>

⁷ <http://verapdf.org/project/>

⁸ <https://www.iso.org/standard/38920.html>

2. Methodology

Within this dissertation, the following research questions will be investigated:

- What are the main challenges faced by local authority Archives with reference to caring for their digital collections?
- Would limiting the ingress of PDF files to those of the PDF/A variety to the collection be the best approach to help meet legislative responsibilities?
- Does saving other digital file formats to PDF/A incur any loss of original content, appearance, meaning, authenticity or value of the original file?
- What is more important within an Archival repository; the appearance or the content of the digital file?

In an attempt to answer these questions, a variety of methods will be used. An analysis of the recent NRS digital preservation Skills for the Future projects will be employed to gauge the current issues which exist within local authority repositories. Additionally, statutory regulations required by local authorities to preserve their collections will be considered, drawing sources from relevant current legislation, advice from the NRS, and any recommendations for fulfilling these obligations based on current digital preservation advice, particularly sought from the seminal “Digital Preservation Handbook” produced by the Digital Preservation Coalition.⁹

⁹ <http://dpconline.org/handbook>

For the purposes of this research, an investigation of preservation risks that may exist within PDF files was carried out using the veraPDF validation tool. The outcomes from this study will be discussed along with any conclusions drawn from the results achieved.

Furthermore, the authenticity of a digital file and whether the nature of a digital object means it should be considered any less valuable or reliable than an equivalent analogue object will be mentioned. Additionally, the issue of content vs the appearance of a digital file will be addressed; with reference to Jeff Rothenberg's paper "Preserving authentic digital information"¹⁰ as quite often either content or appearance has to be sacrificed to maintain future accessibility of a digital object.

Information and an understanding of the topic will be sought from a variety of sources such as the Digital Preservation Coalition [DPC], the veraPDF Consortium, the COPTR registry, the Digital Curation Centre [DCC], Scottish Council on Archives [SCA], Archives and Records Association [ARA] and from publications by Authors such as Jenkinson, Schellenberg, Cook, Cox, Duranti, Dearstyne, Kuny, Theimer and Barata amongst others.

It is hoped that by employing a varied methodological approach, this integration will provide a fuller understanding of the topics raised during the course of this research.

¹⁰ <https://www.clir.org/pubs/reports/pub92/pub92.pdf>

3. Introduction

In his novel *Nineteen Eighty-Four* which was first published in 1949, George Orwell wrote “Who controls the present controls the past.”¹¹ A political statement of the time perhaps, but nevertheless an accurate observation based on how he perceived the future - particularly so when considering the role of memory institutions whilst carrying out their remit as custodians of the historical record. Record-keepers are placed in the role of helping to define the past based on what they choose to retain or dispose of. The notion of being responsible for acquiring, appraising, storing, preserving and making information accessible for future generations is enormous and must not be considered lightly. Leaving aside any discussion on the politics of the collecting process or perceived bias this well-known quote may infer, the issue being considered here is simply that collecting, appraising, storing, preserving and making available information within the context of an Archive is a multifaceted task which carries with it great responsibility for any Archivist, Records Manager or Information Professional.

Records are created for many reasons; usually to support the business needs of that organisation and allow for their core functions to be carried out. Records are valuable in many ways; they can act as evidence and provide documentation of processes, to show accountability and transparency within that organisation, and to help create a corporate memory. Once records are no longer required for the primary function for which they were created, if deemed worthy of long term or permanent retention, those

¹¹ George Orwell, *Nineteen Eighty-Four* (1949) pp 12

records can be transferred to an appropriate repository to carry out a secondary function as historical evidence. This process of transfer within a local authority setting is a compulsory element from the Public Records (Scotland) Act 2011, Section 1 2(b)(iii), Archiving and transfer arrangements (element 7).¹² That this element is compulsory highlights the importance placed on the process of transferring selected records to a suitable repository for permanent retention due to their enduring value. It is worth noting at this point, that in the context of many Archival repositories, these discrete stages are not necessarily adhered to; the record can be treated within the records continuum model which is defined by the International Council on Archives [ICA] as a “consistent and coherent process of records management throughout the life of records, from the development of recordkeeping systems through the creation and preservation of records, to their retention and use as archives.”¹³ Frank Upward proposed this continuum model in the 1960’s as an ongoing process with recurring stages during the life cycle of the record.¹⁴

Historically, Archives have been seen as a repository for paper based documents and physical items, however due to the furtherance in technology, the way in which we create and keep information has changed. Therefore it has become necessary for Archivists to become the custodians of information deposited from a diverse range of formats; which increasingly includes born digital objects, which is defined by the DPC as “materials which are not intended to have an analogue equivalent”.¹⁵ The landscape of

¹² <https://www.nrscotland.gov.uk/record-keeping/public-records-scotland-act-2011/resources/model-records-management-plan/model-plan-guidance-to-element-7>

¹³ <https://recordsandarchives.wordpress.com/2011/07/14/the-records-life-circle-and-continuum-concerpts/>

¹⁴ <https://recordsandarchives.wordpress.com/2011/07/14/the-records-life-circle-and-continuum-concerpts/>

¹⁵ <http://www.dpconline.org/handbook/glossary#B>

archives is transforming in an attempt to meet the needs of the 21st century, thus meaning the role of the Archivist is expanding - and must continue to do so - to meet these ever increasing and changing demands.

User expectations and consumption of media is changing; there is now a ubiquitous expectation of immediate access to information heralded by the digital age. Therefore, in an attempt to both meet these expectations and to help sustain their service, Archives must consider the benefits of making their digital collections available to a wider audience. This can be achieved by creating digital surrogates to make them more accessible, along with any born digital materials held within the collection.

There is a positive side of doing so; the Archive can benefit from increased user engagement, which in turn can be a demonstrable tool to advocate for additional funding by evidencing an appetite for the collection. This can aid an Archive to be able to justify its existence in a climate where heritage and memory institutions are facing increasing financial restrictions. However, by increasing the digital assets, the Archive is faced with the difficult task of caring for a diverse collection comprising a myriad of digital formats as well as physical holdings. This diversity of format in itself presents all sorts of challenges; a particular issue being the availability of storage space and the need for careful forward planning to allow for the deposit of new accessions in whatever format they are received. Storage costs money whether physical or digital, and in a climate of ever decreasing budgets, this can be rather problematic to say the least.

The DPC handbook states that “Digital materials are a core commodity for industry, commerce and government”¹⁶ and within the context of local authorities, this is becoming increasingly so. It is therefore necessary to realise that digital materials are assets and must be created, maintained and stored in a viable way to ensure their long term preservation. Simply creating a digital file and saving it is not enough, therefore much needs to be done to educate document creators and users about the life-cycle of a digital object, as the creation stage of a record is one of the most important stages when considering digital preservation.

In many cases, software upgrades fail to support legacy files, so to prevent obsolescence (i.e. the situation whereby the digital object is no longer able to be rendered whether due to hardware or software becoming outdated), a digital preservation strategy must be adopted. Two strategies most commonly used are:¹⁷

- Migration – To change the original file to another format in order to retain the significant properties of the file whilst allowing access to the content without necessarily preserving the original appearance and context.
- Emulation – To imitate the original environment of the digital object by using software to recreate and preserve the original look and feel whilst allowing full functionality of the object.

¹⁶ <http://www.dpconline.org/handbook/digital-preservation/why-digital-preservation-matters>

¹⁷ <http://www.dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards>

Current thinking appears to suggest that the migration of text based documents and other media files of various formats is the most effective way forward.¹⁸ Although migration may offer a solution if it allows us to access the content of the file in future, the act of migrating the file format into a different format altogether, means changing it into something new. Within some organisations or environments this may not be an issue, however within the archives sector, this would open up discourse about themes such as originality, authenticity, integrity, and trustworthiness, prompting the question; is the artefact or the content more important? Although these are very important considerations to ponder, due to the scope of this paper and the complex nature of the question, addressing these questions will not be the main focus of this paper.

¹⁸ <http://www.dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards>

4. The changing role of the Archivist

There are 32 Local Authorities which exist in Scotland.¹⁹ All of which have a duty of compliance to various pieces of legislation such as the Public Records (Scotland) Act 2011 [PRSA],²⁰ the Freedom of Information (Scotland) Act 2002 [FOISA],²¹ the Data Protection Act 1998 [DPA],²² the Local Government etc. (Scotland) Act 1994,²³ the Local Government (Access to Information) Act 1985,²⁴ the Re-use of Public Sector Information Regulations (2005)²⁵ and the Environmental Information (Scotland) Regulations 2005 [EIRs].²⁶ Issues of copyright compliance are also most pertinent within an archive service and additionally, local authority Archives must comply with the aims and objectives of their organisation as set out in their corporate mission statement.

The PRSA states that all local authorities must create a records management plan [RMP] specifying the *proper arrangements* for managing the authority's public records which must be approved by the Keeper of the Records of Scotland.

Sections 53 and 54 of the Local Government etc. (Scotland) Act, states that all local authorities have a statutory obligation to make *proper arrangements* to preserve and manage all records – in all formats – which have been created by that local authority and those inherited from any predecessor authorities.

¹⁹ <http://www.gov.scot/Topics/Government/local-government/localg/usefullinks>

²⁰ http://www.legislation.gov.uk/asp/2011/12/pdfs/asp_20110012_en.pdf

²¹ http://www.legislation.gov.uk/asp/2002/13/pdfs/asp_20020013_en.pdf

²² http://www.legislation.gov.uk/ukpga/1998/29/pdfs/ukpga_19980029_en.pdf

²³ http://www.legislation.gov.uk/ukpga/1994/39/pdfs/ukpga_19940039_en.pdf

²⁴ http://www.legislation.gov.uk/ukpga/1985/43/pdfs/ukpga_19850043_en.pdf

²⁵ http://www.legislation.gov.uk/uksi/2015/1415/pdfs/uksi_20151415_en.pdf

²⁶ http://www.legislation.gov.uk/ssi/2004/520/pdfs/ssi_20040520_en.pdf

Additionally, there are a number of British Standards [BS]²⁷ and International Electrotechnical Commission [IEC] standards²⁸ relating to record-keeping which are also worthy of consideration within an Archive setting. Although these international standards are not required to be met by current legislation, adherence to them constitutes best practice:²⁹

- BS 5454:2000 – relates to the storage and exhibition of materials
- BS 4783-8:1994 – relates the care of materials in storage and transportation
- BS ISO 15489-1:2001 – relates to managing information and documentation
- BS ISO/IEC 27001:2005 – relates to information technology and security
- BS ISO/IEC 27002:2005 – code of practice for securing information
- BS 10008:2008 – relates to legal admissibility of digital information
- BS EN 15713:2009 – relates to secure destruction of information

Putting adherence to rules and regulations aside, Archivists are expected to carry out a multitude of tasks, many having the dual job title of Records Manager and Archivist, occasionally Local History Librarian and even some Information Professionals in this field are tasked with the responsibility for compliance and FOISA enquiries.

In Scotland, all local authorities have a professional record-keeping professional in post; however the role of these individuals can vary enormously depending on the

²⁷ <https://www.bsigroup.com/en-GB/standards/>

²⁸ <http://www.iec.ch/about/activities/standards.htm>

²⁹ <http://www.gov.scot/Resource/Doc/933/0124124.pdf>

requirements of the Council in which the particular Archive or repository is set.³⁰ Some institutions are part of the Council, whereas others are run by Trusts as *Arms-Length External Organisations* [ALEOs]. These ALEOs are usually set up by the local authority to provide services such as heritage and leisure and are considered as an external organisation related to the LA rather than actually being part of it.³¹

Common tasks for the record-keeping professional within these organisations include acquiring, appraising, accessioning, cataloguing, storing, preserving, and documenting the Archive collection. Additionally, research, writing policies, managing staff and volunteers, dealing with enquiries, arranging public access to requested items, outreach activities, arranging exhibitions, care and conservation of the collection, advocacy and helping to publicise the service are all important additional tasks very often required with this role. It can be a stressful environment; for the most part due to the demands placed by ever changing priorities which quite often are unique to Archivists within a local authority or ALEO environment.

Moreover, if you add the fairly new task of being responsible for acquiring and caring for a digital collection to this already heady mix, it becomes clear that there is a need for additional resources, training and support to aid Archivists and record-keeping professionals to competently carry out the requirements of caring for their digital assets.

