Sustainability Assessments at the British Library: Formats, Frameworks, & Findings.

Maureen Pennock
The British Library
Boston Spa
West Yorkshire
+ 44 (0)1937 546302
Maureen.Pennock@bl.uk

Paul Wheatley
Paul Wheatley Consulting Limited
Leeds
West Yorkshire
@prwheatley
paulrobertwheatley@gmail.com

Peter May
The British Library
Euston Road
London
+44 (0)20 7412 7199
Peter.May@bl.uk

ABSTRACT

File format assessments have been the subject of much debate in and outside of the preservation community in the past decade. Recognizing the unique structural, operational, and collecting context of the British Library, the Library's digital preservation team recently initiated new format assessment work to deliver recommendations on which file formats will best enable the preservation of integral, authentic representations of British Library collection content over the long term. This paper describes the work carried out to review previous assessments, identify appropriate sustainability categories and newly assess formats accordingly.

We posit that the relatively 'fuzzy' nature of a file format requires a relatively open-ended assessment framework and a nuanced understanding of preservation risk that does not solely lie with 'all-or-nothing' format obsolescence. We review other work in this area and suggest that whilst previous format assessment work has addressed a range of subtly different aims, experience has since indicated that some of the criteria used - such as considering number of pages in a format specification as a measure of complexity - may be invalid. British Library assessments are made on documented points of principle, for example, an emphasis on evidence-based preservation risks and the avoidance of numerical scores leading to comparisons between formats, and these have formed the base upon which sustainability categories are defined. We present these categories, which help to identify preservation risks or other challenges in the management of digital collections, and provide an overview of initial assessments of three formats: TIFF, JP2, and PDF. We acknowledge however, that implementation of preservation requirements, e.g., the use of particular preservation-justified file formats, must be balanced against other business requirements, such as storage costs and access needs, and argue that transparency of this format assessment process is fundamental if the resulting recommendations are to be fully understood in the future.

General Terms

Preservation strategies and workflows, specialist content types, case studies and best practice

Keywords

British Library, file formats, sustainability, assessments, transparency, preservation master

1. INTRODUCTION

The British Library is increasingly a digital library. Our long term digital repository already holds over 11,500,000 items and more are added every day. With acquisition comes responsibility: we must preserve and make this content accessible for our future users - as a national library, this is at the heart of our mission. Yet preservation of digital content is not straightforward, requiring action and intervention throughout the lifecycle far earlier and more frequently than for our physical collection. The digital preservation team at the British Library is responsible for addressing this to ensure that despite the challenges, we are able to preserve our digital collections for the very long term.

The nature of the work carried out by the digital preservation team has changed since it was established in 2005. This is due in part to changes in leadership and organisational structure, but more significantly as a result of growth in our knowledge and changes in operational context. Furthermore, our digital library system has matured significantly in the past eight years, as has our understanding of key non- or semi-technical digital preservation needs in the Library. In 2013, the Library published a new digital preservation strategy that addressed these changes. The strategy identified four strategic priorities that must be addressed by 2016 [1]:

1. Ensure our digital repository can store and preserve our collections for the long term - enhancing its preservation capabilities and devising preservation plans for collections stored within;

2. Manage the risks and challenges associated with digital preservation throughout the digital collection content lifecycle - clearly defining our preservation requirements and implementing preservation risk management practices across the lifecycle;

3: Embed digital sustainability as an organisational principle for digital library planning and development - planning and budgeting for preservation and sustainability from the point of acquisition;

4: Benefit from collaboration with other national and international institutions on digital preservation initiatives - embarking on appropriate collaborative endeavours and achieving maximum return on investment in terms of time, effort and financial commitment.

These strategic priorities are addressed in a programme of work led by the digital preservation team that identifies and aligns eleven core workstreams with one or more priorities. Workstreams are highly interdependent; most are collaborative and require input from colleagues in other areas of the Library (e.g. curators, content owners, developers, architects, and processing staff), though a small number are driven and delivered by the digital preservation team alone.