It is not feasible to limit the accession of objects to traditional formats such as paper, photographs and manuscripts due to the perceived difficulties in holding digital assets

³⁰ Information supplied by ASLAWG Chair Janice Miller, Aug 2017

³¹ <http://www.gov.scot/Topics/Government/local-government/localg/whatLGdoes>

- the proliferation and rate of creation of digital records means *like it or not*, the Archivist and record-keeper must be prepared to accept them and care for them with as much consideration and attention to detail as is afforded to more traditional deposits.

5. Inverclyde Archives

Inverclyde Archives is the public Archival repository that exists within Inverclyde Council as part of the local authority, and as such, must comply with the aforementioned statutory requirements and legislation, such as contributing to the implementation of the records management plan [RMP]³² as set out in the PRSA.

Currently, the Archival collection consists of both physical and digital assets and holds in the region of 80 linear metres of physical records with approximately 50GB of digital assets³³. Despite the pervasiveness of digital record creation within Inverclyde Council, the current digital holdings within the Archive predominately consist of surrogates rather than born digital files. However, it is anticipated that a large quantity of born digital material will be accessioned from within the Authority over the next year and beyond due to the type of records being created with the Council and the ongoing implementation of element 7 *Archiving and transfer arrangements* of the RMP as set out in the PRSA.

The nature of local authorities means that there are numerous directorates, services and departments which exist, each carrying out different functions in their remit to deliver services as part of local government.

In Inverclyde Council, there are three directorates³⁴;

³² <https://www.nrscotland.gov.uk/record-keeping/public-records-scotland-act-2011/resources/model-records-management-plan/model-plan-guidance-to-element-7>

³³ CIPFA Statistical Return March 2017 <https://www.cipfastats.net/cipfastats/>

³⁴ <https://inverclyde.gov.uk/assets/attach/2075/.pdf>

- Education, Communities & Organisational Development
- Health & Social Care Partnership
- Environment, Regeneration & Resources

Each directorate is divided into a service which is further sub divided into departments, depending on the corporate structure of each local authority. Some examples of those services or departments are education, property services, regeneration and planning, housing, economic development and environmental health - to name but a few. Each department creates records for different purposes and in different methods, and many of which are dependent on particular proprietary software applications.

For example, most Scottish LA's use an educational management information system called SEEMiS to document everything from pupil and staff records, attendance and pastoral notes.³⁵ At present, it is anticipated that SEEMiS records relating to Inverclyde educational institutions will be transferred to Inverclyde Archives in the foreseeable future. Software dependency and content file formats are anticipated challenges that this transfer will present and must be addressed before any such transfer is allowed to occur.

In 1995, Richard Cox anticipated the concerns that this transition from paper to electronic records would bring; "about the continuing management of such records, access and privacy, the notion of a record".³⁶

³⁵ <https://www.seemis.gov.scot/site3/index.php/about>

³⁶ <http://www.jstor.org.ezproxy.lib.gla.ac.uk/stable/41101910>

Although the transfer of born digital files to Inverclyde Archives thus far has been negligible, the sorts of born digital records expected to be received in future may include the following types:

- Text documents
- Images
- Sound files
- Video Files
- Spreadsheets
- Databases

From this list, each type of record will likely consist of various formats, for example; text documents may be of notepad, Microsoft Word or PDF format. Similarly, images may consist of jpeg, TIFF or PNG files. From this initial observation, it becomes clear that not only may the future digital collection contain numerous categories of records, but within each category, several file formats may exist - each with their own unique requirements.

A pertinent matter to consider is this; upgrades to hardware and software happens regularly and often, therefore the risk of both types of obsolescence is high.

Additionally, some software dependencies may exist which could prove to be problematic in future. Spreadsheets or databases may have been created using proprietary software, which creates an added complication where it becomes necessary to extract the information contained within such a system to allow for future

access without any software dependency. If the content cannot be extracted or understood without the need for the original software in which it was created, the file will eventually become useless and the information lost forever. This is a great risk, and one reason why digital material is considered more fragile than analogue material.

It is therefore absolutely necessary to understand the nature of any digital records which may be received by the Archive, and this starts with the first stage of the records life-cycle; the point of creation. Good communication between the record creator and the Archivist is necessary to allow for preservation of the digital object, capture of metadata, and future access. Without these things, the digital object can lose its value. To illustrate this point, imagine a photograph has been found within an Archive with no accompanying information. Without any context, how can we know who or what the subject(s) is/are, what date the photograph was taken, who the photographer was (and if there are any copyright restrictions for re-use) or indeed why it was taken (as often photographs are taken to document a point in time). Without any context (i.e. metadata), the photograph becomes useless and in the context of an Archive, cannot be reused and has diminished value.

The Digital Curation Centre describes the ongoing processes of digital curation in more detail with reference to the digital curation lifecycle. They state that this lifecycle consists of the following steps:³⁷

- Conceptualise
- Create

³⁷ <http://www.dcc.ac.uk/digital-curation/what-digital-curation>

- Access and use
- Appraise and select
- Dispose
- Ingest
- Preservation action
- Reappraise
- Store
- Access and reuse
- Transform

These steps can be used as a practical guide to assess and create processes and policies within any organisation that looks after digital data. In some instances, all of the steps in the sequence may not be necessary, and in some cases, additional steps may need to be added; however this is an excellent starting point to allow those who care for digital records to consider what steps are necessary when caring for a digital collection.

Within the context of Inverclyde Archives - and indeed all local authority repositories and other archival institutions - identifying the issues faced by collecting digital objects is a necessary first step to addressing the digital preservation challenge.

For the purposes of testing the veraPDF software, a wide range of sample files from within the Archive department were selected to carry out checks using the recently developed veraPDF software. The aim was to identify the extent of preservation risks

which may be present within PDF files are created to support the running of this service.

6. Challenges of born digital accessions within local authority Archives

Adrian Brown defines digital preservation as the “process of maintaining a digital object for as long as required, in a form which is authentic, and accessible to users.”³⁸ This definition is incredibly useful as it points out the inherent challenges which exist within digital preservation:

- Maintaining a digital object
- Looking after it for as long as is required
- Keeping it in a form which is authentic
- Making sure the digital object is accessible to users

All of these points are important, but pose the question, *how do we actually achieve this?*

The National Records of Scotland (NRS) was formed in 2011 following the merger of the General Register Office for Scotland (GROS) and the National Archives of Scotland (NAS).³⁹ In 2014, they published their digital preservation strategy “The National records of Scotland and born digital records – a strategy for today and tomorrow”. In an attempt to meet the objectives of this strategy, they formed a five year digital preservation programme with the primary aim to acquire a digital repository.⁴⁰

³⁸ Adrian Brown, Practical digital preservation: a how to guide for organisations of any size (Facet Publishing 2013 - glossary, pg XII)

³⁹ <https://www.nrscotland.gov.uk/about-us>

⁴⁰ NRS digital preservation strategy 2014 <https://www.nrscotland.gov.uk/files/record-keeping/nrs-digital-preservation-strategy.pdf>

One of the main topics currently being addressed by the Archivists of Scottish Local Authorities Working Group [ASLAWG] - a sub group of the Archives and Records Association [ARA] - is the preservation of born digital records held by local authorities.⁴¹ The NRS, SCA and ASLAWG have been in discussion about the possibility of collaborating to form a joint digital repository to house these aforementioned born digital files.⁴²

However, although a lack of suitable storage is certainly an issue that affects many – if not all – local authorities, but it is certainly not the only concern. Therefore the NRS commissioned a project to address the problems faced within local authority repositories when caring for a digital collection.

⁴¹ <http://www.archives.org.uk/about/sections-interest-groups/aslawg-archivists-of-scottish-local-authorities-working-group.html>

⁴² ASLAWG minutes from meeting August 2016

6.1. NRS digital preservation Skills for the Future projects

Following from their digital preservation strategy, the NRS began working on a digital preservation guidance and capacity planning project and recruited two *Skills for the Future* trainees to carry them out. Initially, the projects were scoped during a day long workshop at New Register House in July 2016, led by Susan Corrigall and Tim Gollins from the NRS. As local authority Archivists and record-keepers were identified as stakeholders in this project, they were invited to attend a day long workshop to help plan and structure the project to help meet their needs in advance of its commencement in autumn 2016. The two successful candidates Penny Wright and Ruth Marr began working at the NRS on the digital preservation guidance project and capacity planning project respectively. Local Authority Archivists and record-keepers became an integral part of the consultation process and the trainees met with them to introduce the project at an ASLAWG meeting in Kirkintilloch in August 2016. Following from this, a survey to identify the number and types of digital holdings within each local authority was distributed for completion amongst the members; the results of which highlighted the need for record keeping professionals within local authorities to have access to training and resources in this area. A few authorities were selected to become more closely involved in the project and from these information gathering sessions over a number of months, the projects took shape.

The project outcomes were delivered at two launches; one in Glasgow in July 2017 and another in Aberdeen in August 2017. The trainees have also delivered presentations at other events including a recent Edinburgh Preserves meeting in August 2017.

Penny delivered the 'NRS digital preservation for local authorities guidance' presentation which highlighted the range of issues faced by LA's and ALEO's such as having to care for a diverse range of digital data created by different departments for different functions. She used Brigadoon Life as a case study and described the added challenges faced there as an ALEO rather than being part of Brigadoon City Council. The suggestion was that the Council created and were responsible for digital records within records and information management, whilst the Archives historically dealt with paper records rather than digital. The retention schedule mandated the time of transfer of the digital records to the Archive for permanent preservation, but this posed a problem such as the lack of available digital storage. As there was nowhere suitable to keep the digital material, they are currently located in several other locations such as shared drives, hard drives, internal servers and external systems. Obviously this is not an ideal scenario as the best practice would be to keep the digital files together in a secure and identifiable location where the files can be easily located, made accessible and able to be checked for integrity. Therefore the information contained within this digital material is at risk because of the lack of appropriate storage.

The recommendations in this instance were to identify the key stakeholders and facilitate joint working across the Council and Brigadoon Life to address the issues. An assessment of the current capability to preserve the digital records was suggested and the need to estimate storage requirements, along with a discussion about any other requirements, strategic drivers and agreement on who should take ownership of the transfer and subsequent storage of the records. Having done this, the aim would be to secure resources to allow the records to be transferred to a suitable location whilst

allowing for access (to comply with FOISA and other legislation) and to maintain the records for as long as required.

Furthermore, Penny discussed file formats and recommended restricting the numbers of formats for transfer, would make calculating and managing the digital data easier. However, it was highlighted that by migrating the files from one format to another, it posed the risk of losing data. The best practice recommended for storage was to save three copies of each file in different locations, similar to the LOCKSS idea of *lots of copies keeps stuff safe*.⁴³

Once permanent storage was made available, the next step would be to deposit the data, and analyse the ingested files by using software such as DROID, which is created by the National Archives as a tool to identify file formats.⁴⁴ Further to this, metadata needs to be captured and included and some form of content management system would be necessary in order to allow for future access to the digital records. These necessary steps highlight the need for training amongst Archivists and record-keeping professionals, as there is the likelihood that many in this field may not feel confident enough to tackle the required steps to transfer, store and maintain digital records in their custody. However, if the steps are carried out, the results would mean that the digital records would be securely managed and accessible.

Ruth presented the findings from her project in the presentation titled 'Introduction to the capacity planning tool.' The target audience for this tool was local authority

Archivists and other professionals responsible for record-keeping and digital

⁴³ <https://www.lockss.org/>

⁴⁴ <http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/>

preservation within the organisation. The aim was to produce a tool that would allow LA's to calculate their storage requirements for those digital files identified as being retained for long-term retention. Similar challenges about the disparity of file formats being held in different locations were discussed, and it was mentioned that in many cases, the current practices and policies of depositing paper based or analogue documents had not been extended to include digital records as yet. Challenges such as a lack of time, lack of knowledge and expertise, separation from document creators and lack of support from IT operatives were discussed and it appeared that these challenges were fairly representative within Scottish local authorities in general.

The capacity planning tool is based on a Microsoft Excel spreadsheet as this is a common software application widely available within local authorities and ALEO's. It was explained that there are two approaches to using the tool; either by inputting information based on existing paper or digital records *or* by taking an overall approach based on the volume of records created on a *function-by-function* basis. It was suggested that the first option to base the calculations on existing records (either paper or digital) would have the advantage of being quicker, easier and more accurate, but the downside would require a strong channel of communication with the Council IT department, which may not always be possible, especially in the situation where the Archive is part of an ALEO. The second option of using a *function-by-function* approach was discussed, with the advantages of being able to consider different applicable retention schedules and the potential to discount departments who would not be likely to deposit records for long term retention, thus making for a more accurate estimate of required storage needs. However, this is far more time consuming than the first option

and still requires good communication channels to be in place, although in this instance with different departments other than IT across the Council, who may deposit their records in the Archive in future. The suggestion was that each local authority would choose the best approach to suit their organisation, and that the tool was only intended as a guide to help estimate future storage requirements rather than to give an exact figure of what those requirements would be.

To conclude, the prevalent challenges faced by Archivists and record-keepers within local authorities are:

- Lack of time; too many other priorities
- Lack of technical knowledge and expertise
- Difficulties in working across different departments

However the main challenge overall is budget constraints; this means that often resources are limited, meaning a lack of staffing and storage can be important issues that prevent local authorities being able to successfully look after their digital holdings, despite it being a requirement under several pieces of legislation (as mentioned previously). It has been argued by Osbourne and Gaebler in Richard Cox's 'Archives and Archivists in the twenty-first century: what will we become?' that "Most public organisations are driven not by their missions, but by their rules and budgets."⁴⁵

⁴⁵ <http://www.jstor.org.ezproxy.lib.gla.ac.uk/stable/41101910>

Therefore at this juncture when most local authority repositories are finding it necessary to prepare for the onset of caring for digital records, advocacy is required to put a business case forward to decision-makers for additional funding and resources to allow Archivists and information professionals meet their statutory requirements.

7. Setting the scene: portable document format [PDF]

The portable document format was invented by Adobe in 1992 as a way of creating a paper free document sharing system that was regarded as more reliable than the multitude of other document file formats available at the time.⁴⁶ *Adobe Acrobat* - the original application used to create PDF files - was released in 1993,⁴⁷ though subsequently, the original proprietary file format changed to open source in January 2007 when Adobe released the specifications of PDF 1.7 for ISO [International Organization for Standardization] certification.⁴⁸

The Library of Congress describes PDF/A as a “family of ISO standards for constrained forms of Adobe PDF intended to be suitable for long-term preservation.”⁴⁹ The first published standardised version of PDF/A is PDF/A-1 from 2005, which is based on PDF version 1.4. The second is PDF/A-2 based on PDF version 1.7 from 2011. The most recent is PDF/A-3 from 2012 which is also based on the 1.7 version and improves on the former versions due to the inclusion of permissions to embed files of other arbitrary file formats within the file whilst still ensuring the PDFs overall conformance to the ISO 19005-1 standard.⁵⁰ Further to this, PDF/A-1 has two conformance levels; level A for *accessible* which maintains the file structure, with level B for *basic*, thereby only maintaining the visual appearance of the file. Level A is the preferred conformance level for born digital files. PDF/A-2 and PDF/A-3 conform to U for *Unicode* which enables the embedding of Unicode information within the PDF document.⁵¹

⁴⁶ <https://www.prepressure.com/pdf/basics/history>

⁴⁷ <https://www.prepressure.com/pdf/basics/history>

⁴⁸ <https://www.prepressure.com/pdf/basics>

⁴⁹ <https://www.loc.gov/preservation/digital/formats/fdd/fdd000318.shtml>

⁵⁰ <https://www.loc.gov/preservation/digital/formats/fdd/fdd000318.shtml>

⁵¹ <https://blogs.bodleian.ox.ac.uk/archivesandmanuscripts/2017/08/14/pdfa-challenges-meeting-the-iso-19005-standard/>

There are several types of PDFs which are ISO regulated - for example; PDF/A for archiving, PDF/X for graphic arts, desktop publishing and printing, and PDF/E for engineering.⁵² Although in theory this move to open source was a good idea, it also means that PDFs can be created in all sorts of ways by using any one of the numerous PDF creation, editing or conversion tools currently available on the market.⁵³

Due to the inherent nature of PDFs as being “independent of software, hardware, or operating system,”⁵⁴ the variety of creation methods means that there are numerous versions and *flavours* of PDFs in use - not all of which comply with ISO standards. Although this flexibility of formation may be an asset to the creator or user of the file, conversely it can become a liability when considering future preservation risks to the document.

⁵² James C. King “Document Formats and Image Formats”

<http://www.dpconline.org/docs/miscellaneous/events/163-document-formats-and-image-formats-king/file>

⁵³ <https://tools.ietf.org/html/rfc3778>

⁵⁴ <https://acrobat.adobe.com/uk/en/why-adobe/about-adobe-pdf.html>

8. Solving the PDF Preservation Problem: DPC Student Project

The Digital Preservation Coalition (DPC) ran a student project in conjunction with the veraPDF Consortium between 2016 - 2017 with the objective “to apply preservation tools, analyse what they report, investigate preservation risks and develop best practice”.⁵⁵

veraPDF is an open source PDF/A validation software which can be downloaded in the form of a desktop graphic user interface [GUI]⁵⁶ and/or as a command line interface [CLI]⁵⁷ for Windows, Mac or Linux users.⁵⁸ It has been developed by the veraPDF Consortium in association with the Open Preservation Foundation [OPF], the PDF Association and the DPC and funded by the European Commission’s PREFORMA Project.⁵⁹ The GUI is the simplest way to use the tool; however it requires the user to input each file separately, which in the case of a large corpus of files, is not necessarily the most effective and timely option. Therefore the developers also made a CLI version available which allows for batch processing, thus making the process more automated and quicker. In addition to this, users could try out the software using a web tool available on the website which worked in a similar way to the downloaded GUI.⁶⁰

The veraPDF software was still in the relatively early stages of development at the time this student project commenced in summer 2016, having only been released as a

⁵⁵ DPC student project outline (2016)

⁵⁶ <http://docs.verapdf.org/gui/>

⁵⁷ <http://docs.verapdf.org/cli/>

⁵⁸ <http://verapdf.org/software/>

⁵⁹ <http://verapdf.org/>

⁶⁰ <http://staging.verapdf.org/>

prototype to the public the previous year, in July 2015.⁶¹ Interested parties were invited to trial the tool and offer feedback to the developers, thus allowing the software to be improved upon to overcome any initial problems and function better overall.

For the purposes of the student project, the GUI and CLI versions for Windows were downloaded in June 2016. However, for the actual trialling phase, a newer version - 0.18.1 - was used for the duration of testing to allow for consistent and comparable results, despite newer updates subsequently becoming available.

Initially there were a few problems installing the software (it was first necessary to update Java⁶²), but this was remedied following guidance from DPC Project Officer Sara Day Thomson, who provided mentorship for the duration of the student project.

PDF files from various online sources were obtained in order become more familiar with the software initially. The GUI was chosen as the preferred method due to the user friendly interface, and several validation and features reports were executed using xml⁶³ as the report output. It became clear early on that not all files could be successfully checked. Error messages such as “could not finish validation due to unexpected error” and “error in reading PDF file: document doesn’t contain startxref keyword couldn’t parse stream” appeared (see figures 1 and 2 respectively, below).

⁶¹ <http://verapdf.org/roadmap/>

⁶² Java: computer programming language as defined by Gosling in *The Java Language Specification* found at: <https://docs.oracle.com/javase/specs/jls/se8/jls8.pdf>

⁶³ XML: *eXtensible Markup Language* which can be deciphered by humans and machines. Definition found at: <https://www.w3schools.com/xml/>

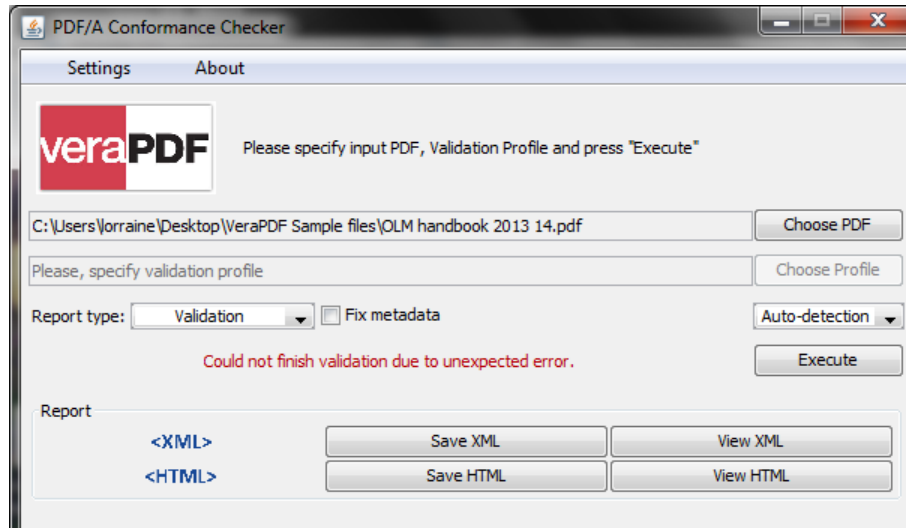


Figure 1

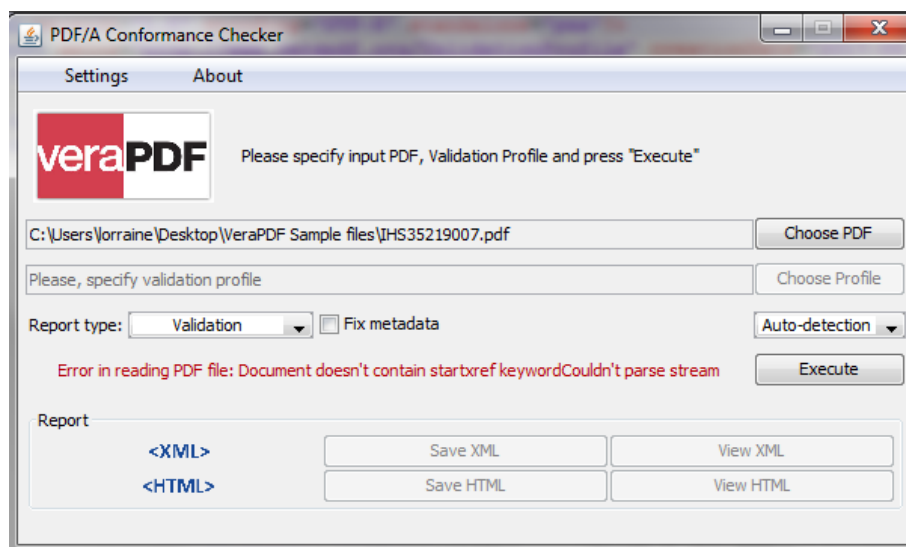
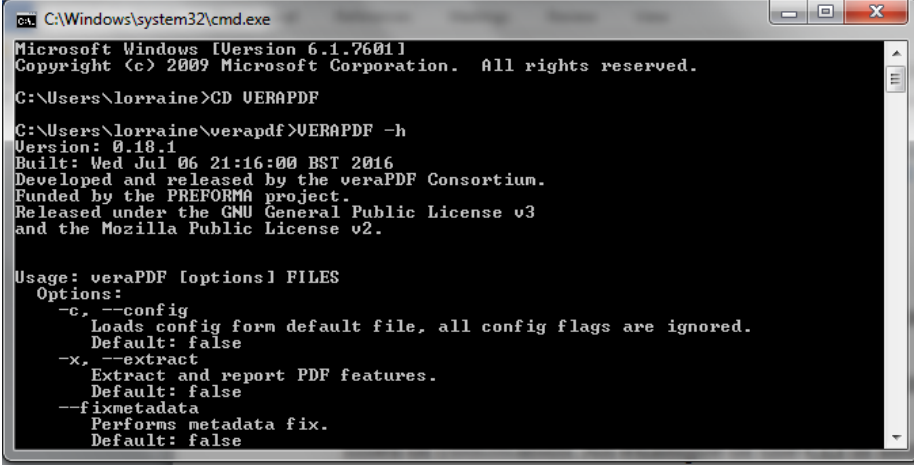


Figure 2

Continuing with the GUI, many of the downloaded PDFs were successfully checked and reported as either “compliant with validation profile requirements” or “not compliant with validation profile requirements”. After becoming more comfortable using the software, the same sample files were checked a second time using the CLI instead.