The remainder of this paper is focused on the *File Format Assessment* workstream, which is delivered by the digital preservation team and aligns primarily with strategic priorities one

and two. The workstream assesses file formats for long term preservation risks and identifies preferred formats for the preservation of collection content stored in our long term digital repository. It should be noted that the File Format Assessment work described in this paper is only one of several activities (including Policy Development and Collection Profiling) that provides input to preservation planning exercises. File format assessments should not be used in isolation to drive preservation decision making.

2. FILE FORMATS & LONG TERM PRESERVATION

Despite many years of global digital preservation research, experimentation and practice, fundamental questions about file formats and long term preservation remain under discussion. This section will attempt to assess work, thought and comment from the wider digital preservation community in order to inform a sensible and practical approach to assessing file formats and ultimately preserving digital collections.

2.1 What is a "File Format"?

A number of sources in the digital preservation sphere, for example the Global Digital Format Registry (GDFR) [2], have defined a file format as a representation of an information model, typically with an implied assumption that a file format is a method of structuring information in a sensible way for storage and ultimately retrieval and use. In the case of some file formats, such as TIFF[1], specifications have been created that do describe a reasonably sensible information model, as well as how it should be realised into an instance of the format. This concept has been identified and exploited for preservation purposes and is evident in the design and usage of the JHOVE tool, which compares a file against its respective file format specification and reports discrepancies.

More recently, some within the preservation community have observed that the software that is used to create instances of file formats also plays a role in defining what a file format is. Furthermore, a reference implementation of a viewer for a particular format could provide a different definition of the format itself. For example, Sheila Morrissey describes the "violations" of Adobe's specification for PDF that are tolerated by the Adobe Acrobat Reader [3]. Some of these were described in an appendix to the PDF specification, but were subsequently removed when PDF received ISO standardization. Morrissey states "...these notes, while helpful, beg the question as to what we are to consider authoritative with respect to PDF format instances: the specification, or the behavior of the Acrobat reader application."

An alternative definition proposed by Andy Jackson defines a file format as "a formal language defined for the purpose of persisting and transmitting the state of computer programmes" [4]. This position has been illustrated particularly well with extreme examples, such as that of the early binary office formats which effectively provided a dump of the application's internal data structures [5]. Rather than representing a cleanly structured information model, these formats were little more than a dump of application memory to enable faster loading and saving on sluggish 1990's-era PC hardware. The lack of a preservation community created format validation capability is hardly surprising in these cases.

Defining an appropriate scope for what we understand as a single file format is challenging. Many versions of a single format can exist, sometimes maintaining a degree of backward compatibility but sometimes involving wholesale redesigns over time (e.g., Office formats). Other formats allow embedding or attaching of yet other formats, leading to the possibility of veritable Pandora's Boxes of multi-format data waiting to be opened by reluctant preservationists.

Clearly the concept of a file format is difficult to tie down, and is perhaps most usefully considered as a somewhat amorphous entity. Assessment mechanisms (and indeed the preservation work they inform) will therefore need to take into account the somewhat imprecise nature of the main target of this work.

2.2 What is File Format Obsolescence? Does it Exist? And if so, to what Extent?

The digital preservation challenge was clearly identified and addressed by the Museums, Libraries and Archives (MLA) community in the latter part of the 1990s. A central theme that emerged from this early work was the danger of format obsolescence. This was characterised in a widely referenced piece by Jeff Rothenberg in Scientific American in 1995 where he stated "digital documents are evolving so rapidly that shifts in the forms of documents must inevitably arise. New forms do not necessarily subsume their predecessors or provide compatibility with previous formats" [6]. At the time, the IT market was emerging from an era characterised by a multitude of computer platforms, many of which had disappeared in a relatively short space of time. This was particularly evident in the home computing market. In this climate, the message that file formats were at risk of obsolescence unsurprisingly took hold. It can still be found today as a core part of many digital preservation training resources.