The process was much more straightforward than anticipated and simply required a few lines of simple commands. It was necessary to create an *in* and *out* folder; the *in* folder to contain the source file to be checked and the *out* folder was specified as the destination for the report to be generated by the process.

An example of the CLI is shown below in figure 3:



```

C:\Windows\system32\cmd.exe
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\lorraine>CD VERAPDF
C:\Users\lorraine\verapdf>VERAPDF -h
Version: 0.18.1
Built: Wed Jul 06 21:16:00 BST 2016
Developed and released by the veraPDF Consortium.
Funded by the PREFORMA project.
Released under the GNU General Public License v3
and the Mozilla Public License v2.

Usage: veraPDF [options] FILES
Options:
  -c, --config
    Loads config from default file, all config flags are ignored.
    Default: false
  -x, --extract
    Extract and report PDF features.
    Default: false
  --fixmetadata
    Performs metadata fix.
    Default: false

```

Figure 3

After becoming more familiar with testing single PDFs, the next stage was to perform batch testing, which (as mentioned before), is only possible using the CLI rather than the GUI version. The command line required to do so was based on the name of the folders which were created for the source files and generated reports (i.e. 'dpcin' and 'dpcout':

verapdf -r dpcin > dpcout/out.xml

Initial batch testing was unsatisfactory as the tool failed in most instances, which was caused by running out of memory. This issue was remedied by updating to veraPDF

version 0.18.1 in July 2017 as advised on the veraPDF Github website; an online forum for users to report issues to the developers and gain new insights.⁶⁴

Additionally, I extended my initial testing to include PDF e-books to see whether they would be PDF/A compliant or not. However, after testing several different ones from different sources, the tool failed and could not offer a result. The next step was to convert the original e-book to the PDF/A version in the hope that the tool would work – and this proved to be successful; converting an e-book to the PDF/A version meant that after testing, the file was compliant and therefore less likely to show preservation risks in future. However, a very important outcome of converting these e-books must be mentioned; the act of migration resulted in a considerable loss of data. The original files contained hyperlinked content pages which allowed the user to navigate the e-book. Upon conversion to PDF/A, these links were removed by the process and the resulting document was a flat, read-only file with no functionality retained. The act of migration may well have mitigated the potential preservation risks, but in doing so removed the essence of what the original digital object was.

⁶⁴ <https://github.com/veraPDF/veraPDF-library/issues>

8.1. Trialing the veraPDF toolset

After completion of the familiarisation stage, the trialling stage began in earnest and to tie in with this dissertation research, a corpus of 100 PDFs obtained from Inverclyde Archives were used as source files to both test the software, and to glean information about any preservation risks which may exist in PDFs within the Archive department. Semi current and current PDF files created by numerous personnel between 2013 - 2017 were chosen as they were readily available within the context of my working environment, and broadly speaking, would likely have been created using different methods and different software for different purposes over a four year span. The rationale behind this decision was that it was considered a more worthwhile endeavour to concentrate on a variety of PDFs produced by a single organisation rather than from a proliferation of sources, thus making it easier to evaluate the results. The diversity of material obtained from within this single organisation was deemed to be enough of a variant to give a broad perspective despite concentrating on one single department within Inverclyde Council; namely Archives. My supposition was that whatever the results achieved, it would be reasonably representative of the extent of compliant files held within the Libraries, Archives and Museums service.

Appendix III titled “Validation report results” shows the data obtained from testing the aforementioned aggregation of 100 PDFs. The data captured in this table includes a unique identifier (file name), statement (true= file is compliant, false=file is non-compliant), passed rules, failed rules, passed checks and failed checks; all of which are

presented in the form of statistical data showing the extent of passes and fails when comparing each individual file to the base set of PDF/A standards.

From the sample of 100 files, only 9% passed the validation checks with the other 91% failing as shown in figure 4 below:

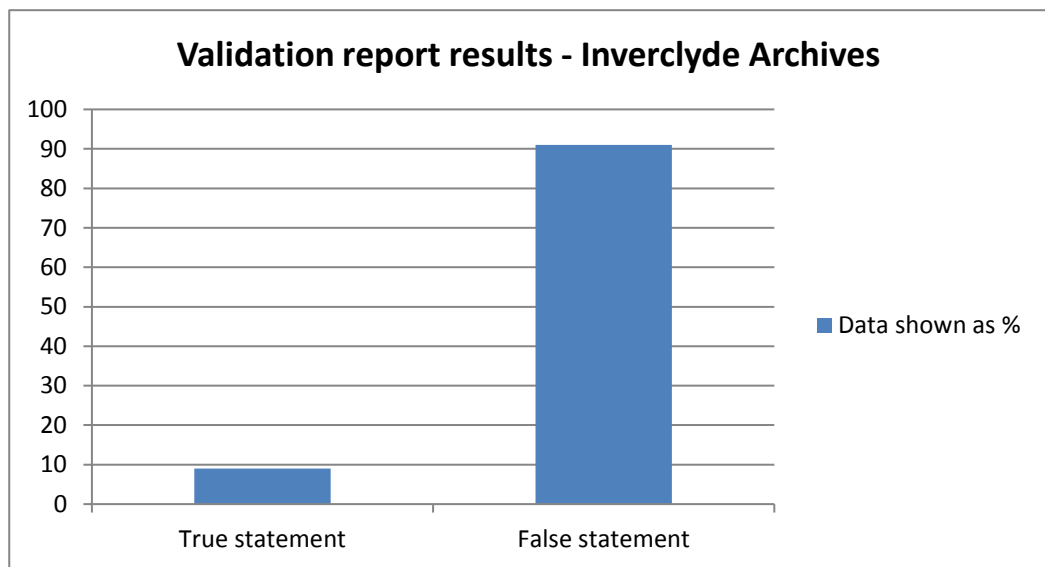


Figure 4

This result was not entirely unexpected; it was imagined that the fails would likely outnumber the passes given the diversity of creation; however the extent of failure were far more prevalent than predicted.

At this stage, the initial testing was only intended to give an idea of the amount of files that were (or were not) PDF/A compliant. The lack of information about the source of the files or any information regarding the creation was limited or non-existent, thus

making further evaluation difficult without drilling down to the metadata in more detail and having to make assumptions where information was not available.

Therefore a second test was carried out using eight files which I created for the purposes of testing and evaluation. Appendix IV shows the data for the second round of testing, which includes the unique identifier, the file name, the method of creation, additional information, file size, statement, and passed and failed rules and checks respectively. By having information about the method of creation, analysis of the results should be more comprehensive. The result of this second test was significantly different to the initial testing; 50% pass and 50% fail rate.

Depending on the way the PDF file is created, it is usually possible to save the version as a PDF/A file which is the version deemed most appropriate for long term preservation.⁶⁵

Within the context of Inverclyde Archives, PDFs are usually created in one of two ways; either by saving from an existing Microsoft Word document or by scanning from a corporate multifunctional device [MFD]. In the case of the former, when saving a MS Word document to PDF, the document creator can specify the version of PDF to be used. Within the 'PDF options', a box can be ticked to save the document as PDF/A, thus making it ISO 19005-1 compliant as shown in figure 5 (below):

⁶⁵ <https://www.loc.gov/preservation/digital/formats/fdd/fdd000318.shtml>

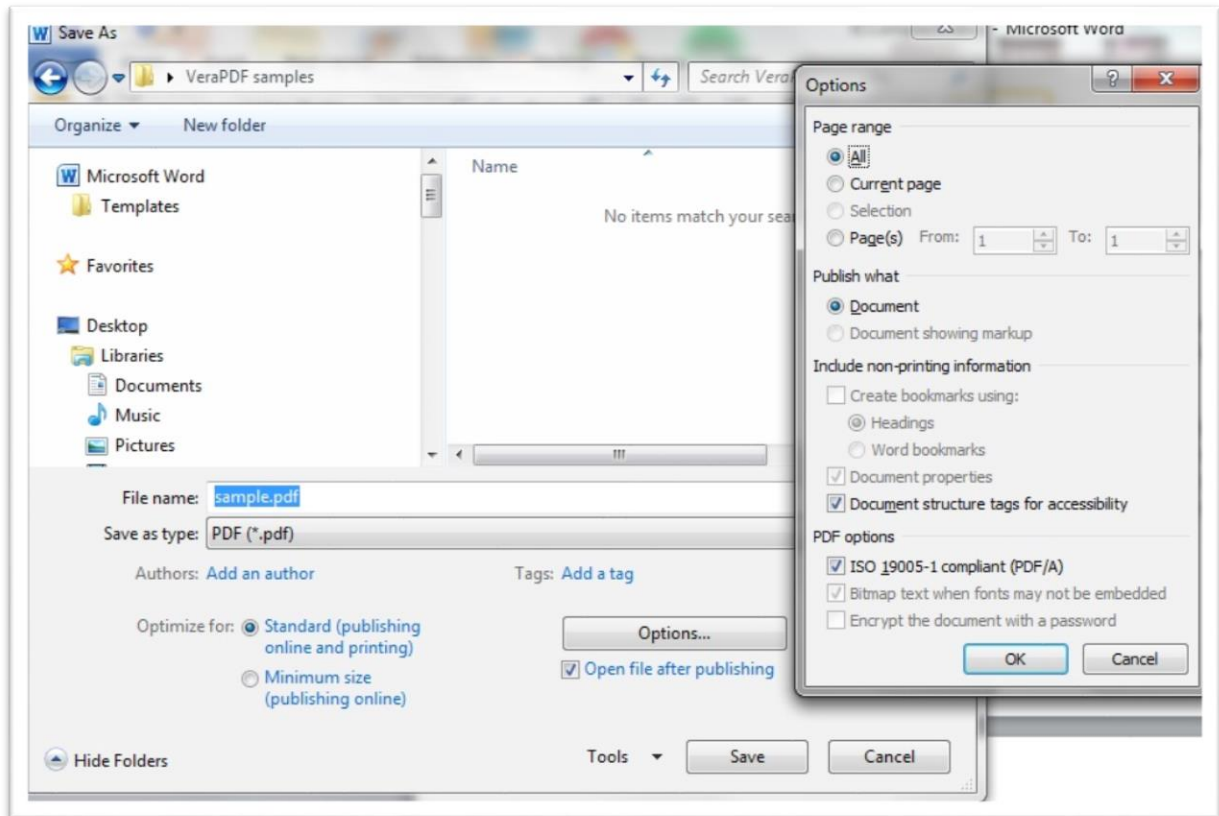


Figure 5

Figure 6 (below) confirms the saved file complies with the PDF/A standard when the document is opened for viewing.

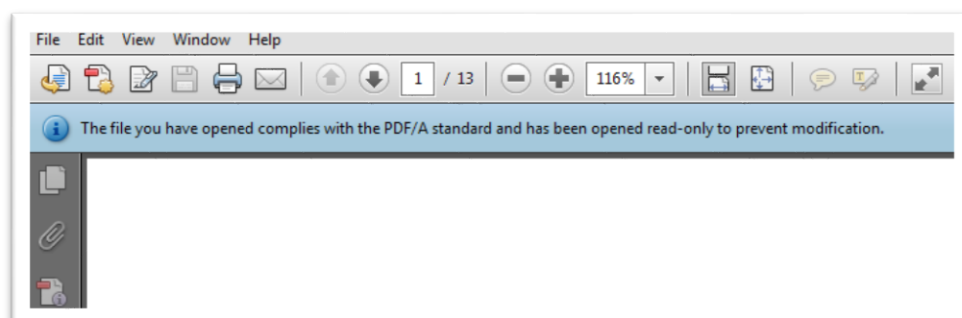


Figure 6

From the eight files produced and tested, four were deliberately saved as PDF/A versions using the option in Microsoft Word as detailed above. Those files were as follows; PDF02, PDF03, PDF05 and PDF06. However, despite choosing the PDF/A version, only two of these four files passed the test; PDF02 and PDF03. I did not expect PDF02 to pass as it contained embedded images, however it did pass. And although I did expect PDF03 to pass (which it did), as it only contained Arial font text, the next two fails (namely PDF05 and PDF06) were unexpected fails. Both of these files were almost identical to PDF03, however in PDF05, I included a hyperlink and in PDF06, I added some special characters. From this, I must conclude that hyperlinks and special characters do not conform to the PDF/A set of specifications, so must be considered a potential preservation risk. However as PDF02 passed compliancy, it would appear that embedded media is not considered a preservation risk.

PDF01 was created by scanning an image using one of our corporate MFD's (which is a method of PDF creation used to great extent within Inverclyde Council). This document failed the compliancy test; either due to the creator being unable to specify the version of PDF to save the scanned file to, or for another unknown reason entirely. After some investigation of the MFD settings, if an option to save as PDF/A was available (I was unable to find it), it certainly was not the default setting and I strongly believe any individual creating files using this method would simply *scan to PDF* as I did in this instance without further investigation. Therefore the conclusion drawn here is that using a scanner to save as a PDF file in this manner results in a file that has potential preservation risks.

PDF04 was exactly the same document as PDF03, however I unchecked the option to save the file as PDF/A, meaning it saves as another version of PDF, consequently

resulting in a fail, which was entirely expected. Interestingly enough, by not specifying this file as a PDF/A compliant document, it created a smaller sized file (94.8 KB for the non-compliant file as opposed to 103KB for the compliant file), thereby concluding that although PDF/A files have less potential preservation risks, they require more storage space.

PDF07 was created using Microsoft PowerPoint – another common way of creating PDFs – and despite not having the option to specify what type of PDF to save this document as, it appears to have been saved as PDF/A by default, resulting in an unexpected pass. However, the original PowerPoint document contained notes under each slide which did not appear in the created PDF, thereby resulting in a loss of data.

The final file to examine is PDF08 which was saved as a PDF using Microsoft Excel. This document contained a set of tabular numerical data, formulas and text. Because of this, I fully expected this file to fail; however this was not so. When the the original file was saved as a compliant PDF/A file, it lost all of its functionality, making it a flat, read-only document, thereby incurring a loss of data and changing the essence of the file.

The changes that occurred in the two aforementioned files when saving the original document to a different format (and similarly in the conversion of e-books to PDF/A), would appear to result in a diminished value of the original file, thus prompting the question of whether the potential risk of saving the fully functional Excel or PowerPoint document outweighs the perceived benefits of saving the document as PDF/A.

8.2. Project conclusions and recommendations

As mentioned before, the first round of testing was fairly basic, but twofold; essentially to try out the tool and offer feedback to the developers within the remit of the DPC student project, but also to gauge the likelihood that PDFs contained within Inverclyde Libraries, Archives and Museum would pose potential preservation risks in future. On the latter point, the results were conclusive; with only a 9% compliancy pass rate from a sample of 100 documents, it would appear that the PDFs within this department are very much at risk in future, when supporting the hypothesis that PDF/A's are the most robust version of PDF for Archiving and long-term preservation. It would be difficult to say exactly what potential risks these documents may contain and how they could be mitigated without extending the scope of this project, but there is indeed scope to pursue this with further and more detailed examination at a later time.

The second round of testing was intended to give an insight into how the creation of a document or the manner in which it is saved to PDF would affect its long-term viability. Obviously with such a small number of sample files it is difficult to make any conclusive statements, however, it would appear that simply creating a PDF/A from Microsoft Word is not enough in itself to create a compliant document as evidenced by the failure of PDF05 and PDF06. Similarly, it appears that scanning a document to PDF using a MFD should not be recommended either – although this may not be the case universally; it may well depend on the brand (in this instance, Konika) and the settings of the individual MFD. However considering the high failure rate in the first round of testing and the prior knowledge that many PDFs within this department are created in

such a manner, I would argue that this practice should be avoided in the context of local authorities. The final point to make from this secondary round of testing is that saving a different file format to PDF may well mean that the new file created is potentially less at risk in the long-term, but it must be noted that the process of doing so changes the essence of what the digital object is, and – in my opinion – decreases its value. In such cases, my recommendation would be to save the original file format **and** the created PDF to maximise the chances of long-term accessibility and re-use. Despite the fact that the price of storage is decreasing as discussed by Wright, Addis and Miller,⁶⁶ storage still costs money. Therefore within the context of a local authority where budgets are always an important consideration, saving multiple versions of the same file may not be a viable option.

The scope of the DPC student project was fairly broad; I could choose the source material for testing, thus allowing the project to be tailored to both inform my dissertation and my role within Inverclyde Archives, proving to be very useful and informative. The project allowed me to gain experience of working with the veraPDF tool and using the command line interface, which I have only a limited experience of thus far. The software has been developed and tested and the project is now at the final stages. The final product is a tool that can not only test for PDF/A compliancy, but fix metadata, has a policy checker to confirm whether the document complies with institutional policy and creates machine readable reports which can also be analysed by humans⁶⁷ with the help of the veraPDF library.⁶⁸

⁶⁶ <http://www.ijdc.net/index.php/ijdc/article/view/138/173>

⁶⁷ <http://verapdf.org/home>

The software is open source, meaning that is freely available and the GUI version is very user friendly making it a tool that even a beginner could competently master.

However, there are concerns about the outcomes from having tested the tool, though they do not stem from the software – which I feel is a very useful tool – but rather the underlying perception about the PDF/A format. It is widely considered that PDF “is the *de facto* standard for electronic documents worldwide”⁶⁹ and the notion of PDF/A being categorically the best PDF format for Archiving purposes is perpetuated.⁷⁰

Therefore any conclusions previously drawn about preservation risks from these rounds of testing are based on the assumption that PDF/A is the most robust and reliable version of PDF to use for long-term digital preservation. But the question must be asked, is PDF/A really the best choice of format to use when considering carrying out a digital preservation strategy? Certainly this is not a question that can be answered within the scope of this research - if indeed at all - however the point to be made here is that the very act of migrating or converting another file format to PDF/A will almost certainly result in a loss of data to some extent. Is it acceptable to lose data in order to preserve the file for future use, or is it worth taking the risk on the original file which has all the functionality and elements present at the point of creation?

In other words, is the artefact or the content more important?

⁶⁸ <https://github.com/veraPDF/veraPDF-library>

⁶⁹ <https://www.pdfa.org/pdf-association-2/>

⁷⁰ <https://www.iso.org/news/2005/10/Ref974.html>

9. Conclusion

Anna Oates recently presented a paper called “A Case Study on Theses in Oxford’s Institutional Repository: Challenges Meeting the ISO 19005 Standard” in autumn 2017 at the Bodleian Library. She noted that PDF/A is better suited to long term preservation than other variants of PDF as it “maintains the integrity of the information included in the source files”⁷¹ Though despite this, she argued that PDF/A was not necessarily the best format for long term preservation at this point due to a number of areas that she considers require improvement. One issue highlighted by Oates was the differences that exist between comparable embedded images when converting PDF to PDF/A files (i.e. changes to the colour of pixels), which casts doubt on the authenticity of the converted file. Another issue was the increase of the original file size when converting original PDF files to PDF/A, thus taking up more space within the host repository, which in turn makes it more expensive to store long-term. Both of these issues have been already considered within this research. Overall it was concluded by the attendees at this presentation, that the best option would be to keep both the original PDF file as the primary source and the created PDF/A version as the stable version for preservation purposes⁷² which is an option I consider worthwhile if the situation and landscape allows for it.

Marco Klindt recently presented a paper at iPRES; the 14th International Conference on Digital Preservation, in Japan on 27th July 2017 titled “PDF/A considered harmful for

⁷¹ <https://blogs.bodleian.ox.ac.uk/archivesandmanuscripts/2017/08/14/pdfa-challenges-meeting-the-iso-19005-standard/>

⁷² <https://blogs.bodleian.ox.ac.uk/archivesandmanuscripts/2017/08/14/pdfa-challenges-meeting-the-iso-19005-standard/>

digital preservation”.⁷³ As this is such a recent paper and therefore not available to consult, it would be an interesting addition to this discourse, especially when PDF/A is the recognised preferred format for archiving purposes and accepted as such. Any argument posed by Klindt as to why PDF/A’s may be harmful would appear to offer a contrary view to the accepted norm and is well worth addressing in future research on this topic.

It is worth noting that the research carried out for the purposes of this dissertation only scratches the surface of this topic; further testing with a larger corpus of source material may yield some interesting and quite different results, as would carrying out additional testing of PDFs from other sources to compare with the results achieved here. This experiment was limited; the second set of assessments would have benefited from creating a larger number of PDFs using different hardware, different operating systems and different software applications. The results achieved in this instance may not be wholly representative of PDFs created under the same circumstances elsewhere and should be noted as such. However, it does pay homage to the fact that not all PDFs are created equally – and that the method of creation has a substantial effect on the future viability of the file in question.

In short, trialling the veraPDF software was a most useful experiment; it prompts the user to think about potential preservation risks and will perhaps inform document creators on how to create PDFs in future to mitigate any potential risks by considering the method in which they create PDFs. However, this does not address the issue of legacy files and the challenges they present with an Archival repository. Most legacy

⁷³ <https://ipres2017.jp/programme/>

files have been created over a period of time using different hardware and software, resulting in many different *flavours* of PDF being in existence. Migration may well be an easier or more cost effective approach to retaining these files for future use, however in doing so, it must be remembered that the migrated file is, in fact, a surrogate; and not the original digital object. The act of migration changes the essence of the digital object; it can diminish the value, change the meaning, and result in a change of appearance (as evidenced by Oates).

Therefore I would argue that the migrated file it is no longer authentic, and within the context of an Archive, this is an important consideration to be mindful of. Archives have historically created surrogates to allow for easier access of content contained within primary material; from creating microfilm reels, to digitisation of an analogue counterpart. Using surrogates within an Archive is a useful method to allow for more immediate access, which is of benefit to both service users and information professionals alike. However, I would argue that a migrated or converted file – although worth having – is **not the original object**. It is simply a surrogate which can prove useful for access to the content in future. It cannot be accepted as the authentic, original digital object - and as such - the recommendation would be to keep both the original file and the surrogate – each for different reasons, in the same way we do with other items within the Archive. Do we dispose of the original manuscript or document once it has been digitised since we may not need to consult the original again? Or do we look after it to the best of our ability with the tools we have at our disposal at that time?

The same logic should apply to the original digital entity. The potential threat of loss or corruption should not be the reason why we dispose of the original digital object. Of

course, the original object may well be at risk, and one day it may not behave the way it ought to, or may indeed become corrupt and become incapable of being accessed at all. However the threat or possibility of this happening should not cause us to abandon the original object in favour of the surrogate. Technology is changing rapidly, and there is always the possibility that advances in future could overcome the potential preservation risks we perceive at this point in time. Rothenberg argues that migration is not always the best approach due to the process causing irreversible changes to the file, and suggests that although emulation may be less cost effective, more complex and take more up time and resources, it is the best way of limiting the potential risk of obsolescence by providing the original digital object a way of being rendered in the way it was intended to be. He goes on to say that “it is not sufficient to save the bit stream of a digital informational entity without also saving the intended interpreter of that bit stream. Doing so would be analogous to saving hieroglyphics without saving a Rosetta Stone.”⁷⁴

To conclude this discussion, I will return to the research questions posed at the beginning of this paper. Regarding the question about the main challenges faced by LA Archives when caring for their digital collections, the most appropriate answer to this is that of restrictions placed by resource limitations. If an Archive has access to funding, it is possible to overcome many of the challenges faced such as lack of available storage (this can be purchased), lack of staff time to commit to digital preservation (additional staff can be recruited), and lack of knowledge or expertise (staff training can be procured when budget is not an issue). The only other issue

⁷⁴ <https://www.clir.org/pubs/reports/pub92/pub92.pdf>

which may still exist is the difficulty of working across different departments or arm's length organisations; however this can also be a step that is possible within budgetary constraints; creating a dialogue with colleagues can be done when a shared goal of preserving Public Records is recognised and addressed in a collaborative way.

In response to the next question about restricting the ingress of PDFs to only PDF/A files within an Archival collection, it is believed that this would not necessarily be the best approach to take when considering legacy files. Using the migrating approach may well preserve the overall content of the original file, but it may also result in a loss of data in the process. Local authorities must be seen to be transparent at all times, and I would argue that this process will likely result in making changes to Public Records, which is entirely unacceptable. The best approach in this instance would be to create a surrogate in a more robust format for long-term use, and keep the original document as the primary source.