In the last few years a more sceptical view of file format obsolescence has emerged. David Rosenthal has made the case that format obsolescence simply doesn't exist, and references web-era work that provides some evidence to back up this position [7]. Evidence that makes a case for the format obsolescence lobby is harder to come by. Extreme examples sought and investigated by Chris Rusbridge were quite quickly solvable with help from colleagues and other expertise via the internet [8]. Rusbridge states "It's worth noting that a lot of the 'official' advice on obsolescence that you might find is useless. Various sites will classify formats as obsolete that are still perfectly easy to open and migrate from. Indeed, I suspect that there's no really helpful way to classify obsolescence (I tried and failed)".

Working on the assumption that data in the vast majority of file formats will be readable with some degree of effort does not take into account two crucial issues. Firstly, what is the degree of effort to enable rendering, and what does it mean for an organisation such as the British Library? Secondly, even if a file format is readable, is the resulting rendering, migration or indeed emulation, anything like an authentic reproduction of the original?

As a national memory institution, the British Library must ensure that collections are accessible for future generations. The term "institutionally obsolete" suggests a file format that may be accessible with further effort but will not run on a typical (or perhaps vanilla) computer platform provided by an institution [9]. In terms of the British Library this may relate to the platforms provided in our reading rooms or assumptions made about software available for those accessing our collections remotely[2]. Addressing this challenge may not be straightforward and has been taken into account in the assessment methodology on which this document focuses.

[1] Where interchange between software applications and the need to address the lack of an appropriate non-proprietary still image format was seen as a key aim in its conception.

[2] Increasingly, this means a web browser.

A number of studies have examined the impact of changing methods of rendering over time, where file formats may still be accessible, but with perhaps some degree of change in the results. These include the work of the Digitale Bewaring Project [10] and the Rendering Matters report which concludes that the "choice of rendering environment (software) used to open or "render" an office file invariably has an impact on the information presented through that rendering. When files are rendered in environments that differ from the original then they will often present altered information to the user. In some cases the information presented can differ from the original in ways that may be considered significant" [11]. The effects of the rendering process or environment on files (of particular formats) must be taken into account when considering the viability of a given preservation approach. What aspects of a digital collection item must be preserved and how can a given format support that?

In the necessarily conservative domain of digital preservation, it seems unwise to completely dismiss a concept such as format obsolescence on the evidence presented. However there are genuine and significant preservation risks beyond the black and white delineation of format survivability and they should be taken into account in the assessment methodology.

2.3 The Role of 'Preservation Masters'

It is not uncommon for legal deposit legislation to stipulate that hard copy deposits must be the best available edition of a work[3]. The term 'best' is open to interpretation, though in Library contexts it is generally taken to mean content of the highest quality and most suitable for purpose. For example, archival-quality paper is preferred over low-grade paper, large size books are often preferred over small ones, complete versions are generally preferred over partial ones, and originals are preferred to copies[4]. Best editions are generally selected for their longevity and usability, both of which are important selection criteria for Libraries operating over the very long term.

'Best' editions remain significant in a digital environment. Digital content is liable to degrade in a similar fashion to hard copy, though in a shorter time frame, and although institutional obsolescence may not be imminent, it is inevitable eventually. The potential longevity of content is an essential consideration in institutions preserving for the long term. The same may be said of usability, where high-quality reproducibility and mutability, automated analysis, detailed searching and content enhancement all offer far more potential to the user than with physical copies. Our experience at the British Library is that in a digital environment, versions of collection items are often differentiated by format or format resolution, making format a key factor in determining best quality.

Preservation Masters play the role of our 'best' available digital editions at the British Library. The concept of a Preservation Master is not new, existing already for both physical and digital collections[5]. Preservation Masters are rich representations of a digital collection item with high levels of information content, which serve to meet both preservation needs and user needs by enabling the creation of derived files with minimum loss.

3. FORMAT ASSESSMENTS ELSEWHERE

File format assessments as a means to guide preservation activities have been ongoing in the preservation community since the latter part of the 1990s. They remain a hot topic in the community at the time of writing:

- The SCAPE Project presented a paper describing the File Format Metadata Aggregator (FFMA), an expert system to collate and assess file format information at iPRES2013 [12];
- The University of North Carolina is conducting research to gather expert opinion on file format risk;
- The 2014 National Agenda for Digital Stewardship identified "File Format Action Plan Development" as a specific priority "for infrastructure investment" [13];
- Lee Nilsson, the National Digital Stewardship Resident at the Library of Congress, recently provided an introduction to "File Format Action Plans", which references much of the existing work in this area [14]. While not adding much new to the debate, it does indicate a commitment to follow up on the priority identified by the Library of Congress.

File format assessments have, however, been emerging for a number of years. Other notable work includes:

- The Florida Centre for Library Automation's File Format Background assessments and quite practically focused Action Plans that were developed from 2003 [15];
- The Library of Congress's widely referenced File Format Sustainability Factors [16];
- The National Library of Australia's AONS work, that attempted to score preservation worthiness [17]. The NLA subsequently moved away from this approach;
- Archivematica which realises file format migration on ingest (sometimes referred to as normalisation) based on a Format Policy Registry [18];
- Far less detailed file format guidance (albeit with obvious elements that can be traced back to the more comprehensive works referenced above) can be seen on innumerable sites across the web, for example the MIT Libraries Formats for Long-Term Access [19].

3.1 Theory versus Evidence

Johan van der Knijff notes that the criteria used in assessment approaches, such as that of the Library of Congress and the UK National Archives, "are largely based on theoretical considerations, without being backed up by any empirical data. As a result, their predictive value is largely unknown" [20]. Whilst such theoretical considerations may seem convincing, basing recommendations on real-world evidence provides a much more reassuring approach to preserving digital collections.

Where automated, top down approaches (such as the FFMA expert system) have the potential to replace expert analysis, there is considerable danger of poor, or possibly even catastrophic, preservation actions being taken. There are a number of documented (and anecdotally many more undocumented) examples of PDF migration implemented to ensure JHOVE provided a "valid and well formed" validation result for each preserved file, where there was little or no evidence of the need to

[3] See for example http://www.bl.uk/aboutus/legaldeposit/printedpubs/depositprintedpubs/deposit.html, and http://www.slsa.sa.gov.au/site/page.cfm?c=4702.

[4] United States Copyright Office *Best Edition of Published Copyrighted Works for the Collections of the Library of Congress*: http://www.copyright.gov/circs/circ07b.pdf.

[5] See for example the Preservation Policy of the National Library of Australia, 4th Edition: http://www.nla.gov.au/policy-and-planning/digital-preservation-policy.

take action given the tolerance of PDF viewers to many of the issues JHOVE identifies [21]. Given the potential for loss of important data when unnecessary format migration is applied (particularly given the woefully inadequate facilities for verifying the accuracy or quality of format migrations), this is particularly concerning. Van der Knijff notes alarm at "recurring attempts at reducing format-specific preservation risks to numerical risk factors, scores and indices"[20]. He goes on to provide an example from his own institution where a format assessment model [22] led to the adoption of JP2 instead of TIFF as the preservation format for digitised still image masters. A number of JP2 format risks were simply unknown at the time of the assessment and only became clear when the organisation worked with the format in practice. Van der Knijff summarises that "None of these problems were accounted for by the earlier risk assessment method (and I have a hard time seeing how they ever could be)!" This also lends support for an evidence backed approach, making recommendations based on empirical results; however, care should still be taken not to simply reduce such evidence to a numerical comparison between formats.

Archivematica is an example of a preservation system that implements file format normalisation on ingest to a repository. The Archivematica Format Policy Registry identifies which formats should be normalised, separately noting formats used for preservation and access [18]. They state that their "preservation formats are all open standards. Additionally, the choice of preservation format is based on community best practices, availability of open-source normalization tools, and an analysis of the significant characteristics for each media type". While the Registry usefully links to further detail and results from small scale testing, some of the normalization operations recommended are known to result in loss of fidelity, for example, transforming PDFs to PDF/A (which precludes some interactive content and hence would lead to data loss in files should normalization occur) or transforming GIF to TIFF (where the latter does not support the more unique animation properties of the former). The Registry justifies the PDF transformation by noting that "PDF/A is the only version of PDF recommended for long-term preservation". In a study of file format guidance from academic repositories in the US, Rimkus *et al* [23] reflect on the significant impact of particular sources of guidance, such as the frequently referenced and reused MIT Libraries Formats for Long-Term Access. They go on to state: "Comments made by repository managers during the data gathering period would imply that Archivematica is poised to play a similar role for the growing number of institutions that deploy it....Several digital preservation managers referred to Archivematica's ongoing file format policy registry and associated migration paths as the policies they intended to adopt at their own institutions".