The issue of whether saving other digital file formats to PDF/A may incur any loss of original content, appearance or authenticity of the original file has been addressed and evidenced thus far. The process of migration or conversion does indeed result in loss – though this may not be evident in every case; the file may look the same, appear to have the same content, but it will have changed in some way; the file will be larger than before and the metadata or essential properties may have been affected in some way. Whether this change affects the meaning or value of the digital object is still up for discussion; the new file may well convey the same meaning as before and it may still be

of value as historical evidence. However the essence of the file has changed, and this may mean the completeness, validity or potential to be used in the manner in which it was created has been compromised or lost in some way.

The last two research questions are related; the question of what is considered more important within an Archival repository - the appearance or the content of the digital file and whether the artefact or information contained within the file is more important? Both questions are much more difficult to address as the artefact is of great value to the Archivist, but equally so is the information contained therein. Which relates to the issue of content vs appearance - different properties are equally important for different reason; Archives aim not only to preserve the information, but also the container of that information. Both are equally valuable and therefore such questions would seem impossible to answer.

To sum up, I would recommend that if saving both the original digital file (as the artefact) and the preservation file (as the access copy) is at all possible, this should be the best approach to take. Additionally, there should be at least three copies of said files saved in different locations to minimise the potential loss of these files, even though creating duplicate versions brings into question the issue of originality. These issues are not going away any time soon; and logic dictates that as technology develops, there is the possibility of new challenges becoming apparent in future. However, it is hoped that as improvements to technology occur, this may result in a positive effect on how we create, manage and preserve our digital legacy.

Bibliography

Adobe. (2017). What is PDF? Adobe Portable Document Format | Adobe Acrobat. [online] Available at: <https://acrobat.adobe.com/uk/en/why-adobe/about-adobe-pdf.html> [Accessed 4 Sep. 2017].

Archaeology Data Service. (2015). Guidelines for Depositors: Version 3.0 September 2015. [online] Available at: <http://archaeologydataservice.ac.uk/advice/FilelevelMetadata.xhtml#Documents> [Accessed 4 Sep. 2017].

Archives and Records Association. (2017). ASLAWG (Archivists of Scottish Local Authorities Working Group). [online] Available at: <http://www.archives.org.uk/about/sections-interest-groups/aslawg-archivists-of-scottish-local-authorities-working-group.html> [Accessed 4 Sep. 2017].

Arms, C., Chalfant, D., DeVorsey, K., Dietrich, C., Fleischhauer, C., Lazorchak, B., Morrissey, S. and Murray, K. (2017). *The benefits and risks of the PDF/A format for Archival institutions*. [online] NDSA. Available at: <http://hdl.loc.gov/loc.gdc/lcpub.2013655115.1> [Accessed 4 Sep. 2017].

Barata, K. (2004). Archives in the digital age. *Journal of the Society of Archivists*, [online] Vol 25(Issue 1), pp.63-70. Available at: <http://www.tandfonline.com/doi/pdf/10.1080/0037981042000199151> [Accessed 4 Sep. 2017].

Beagrie, N. and Greenstein, D. (1998). A strategic policy framework for creating and preserving digital collections: a report to the Digital Archiving Working Group. [online] British Library Research and Innovation Centre. Available at: 1998 <http://opus.bath.ac.uk/35448/1/framework.pdf> [Accessed 4 Sep. 2017].

British Standards Institution. (2017). British Standards Group - Standards. [online] Available at: <https://www.bsigroup.com/en-GB/standards/> [Accessed 4 Sep. 2017].

Brown, A. (2008). *Selecting file formats for long-term preservation*. Digital Preservation Guidance Note. [online] The National Archives. Available at: <https://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf> [Accessed 4 Sep. 2017].

Brown, A. (2013). *Practical digital preservation: a how to guide for organisations of any size*. [London]: Facet Publishing.

Chartered Institute of Public Finance and Accountancy. (2017). CIPFA Stats - Local Government. [online] Available at: <https://www.cipfastats.net/cipfastats/> [Accessed 4 Sep. 2017].

Cook, T. (1997). What is Past is Prologue: A History of Archival Ideas Since 1898, and the Future Paradigm Shift. *Archivaria*, [online] Vol 43, Spring 1997. Available at: <http://archivaria.ca/index.php/archivaria/article/view/12175/13184> [Accessed 4 Sep. 2017].

COPTR contributors. (2017). COPTR. [online] Coptr.digipres.org. Available at: http://coptr.digipres.org/Main_Page [Accessed 4 Sep. 2017].

Cox, R. (1995). Archives and Archivists in the twenty-first century: what will we become? *Archival Issues*, [online] Vol 20(Issue 2), pp.97-113. Available at: <http://www.jstor.org.ezproxy.lib.gla.ac.uk/stable/41101910> [Accessed 4 Sep. 2017].

Cox, R. (2000). *Closing an era: Historical Perspectives on Modern Archives and Records Management Westport*. [Westport, Connecticut, London]: Greenwood Press.

Cullen, C., Hirtle, P., Levy, D., Lynch, C. and Rothenberg, J. (2000). *Authenticity in a digital environment*. [online] Washington, D.C.: Council on Library and Information Resources. Available at: <https://www.clir.org/pubs/reports/pub92/pub92.pdf> [Accessed 4 Sep. 2017].

Dearstyne, B. (2002). Effective Approaches for Managing Electronic Records and Archives. *American Archivist*, Vol 65(Issue 2).

Digital Curation Centre. (2017). What is digital curation? | Digital Curation Centre. [online] Available at: <http://www.dcc.ac.uk/digital-curation/what-digital-curation> [Accessed 4 Sep. 2017].

Digital Preservation Coalition. (2017). Digital Preservation Handbook. [online] Available at: <http://www.dpconline.org/handbook> [Accessed 4 Sep. 2017].

Digital Preservation Coalition. (2017). Digital Preservation Handbook Glossary – B. [online] Available at: <http://www.dpconline.org/handbook/glossary#B> [Accessed 4 Sep. 2017].

Digital Preservation Handbook. (2017). 2nd ed. [online] Digital Preservation Coalition, Chapter: File formats and standards. Available at: <http://www.dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards> [Accessed 4 Sep. 2017].

Digital Preservation Handbook. (2017). 2nd ed. [online] Digital Preservation Coalition, Chapter: Why digital preservation matters. Available at: <http://www.dpconline.org/handbook/digital-preservation/why-digital-preservation-matters> [Accessed 4 Sep. 2017].

Dressler, V. (2010). Migration and emulation tools. [online] Digital Preservation: issues and other related topics. Available at: http://blog.case.edu/digitalpreservation/2010/11/29/week_5_migration_and_emulation_tools [Accessed 4 Sep. 2017].

Duranti, L. (1995). Reliability and Authenticity: The Concepts and Their Implications. *Archivaria*. [online] Vol 39, pp.5-10. Available at: <http://archivaria.ca/index.php/archivaria/article/download/12063/13035> [Accessed 4 Sep. 2017].

Gilliland-Swetland, A. (2000). *Enduring paradigm, new opportunities*. Washington, D.C.: Council on Library and Information Resources.

GitHub. (2017). veraPDF/veraPDF-library. [online] Available at: <https://github.com/veraPDF/veraPDF-library/issues> [Accessed 4 Sep. 2017].

Gosling, J., Joy, B., Steele, G., Bracha, G. and Buckley, A. (2013). The Java® Language Specification Java SE 8 Edition. [online] Oracle America. Available at: <https://docs.oracle.com/javase/specs/jls/se8/jls8.pdf> [Accessed 4 Sep. 2017].

Guidance to the form and content of the model records management plan. (2017). [online] National Records of Scotland. Available at: <https://www.nrscotland.gov.uk/files//record-keeping/public-records-act/prsa-guidance-document.pdf> [Accessed 4 Sep. 2017].

Harvey, D, R. (2012). Preserving digital materials. Berlin: De Gruyter Saur.

HM Government and The National Archives. (2011). Public Records (Scotland) Act 2011 (Asp 12) [online] HM Government and The National Archives. Available at: http://www.legislation.gov.uk/asp/2011/12/pdfs/asp_20110012_en.pdf [Accessed 4 Sep. 2017].

HM Government and The National Archives. (2002) Freedom of Information (Scotland) Act 2002 [Asp 13] [online] HM Government and The National Archives. Available at: http://www.legislation.gov.uk/asp/2002/13/pdfs/asp_20020013_en.pdf [Accessed 4 Sep. 2017].

HM Government and The National Archives. (2005) Data Protection Act 1998 (Chapter 29) [online] HM Government and The National Archives. Available at: http://www.legislation.gov.uk/ukpga/1998/29/pdfs/ukpga_19980029_en.pdf [Accessed 4 Sep. 2017].

HM Government and The National Archives. (1994) Local Government etc. (Scotland) Act 1994 (Chapter 39) [online] HM Government and The National Archives. Available at: http://www.legislation.gov.uk/ukpga/1994/39/pdfs/ukpga_19940039_en.pdf [Accessed 4 Sep. 2017].

HM Government and The National Archives. (2004) The Environmental Information (Scotland) Regulations 2004 (No. 520) [online] HM Government and The National Archives. Available at: http://www.legislation.gov.uk/ssi/2004/520/pdfs/ssi_20040520_en.pdf [Accessed 4 Sep. 2017].

HM Government and The National Archives. (1985) Local Government (Access to Information) Act 1985 (Chapter 43) [online] HM Government and The National Archives. Available at: http://www.legislation.gov.uk/ukpga/1985/43/pdfs/ukpga_19850043_en.pdf [Accessed 4 Sep. 2017].

HM Government and The National Archives. (2015) The Re-use of Public Sector Information Regulations 2015 (No. 1415) [online] HM Government and The National Archives. Available at: http://www.legislation.gov.uk/uksi/2015/1415/pdfs/uksi_20151415_en.pdf [Accessed 4 Sep. 2017].

International Organization for Standardization. (2015). ISO 19005-1:2005 - Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A-1). [online] Available at: <https://www.iso.org/standard/38920.html> [Accessed 4 Sep. 2017].

International Organization for Standardization. (2005). New ISO standard will ensure long life for PDF documents. [online] Available at: <https://www.iso.org/news/2005/10/Ref974.html> [Accessed 4 Sep. 2017].

International Electrotechnical Commission. (2017). Standards - International Electrotechnical Commission. [online] Iec.ch. Available at: <http://www.iec.ch/about/activities/standards.htm> [Accessed 4 Sep. 2017].

Inverclyde Council. (2015). Inverclyde Council Corporate Management Structure. [online] Available at: <https://inverclyde.gov.uk/assets/attach/2075/.pdf> [Accessed 4 Sep. 2017].

Ipres. (2017). Programme – iPRES2017. [online] Available at: <https://ipres2017.jp/programme/> [Accessed 4 Sep. 2017].

Jackson, A. (2017). *Formats over Time: Exploring UK Web History*. [online] The British Library. Available at: <http://arxiv.org/pdf/1210.1714v1.pdf> [Accessed 4 Sep. 2017].

Jenkinson, H. (1922). *A manual of archive administration including the problems of war archives and archive making*. [Oxford]: Clarendon Press.

Johnson, D. (2017). The 8 most popular document formats on the web. [online] Duff Johnson Strategy and Communications. Available at: <http://duff-johnson.com/2014/02/17/the-8-most-popular-document-formats-on-the-web/> [Accessed 4 Sep. 2017].

King, J. (2017). Document Formats and Image Formats. [online] Digital Preservation Coalition. Available at: <http://www.dpconline.org/docs/miscellaneous/events/163-document-formats-and-image-formats-king/file> [Accessed 4 Sep. 2017].

Kuny, T. (1997). A Digital Dark Ages? Challenges in the Preservation of Electronic Information. [online] 63rd IFLA Council and General Conference Presentation. Available at: <https://archive.ifla.org/IV/ifla63/63kuny1.pdf> [Accessed 4 Sep. 2017].

Lamb, W. (1966). The changing role of the archivist. *The American Archivist*, [online] Vol 29(Issue 1), pp.3-10. Available at: <http://www.americanarchivist.org/doi/pdf/10.17723/aarc.29.1.c6015k0057756510> [Accessed 4 Sep. 2017].

Library of Congress. (2017). Sustainability of Digital Formats: Planning for Library of Congress Collections. PDF/A, PDF for Long-term Preservation. [online] Available at: <https://www.loc.gov/preservation/digital/formats/fdd/fdd000318.shtml> [Accessed 4 Sep. 2017].