Malcolm Todd's Digital Preservation Coalition Technology Watch Report: "File Formats for Preservation"[24] engages in a detailed discussion on the weighting and reconciliation of numerical scores for assessing formats based on a variety of assessment work. It concludes with support for score-based approaches, though the viability of these was later cast into doubt by Van der Knijff after practical experience with the approach at the Koninklijke Bibliotheek (see above).

There are a number of examples in which the application of assessment factors stop short of examining the practicalities of working with the format, some of which are listed above. Where this practical evidence is not available, proxies have been used without evidence that they are indeed linked to preservation risk - for example, a count of the number of pages in a file format's specification, or the number of applications that support a particular format. The former gives an impression as a crude measure of "format complexity", but arguably nothing more. Counting the huge number of pages in OOXML documentation might perhaps provide some indication of the sheer vastness of these formats but nonetheless it is woefully inadequate as a comparative measure between formats. The latter, on the other hand, can be simply misleading as many applications could rely on a small number of software libraries.

3.2 Clarity of Purpose and Audience

Format guidance to date has appeared to focus on addressing a range of subtly different aims, sometimes without clarity as to what those aims actually are. These include:

- Guidance that records the level of support that will be provided to data preserved within a particular repository (typically ranging from some kind of guarantee or best effort, through to bit preservation only);
- Guidance that targets contributors to digital repositories, sometimes recommending formats in which particular types of data should be submitted;
- Guidance that targets data creators, recommending formats in which particular types of data should be created;
- Justification and guidance for repository/preservation managers in implementing recommendations, possibly addressing format migration or normalisation.

Where these aims are unclear or, perhaps even more significantly, the target audience of the guidance is unclear, the potential for misuse becomes real. This becomes especially concerning where guidance is re-used outside of its original context, such as by another organisation. As the examples in the previous section indicate, file format assessments and resulting guidance can have a significant impact within the wider community, leading to the possibility of mis-informed preservation choices.

3.3 Misleading Measures

Adoption rates and (self)-documentation are common features in the assessment frameworks mentioned above that can be misleading if not properly understood.

A reference to the availability of documentation can be found in most of the existing file format assessment work. In the UK National Archives' "Selecting File Formats for Long-Term Preservation" Adrian Brown states that the "availability of format documentation is not, in itself, sufficient; documentation must also be comprehensive, accurate and comprehensible. Specifically, it should be of sufficient quality to allow interpretation of objects in the format, either by a human user or through the development of new access software" [25]. Brown also suggests that a "detailed judgment of documentation quality will require evaluation of the documentation itself". However the only way to be sure that documentation is sufficiently complete to enable development of new access software would be to develop and test new access software from it. This is a costly approach. Documentation is undoubtedly beneficial to have in some circumstances, but assessing or rating the quality of documentation is clearly problematic and so use in assessing the sustainability of file formats requires careful consideration. As van der Knijff states: "A problem with errors and ambiguities in format specifications is that they can be incredibly easy to overlook, and you may only become aware of them after discovering that different software products interpret the specifications in slightly different ways" [26].

The value of self-documentation (where sufficient metadata is present to aid in understanding and/or use of the format, without the need for additional attached metadata) is debatable for collections that reside within a modern digital repository with comprehensive support for attached metadata. While embedded

metadata may provide some use in the event of catastrophic repository damage that might physically separate collections from their metadata, this is an eventuality that repository design, replication and backups aim to avoid. Conversely, where metadata is both embedded in a file and associated or attached in a repository, should it be kept consistent? To do so may require frequent modification to the collection object - a course of action in itself that introduces preservation risk, and hence is probably undesirable. If embedded and attached metadata is inconsistent its value becomes questionable. It therefore seems sensible not to take self-documentation into account in a format assessment of this kind.

Measuring "adoption" of a format in the wider world is clearly a difficult task. What level of adoption is sufficient? How might it be quantified? Observations about formats residing in niches, perhaps in conjunction with the availability or quality of software to render the format in question, could provide useful insight. The adoption of the JP2 format within the library community provides some interesting observations. At a digital preservation meeting at the Wellcome Library focusing on JP2 in 2010, comments from members of the audience suggested that a number of libraries within Europe had adopted JP2 "because that was what the British Library had done". It should be noted that the BL adopted JP2 for use in very specific high volume collections and otherwise still utilises TIFF. This example worryingly highlights the impact of hearsay and reputation over analysis and evidence. It also poses questions about analysis that might be based on generalised assessments of adoption. Despite growing numbers of MLA organizations adopting JP2 for storing digitized images (noting that the picture is somewhat muddied by JP2's attractiveness in not only reducing storage volume but also in potentially delivering content to remote users, thereby seeing some use as a preservation format, some as an access format and in some cases both), there remain serious concerns about the quality and sustainability of creation and access software [27]. Clearly measures of adoption in isolation can be misleading. Turning an impression of adoption into a numerical rating to facilitate relative scoring of formats could prove to be a dangerous approach. Approaches that draw conclusions based on surveys of existing advice should also be viewed with caution.

4. BRITISH LIBRARY FORMAT ASSESSMENT POINTS OF PRINCIPLE

Discussion around the issues above has been distilled into the following points of principle that inform the implementation of format assessments:

1. Clearly state the aims of the assessment, the target of resulting guidance and the circumstances within which guidance should be acted upon;
2. Be aware of the potential for file format obsolescence but proceed on the basis that catastrophic loss of access to a particular format will not usually be the most pressing preservation risk;
3. Published guidelines, policies and assessments have a ripple effect and are often reused without consideration of the underlying evidence or the influence of unique organisational requirements. Meta assessments that make recommendations based on surveys of what other organisations do, add a further level of obfuscation. Approach with caution

For assessments:

4. Focus on evidence-based preservation risks (for example, non-embedded fonts in PDF);
5. Focus on implications of institutional obsolescence which lead to issues maintaining the content over time;
6. Any recommendations to choose a preservation format different to the format in which the data was received must be backed up by strong empirical evidence of the benefits and risks involved;
7. Avoid assessment based on theoretical factors and avoid format-to-format comparisons using summarised sustainability factors (in particular numerical scoring based approaches).

On specific sustainability factors:

8. Measures of "documentation completeness" or quality are largely meaningless and should be avoided;
9. Self-documentation should not be considered as an assessment factor. Documentation availability should be considered with a view to supporting likely preservation processes rather than as a judgment of preservation worthiness.

Many other organisations have exactly the same challenges in a different context. Assessments are therefore undertaken in an open and collaborative manner in order to increase the effectiveness of the decision making (based on greater contribution from an array of expertise) and minimise the resources required from the British Library.

5. SUSTAINABILITY CATEGORIES

The British Library assessment of file formats against sustainability categories identifies areas for concern rather than rating a format on a comparative scale. Practical guidance on mitigation practices for areas of concern is provided at the end of each assessment, though it should be noted that the capability (e.g., appropriate software tools) will not always exist to address all areas of concern. In some cases it is necessary to identify instead areas for experimentation with software tools and their impact on sample collections.

In summary, each file format assessment aims to provide evidence-based recommendations around use of a specific format, including whether or not a format is suitable as a Preservation Master within the British Library. Risks of using the format are identified and initial mitigation advice listed. Where there is uncertainty, this is clearly stated.

Sustainability categories considered in the assessments are as follows:

Development Status: An overview of the history, ownership, and current status of the file format.

Adoption and Usage: An impression of how widely the file format is used, with reference to usage in other memory institutions and their practical experiences of working with the format.