Lockss.org. (2017). LOCKSS | Lots of Copies Keep Stuff Safe. [online] Available at: <https://www.lockss.org/> [Accessed 4 Sep. 2017].

Lu, M. and Chiueh, T. (2017). Challenges of Long-Term Digital Archiving: A Survey. [online] Available at: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=173B4602749746EEE90CAC918A0B4313?doi=10.1.1.80.9675&rep=rep1&type=pdf> [Accessed 4 Sep. 2017].

Marr, R. (2017). Introduction to the capacity planning tool. National Records for Scotland Presentation, Aberdeen, 23rd August 2017.

Morrissey, S. (2011). More What You'd Call 'Guidelines' Than Actual Rules': Variation in the Use of Standards. *Standards*, [online] Vol 14(Issue 1). Available at: <http://dx.doi.org/10.3998/3336451.0014.104> [Accessed 4 Sep. 2017].

Moss, M. and Tough, A. [Eds.] (2006). *Record keeping in a hybrid environment: Managing the creation, use, preservation and disposal of unpublished information objects in publishing*. 1st ed. Oxford: Chandos Publishing.

Mutero, E. (2017). The Records Life-Circle and Continuum Concepts. [Blog] *records and archives*. Available at: <https://recordsandarchives.wordpress.com/2011/07/14/the-records-life-circle-and-continuum-concerpts/> [Accessed 4 Sep. 2017].

McKemmish, S., Piggott, M., Reed, B. and Upward, F. [Eds.] (2005). *Archives: Record keeping in society*. Centre for Information Studies, Charles Sturt University.

Orwell, G. (1949). *Nineteen Eighty-Four*. New York: Harcourt, Brace and Co.

Owen, M, P. (2017). Procuring Digital Preservation: A Briefing. [online] *Scottish Council on Archives*. Available at: <http://www.scottisharchives.org.uk/news/announcements/procuringdigitalpreservation> [Accessed 4 Sep. 2017].

PrePressure. (2017). PDF Basics. [online] Available at: <https://www.prepressure.com/pdf/basics/> [Accessed 4 Sep. 2017].

PrePressure. (2017). The history of PDF | How the file format and Acrobat evolved. [online] Available at: <https://www.prepressure.com/pdf/basics/history> [Accessed 4 Sep. 2017].

Records management and preservation of Archival records policy. (2013). Version 1.1. [online] Falkirk Council. Available at: <https://www.falkirk.gov.uk/services/council-democracy/access-to-information/docs/records-management/Records%20Management%20and%20Preservation%20of%20Archival%20Records%20Policy.pdf?v=201508101049> [Accessed 4 Sep. 2017].

Reid, G. (2010). The challenge of making archives relevant to local authorities. *Records Management Journal*, Vol 20(2), pp.226-243.

Rothenberg, J. (2000). Preserving Authentic Digital Information. In: C. Cullen, P. Hirtle, D. Levy, C. Lynch and A. Smith, ed., *Authenticity in a Digital Environment*. [online] Washington, D.C: Council on Library and Information Resources, pp.51-68. Available at: <https://www.clir.org/pubs/reports/pub92/pub92.pdf> [Accessed 4 Sep. 2017].

Safdar, I. (2017). PDF/A: Challenges Meeting the ISO 19005 Standard. [Blog] *Archives and Manuscripts at the Bodleian Library*. Available at: <https://blogs.bodleian.ox.ac.uk/archivesandmanuscripts/2017/08/14/pdfa-challenges-meeting-the-iso-19005-standard/> [Accessed 4 Sep. 2017].

Schellenberg, T. (1956). *Modern archives; principles and techniques*. Chicago: Society of American Archivists.

Scottish Government. (2011). Scottish Ministers' code of practice on records management by Scottish local authorities under the freedom of Information (Scotland) Act 2002. (SG/2011/233). [online] Scottish Government. Available at: <http://www.gov.scot/Resource/Doc/933/0124124.pdf> [Accessed 4 Sep. 2017].

Scottish Government. (2017). Scottish local authorities. [online] Available at: <http://www.gov.scot/Topics/Government/local-government/localg/usefullinks> [Accessed 4 Sep. 2017].

Scottish Government. (2017). What local authorities do. [online] Available at: <http://www.gov.scot/Topics/Government/local-government/localg/whatLGdoes> [Accessed 4 Sep. 2017].

Schieber, S. (2017). Digital archiving in the Hessisches Landesarchiv. *Journal of Automation, Building and Technology in the Archives, Library and Information Technology*, [online] Vol 37(Issue 2). Available at: <https://doi.org/10.1515/abitech-2017-0022> [Accessed 4 Sep. 2017].

SEEMiS. (2017). SEEMiS Group - Education Management Information System. [online] Available at: <https://www.seemis.gov.scot/site3/index.php/about> [Accessed 4 Sep. 2017].

Taft, E., Pravetz, J., Zilles, S. and Masinter, L. (2004). *RFC 3778 - The application/pdf Media Type*. [online] Tools.ietf.org. Available at: <https://tools.ietf.org/html/rfc3778> [Accessed 4 Sep. 2017].

Theimer, K. (2014). A Distinction Worth Exploring: 'Archives' and 'Digital Historical Representations'. *Journal of digital humanities*, [online] Vol 3(No. 2). Available at: <http://journalofdigitalhumanities.org/3-2/a-distinction-worth-exploring-archives-and-digital-historical-representations/> [Accessed 4 Sep. 2017].

The National Archives. (2017). *Preserving Digital Records / Guidance - The National Archives*. [online] Available at: <http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/guidance/> [Accessed 4 Sep. 2017].

The National Archives. (2017). *DRROID: file format identification tool - The National Archives*. [online] Available at: <http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/> [Accessed 4 Sep. 2017].

The National Records of Scotland and born digital records – a strategy for today and tomorrow. (2015). [online] The National Records of Scotland. Available at: <https://www.nrscotland.gov.uk/files/record-keeping/nrs-digital-preservation-strategy.pdf> [Accessed 4 Sep. 2017].

The National Records of Scotland. (2017). National Archives of Scotland Archive content – Public records act introduction. [online] Webarchive.nrscotland.gov.uk. Available at: <http://webarchive.nrscotland.gov.uk/20170106021747/http://www.nas.gov.uk/record-keeping/publicrecordsactintroduction.asp> [Accessed 4 Sep. 2017].

The National Records of Scotland. (2017). Public Records (Scotland) Act 2011: Report by the Keeper of the Records of Scotland, 2016. (SG/2016/129). [online] National Records of Scotland. Available at: <https://www.nrscotland.gov.uk/files//record-keeping/public-records-act/keepers-prsa-annual-report-2016.pdf> [Accessed 4 Sep. 2017].

The National Records of Scotland. (2017). Model Plan Guidance to Element 7 | National Records of Scotland. [online] Available at: <https://www.nrscotland.gov.uk/record-keeping/public-records-scotland-act-2011/resources/model-records-management-plan/model-plan-guidance-to-element-7> [Accessed 4 Sep. 2017].

The PDF Association. (2017). The PDF Association. [online] Available at: <https://www.pdfa.org/pdf-association-2/> [Accessed 4 Sep. 2017].

veraPDF Consortium (2017). veraPDF. [online] Available at: <http://verapdf.org/> [Accessed 4 Sep. 2017].

veraPDF Consortium. (2017). veraPDF Docs | Desktop GUI Quick Start Guide. [online] Available at: <http://docs.verapdf.org/gui/> [Accessed 4 Sep. 2017].

veraPDF Consortium. (2017). veraPDF Docs | CLI Quick Start Guide. [online] Available at: <http://docs.verapdf.org/cli/> [Accessed 4 Sep. 2017].

veraPDF Consortium. (2017). veraPDF Software Releases. [online] Available at: <http://verapdf.org/software> [Accessed 4 Sep. 2017].

veraPDF Consortium. (2017). veraPDF. [online] Available at: <http://staging.verapdf.org/> [Accessed 2 Aug. 2016].

veraPDF Consortium. (2017). veraPDF Roadmap. [online] Available at: <http://verapdf.org/roadmap/> [Accessed 4 Sep. 2017].

Waters, D. and Garrett, J. (1996). Preserving Digital Information: Report of the Task Force on Archiving of Digital Information. [online] Available at: <https://www.clir.org/pubs/reports/pub63watersgarrett.pdf> [Accessed 4 Sep. 2017].

Web.library.yale.edu. (2017). *Born digital / Yale University Library*. [online] Available at: <http://web.library.yale.edu/digital-initiatives/digitization-standards-and-guidelines/born-digital> [Accessed 4 Sep. 2017].

Wheatley, P., May, P., Pennock, M. and Whibley, S. (2015). PDF format preservation assessment [V 1.3]. British Library Digital Preservation team assessments. [online] The British Library. Available at: http://wiki.dpconline.org/images/e/e8/PDF_Assessment_v1.3.pdf [Accessed 4 Sep. 2017].

Wright, P. (2017). NRS Digital Preservation for local authorities guidance. [online] National Records for Scotland Presentation, Glasgow, 10th July 2017. Available at: <http://www.scottisharchives.org.uk/sff/blog/cohort-3/penny-wright-digital-preservation-guidance.pdf> [Accessed 4 Sep. 2017].

Wright, R., Miller, A. and Addis, M. (2017). The Significance of Storage in the “Cost of Risk” of Digital Preservation. [online] Ijdc.net. Available at: <http://www.ijdc.net/index.php/ijdc/article/view/138/173> [Accessed 4 Sep. 2017].

W3schools.com. (2017). XML Tutorial. [online] Available at: <https://www.w3schools.com/xml/> [Accessed 4 Sep. 2017].

Yeo, G. and Shepherd, E. (2005). Managing records: a handbook of principles and practice. London: Facet Publishing.

Appendices

- I NRS: Skills for the Future Local Authority projects Executive Summary
- II Digital Preservation Student Project
- III veraPDF validation report results – Inverclyde Archives sample files
- IV veraPDF validation report results - Created sample files

Appendix i: National Records of Scotland: Skills for the Future Local Authority projects

Counting the Bits – Local Authority Capacity Planning and Good Foundations – Local Authority Digital Preservation Guidance

Executive Summary

Within Scottish local authorities we see a channel shift in service provision which has resulted in a growing number of digital by default records. Local authority record keepers deal with increasingly federated record creation systems which often support common functions and produce national datasets as a result of local decisions.

The National Records of Scotland (NRS) Digital Preservation Programme will lead two collaborative projects which focus entirely on supporting and enabling Scottish local authorities to deliver cost effective digital preservation solutions. While there is a plethora of digital preservation guidance available, very little of it is focused on either a Scottish or a local authority context.

The project Counting the Bits – Local Authority Capacity Planning will take the methodology developed by NRS to quantify the proportion of digital records selected for permanent preservation which it can expect to receive up to 2025, and optimise the technique for use in a Scottish local authority context. For local authorities the project outcomes will mean that they will be better able to understand and plan for their fast-growing, complex business capacity requirements. A further benefit of the project will be the creation of an evidence base quantifying the digital records within local authorities in Scotland which have long-term (archival) value. This could be used to inform future projects in this area.

The project Good Foundations – Local Authority Digital Preservation Guidance will develop a suite of practical digital preservation guidance focused on high priority record areas and systems common across local authorities. The package is likely to cover: planning, education, social care and the administration of meetings. Project benefits for local authorities will include: an improved ability to meet the requirements of the Public Record (Scotland) Act 2011 Element 7 (Archiving and Transfer Arrangements) for digital records; digital preservation efficiencies through joint working and sharing of expertise; improved business continuity and more streamlined legislative compliance.

Both projects will operate on the basis of a 'discovery' phase followed by a workshop(s) to evaluate and sign-off findings. A rough indication of the time commitment for a project strand (eg Good Foundations, social care) would be: one to two days for the 'discovery phase' and a day for the workshop. There is no cost to take part but participating authorities will be expected to make key staff (eg system administrators, system owners, records managers, archivists) available to the projects at agreed points. Different project elements may involve different authorities and different key staff. For instance, an authority may participate in Counting the Bits only, or in the Good Foundations topics where they can add most value.

The projects are designed so that all Scottish local authorities can realise their benefits. Authorities that choose to actively participate in the projects will act as pace setters. By contributing their local good practice and expertise pace setter authorities will accrue additional benefits through actively shaping the projects and their outcomes.

For further information, or to participate, please contact Susan Corrigall, Business Change Manager, Digital Preservation Programme, National Records of Scotland (susan.corrigall@nrsotland.gov.uk).

Solving the PDF Preservation Problem

Summary

This project represents an exciting opportunity to take on a substantial digital preservation challenge and help define best practice for memory institutions around the world. You will work closely with the Digital Preservation Coalition, its members, and other international partners to apply preservation tools, analyse what they report, investigate preservation risks and develop best practice for their use. The project will provide the opportunity to gain hands on digital preservation experience while working with world leading digital preservation organisations.

The challenge

Despite being one of the most popular file formats held by memory organisations, PDF remains a complex and uncertain preservation challenge. A number of problems have contributed to this situation:

1. There are a vast number of different applications that can create PDF files¹
2. The PDF specification has lacked sufficient precision to avoid diverging implementations
3. PDF viewers have become more tolerant of badly constructed PDF files²

This has led to a vicious cycle, fuelling a deterioration in the quality of PDF files in circulation. In turn this has exacerbated existing concerns about long term preservation and our ability to view and gain access to the information within PDF files in the future.

Although there is some understanding of the likely preservation risks to be found in PDFs³, the absence of a reliable, precise and clearly understandable validation tool has left a considerable knowledge gap. Memory organisations are currently unable to check their PDFs for preservation risks and are uncertain what risks to look for. In partnership with the veraPDF Project⁴, this student project will aim to solve this challenge.

The project

The veraPDF consortium, with support from Preforma and the European Commission, is developing a set of new PDF validation tools. These tools provide the ability to check a particular PDF against one of the PDF standards, and report any areas of non-compliance. The student undertaking this project will trial the veraPDF toolset and generate validation data from large collections of PDF files. This data will be analysed to seek out problematic PDF files which will then be investigated for preservation risks. As well as generating useful feedback on the design and effectiveness of the veraPDF tools, the analysis work will be used to

1 Formats over Time: Exploring UK Web History, Andrew N. Jackson, <http://arxiv.org/pdf/1210.1714v1.pdf>

2 "More What You'd Call 'Guidelines' Than Actual Rules"[1]: Variation in the Use of

Standards, Sheila M. Morrissey, <http://dx.doi.org/10.3998/3336451.0014.104>
3 http://wiki.dpconline.org/images/e/e8/PDF_Assessment_v1.3.pdf
4 <http://verapdf.org/>

gain a greater understanding of the impact and incidence of preservation risks in typical PDF files archived by memory institutions. This evidence base will support these institutions (and veraPDF) in developing PDF best practice, and ultimately formal preservation policies for handling PDFs for the long term.

What will trialling veraPDF involve?

The work is likely to be led by what is discovered in the research, so it is difficult to precisely describe what it will involve. However, a particular focus will be made on identifying the incidence of preservation issues in collections of real PDFs and then investigating the likely impact of that incidence. So for example, if a PDF uses fonts that have not been embedded, this *could* have an impact on the preservation of that PDF and whether the information within it can be used in the future. An investigation could begin by running VeraPDF on a large collection of PDF files. Those PDF files without embedded fonts could then be identified by analysing the metadata generated by veraPDF. These PDFs could then be examined to see if the lack of font embedding is likely to have an impact on their preservation.

Supervision

Supervision will be provided by Sharon McMeekin, Head of Workforce Development at the DPC, and Paul Wheatley (@prwheatley), Head of Research and Practice at the DPC.

Timing, effort and location

The project will need to run during 2016. The length of the project can be adjusted to suit particular circumstances of the person undertaking it. As a minimum, a number of weeks would be necessary. The open ended nature of the problem means that sufficient research could be found for a far longer project. The work can be undertaken remotely, although some site visits may be necessary. There may also be an opportunity to present results at a conference or other event.

Technical skills required

The following technical skills will be required by the student undertaking the project, although note that the supervisors will be able to provide some support:

- Ability to install and execute simple command line applications
- Ability to process, manipulate and analyse structured data, particularly XML
- Attention to detail, methodical working

Applying to Participate

See HATII General Information 2015-2016 Moodle section on Information Management and Preservation. Enquiries can be sent to Sharon McMeekin (sharon.mcmeekin@dpconline.org).

veraPDF validation report results – Sample of 100 PDFs obtained from Inverclyde Archives

File name	Statement	Passed Rules	Failed Rules	Passed Checks	Failed Checks
IT0028.pdf	FALSE	98	4	1159	25
100616.pdf	TRUE	105	0	2372	0
trainingchecklist.pdf	FALSE	99	3	456	29
media_list.pdf	FALSE	92	10	23265	3675
SKM_09754.pdf	FALSE	97	5	1866	297
draftsubmission.pdf	FALSE	99	3	65788	293
Handling_Guidelines.pdf	FALSE	98	4	628252	16282
enquiry_slater.pdf	FALSE	100	2	4656	1004
stats0616.pdf	FALSE	88	14	23200	265
FOISA-guidance-for-local-authorities.pdf	TRUE	105	0	2541	0
Web-policy.pdf	TRUE	105	0	16732	0
socialmedialist.pdf	FALSE	99	3	203	196
localstudies(1).pdf	FALSE	95	7	558	232
Christies.pdf	FALSE	83	19	11368	26472
AlanParkPatoninfo.pdf	FALSE	100	2	851	277
localstudiescollection.pdf	FALSE	86	16	23574	405

File name	Statement	Passed Rules	Failed Rules	Passed Checks	Failed Checks
ASLAWGlist.pdf	FALSE	99	3	591	3409
working_group_schedule.pdf	FALSE	98	4	993	35
localanniversaries.pdf	FALSE	95	10	304	21
WattLibraryHistory.pdf	FALSE	93	9	1165	3986
HLFcopy.pdf	FALSE	99	3	1056	329
FNLreport.pdf	FALSE	100	2	457	83
ASLAWGnotesLM.pdf	FALSE	98	4	3688	47
Grimshawlist.pdf	FALSE	87	15	234	56
keycabinetlegend.pdf	FALSE	85	17	462	18
Schedule0716.pdf	FALSE	102	3	2109	265
GIRFECCC.pdf	TRUE	105	0	1388	0
HarbourTrustBox1.pdf	FALSE	98	4	231	91
HarbourTrustBox2.pdf	FALSE	97	5	258	87
TI2359.pdf	FALSE	97	5	63271	5900
media_06532.pdf	FALSE	98	4	36573	2988
policy_for_retention_and_disposal.pdf	TRUE	105	0	3654	0
SKM_0954.pdf	FALSE	99	3	1683	34
loan agreement.pdf	TRUE	105	0	1265	0

File name	Statement	Passed Rules	Failed Rules	Passed Checks	Failed Checks
draft_transfer.pdf	FALSE	93	9	11385	404
GYR5578.pdf	FALSE	97	5	4365	993
wattlibraryinfo.pdf	FALSE	89	13	2168	2844
accidentreportblank.pdf	FALSE	99	3	4005	159
localhistoryresources.pdf	FALSE	88	14	670	92
IHS3522.pdf	FALSE	98	4	3659	663
HAA106.pdf	FALSE	97	5	15957	3488
GRM-HF1001.pdf	FALSE	97	5	5428	4847
wattlibraryarchives.pdf	FALSE	88	14	644	1103
JamesWattDock.pdf	FALSE	98	4	1223	896
ARASFNP.pdf	FALSE	89	13	6960	790
FNApplicationletter.pdf	FALSE	99	3	12559	584
NLStrainingagenda.pdf	FALSE	98	4	9743	989
ICequality.pdf	FALSE	87	15	1253	4847
historicalboundaries.pdf	FALSE	85	17	1706	170
jointstrategy.pdf	FALSE	101	1	3965	102
statement_of_liability.pdf	FALSE	90	12	4720	227
media_345.pdf	FALSE	93	9	17739	23051

File name	Statement	Passed Rules	Failed Rules	Passed Checks	Failed Checks
chapter1.pdf	FALSE	99	3	653	41
hub_timetable.pdf	FALSE	99	3	7650	2604
rota_final.pdf	FALSE	98	4	5530	75
expenses_0516.pdf	FALSE	96	8	2512	147
cometpamphlet.pdf	FALSE	96	8	5607	78
watthistory.pdf	FALSE	99	3	12468	105
download03.pdf	FALSE	103	2	11034	674
lcplannotes.pdf	FALSE	98	4	3351	591
businessstoreplan.pdf	FALSE	90	12	12468	9855
performanceblank.pdf	FALSE	98	4	31276	35388
openinghrs.pdf	FALSE	96	9	26442	986
librarycontacts.pdf	FALSE	94	8	21167	395
accessionform.pdf	FALSE	100	2	1159	25
valuationlist.pdf	FALSE	90	12	23133	2884
TOIL_LM.pdf	FALSE	93	9	4004	191
museum_collections.pdf	FALSE	95	7	648252	16180
PhilSoc2016.pdf	FALSE	97	5	8973	429
performance appraisal.pdf	FALSE	95	7	12664	8754

File name	Statement	Passed Rules	Failed Rules	Passed Checks	Failed Checks
ICPlanning.pdf	FALSE	90	12	165699	5477
ICMuseum.pdf	FALSE	89	13	4376	608
Volunteerweekplan.pdf	FALSE	87	15	13499	6411
volunteerlog.pdf	FALSE	93	9	8764	802
volunteer_protocol.pdf	FALSE	101	1	35701	4550
ingressprocedure.pdf	FALSE	100	2	8733	397
expenses_0616.pdf	FALSE	101	1	650	74
draft_localhistory.pdf	FALSE	98	4	9972	581
doors_open.pdf	FALSE	92	10	5876	15
handbookV1.pdf	FALSE	94	8	12477	943
SSminutesdraft.pdf	FALSE	95	7	1973	2309
media_2345.pdf	FALSE	95	7	7465	132
PrincesPier.pdf	FALSE	99	3	4325	567
BWS events.pdf	FALSE	100	2	19221	14325
Blore_bio.pdf	FALSE	97	5	981	2488
McLeanINFO.pdf	FALSE	96	6	5693	310
PD5454 info.pdf	FALSE	99	3	462	1189
searchroomlayout.pdf	FALSE	91	11	1547	98

File name	Statement	Passed Rules	Failed Rules	Passed Checks	Failed Checks
searchroom guidelines.pdf	FALSE	99	3	1222	301
transferpolicy.pdf	FALSE	97	5	4753	690
leaflet_watt.pdf	FALSE	95	7	899	1476
reminiscence group list.pdf	FALSE	96	6	12658	235
media_5454(1).pdf	FALSE	98	4	259	27
dementia awareness.pdf	TRUE	105	0	1632	0
media_1236.pdf	FALSE	96	6	643	14
digitalpolicy.pdf	TRUE	105	0	1674	0
17655D.pdf	TRUE	105	0	2324	0
volunteer agreement(1).pdf	FALSE	95	7	2354	73
local studies 2015.pdf	FALSE	75	27	133	1
accession 2017 01.pdf	FALSE	84	8	23267	35642

Created sample files – veraPDF validation report results

Unique Identifier	File Name	Method of Creation	Additional Information	File Size	Statement	Passed Rules	Failed Rules	Passed Checks	Failed Checks
PDF01	SKM_C224e17083113580.pdf	scanned from Konika MFD to PDF	contains images and text	379KB	FALSE	99	3	177	3
PDF02	DARK SIDE O.pdf	created using PDF MS Word	Contains images and text	1.86MB	TRUE	105	0	3183	0
PDF03	sample text pdfa.pdf	created using MS Word 2010	contains text - arial font only	103KB	TRUE	105	0	4017	0
PDF04	sample text not pdfa.pdf	created using MS Word 2010	contains text - unchecked PDF/A box	94.8KB	FALSE	96	6	3896	43
PDF05	sample hyperlinks.pdf	created using MS Word 2010	contains text and hyperlinks	99.1KB	FALSE	104	1	3200	2
PDF06	sample special characters.pdf	created using MS Word 2010	contains text and special characters	99.1KB	FALSE	104	1	3200	2
PDF07	Launch program.pdf	created using MS Powerpoint 2010	contains images and text	251KB	TRUE	105	0	2986	0
PDF08	Book1.pdf	created using MS Excel 2010	contains text within fields	193KB	TRUE	105	0	4847	0