Software Support: *Rendering Software Support* - an overall impression of software support for rendering the format with reference to a) typical desktop software and b) current support on British Library reading room PCs; *Preservation Software Support* - an impression of the availability and effectiveness of software for managing and preserving instances of the file format, including a) Format Identification, b) Validation and Detecting Preservation Risks, c) Conformance Checking, d) Metadata Extraction, and e) Migration.

Documentation and Guidance: An indication of the availability of practical documentation or guidance with specific reference to the facilitation of any recommended actions

Complexity: An impression of the complexity of the format with respect to the impact this is likely to have on the organisation managing or working with content in this format. What level of expertise in the format is required to have confidence in management and preservation?

Embedded or Attached Content: The potential for embedding or attaching files of similar or different formats, and the likely implications of this.

External Dependencies: An indication of the possibility of content external to an instance of the file format that is complimentary or even essential to the intellectual content of the instance.

Legal Issues: Legal impediments to the use, management or preservation of instances of the file format.

Technical Protection Mechanisms: Encryption, Digital Rights Management and any other technical mechanisms that might restrict usage, management or preservation of instances of the file format.

Other Preservation Risks: Other evidence based preservation risks, noting that many known preservation risks are format specific and do not easily fit under any of the sustainability categories above.

Categories were defined prior to assessment and without consideration of any specific formats, in order to deliver a 'vanilla' set with no specific format bias. The detail of each category has been elaborated upon as a result of our experience in the initial assessments, but none have been deleted.

6. RESULTS

Six formats have been assessed to date: TIFF, JP2, PDF (including PDF/A), NTF (Ordnance Survey), JATS and ePub. Assessments typically take between 4 – 6 working days to complete, including background research. Results are issued in the first instance in the form of a report, which is subsequently condensed into a summary table for clarity and ease of dissemination. Due to space restrictions in this paper it is not possible to include more than summary discussions for the first 3 formats assessed. The full reports will be published elsewhere by the British Library in due course.

The TIFF assessment concluded that TIFF remains reasonably well suited to the simple task of the storage of digitised preservation masters, despite lacking many new bitmap file format features that have developed to support advances in graphics applications since the last significant changes to the format. Although there are preservation concerns with less well supported features that were introduced in version 6, baseline tags are well supported by software and well tested by many users both within and beyond the MLA sector. Implementation of a TIFF parser/profile conformance checker of a similar form to Jpylyzer [28] would be useful in performing assessments of trial runs in new digitisation projects and allow automated checking of subsequent production runs to the same standards. Detection of poorly supported TIFF extensions would also enable identification of problem content in deposited collections. Further investigation and/or collaboration with institutions interested in developing a "TIFFylyzer" and developers of the Kost-val validation application [29] should be explored.

JP2 fared less favourably than TIFF as a format for digitised preservation masters. Based on the evidence collected, the assessment concluded that JP2 is undesirable from a purely preservation-oriented perspective. JP2 is a niche format that has failed to see widespread adoption. As a consequence there is poor tool support and significant numbers of issues have been reported, despite the low rate of adoption. Obvious bugs in both the format and in software were not fixed before the preservation community adopted JP2 [30]. It is hoped that growing use by memory organisations and associated experience in working with JP2 will eventually lead to mitigation of most issues, but other problems may remain. In the meantime, if the benefits of JP2 (compression and delivery) are sufficient that it remains a desirable solution for storing digitised preservation masters, use of the format must be considered a significant risk. Ideally, mitigation of this risk requires investment in tools such as OpenJPEG to address the tool support concerns, and very thorough checking of all files in production settings. Mitigating JP2 preservation concerns comes with an associated cost and this should be taken into consideration in preservation planning activities where storage cost savings are likely to be significant.

PDF is a ubiquitous format in the contemporary computing world but widespread adoption, usage and software support has not led to the universal mitigation of preservation risks associated with this format. PDF files are frequently found to be invalid or badly formed and whilst the tolerance of most PDF rendering applications makes the impact of this situation difficult to measure, it should nonetheless raise a red-flag for preservation over the long term. A number of the other identified PDF risks have the potential to be catastrophic from a preservation point of view (such as encryption or missing font information, which could prevent access to content altogether). Strengthening our ability to detect these risks and ultimately developing trusted (and verifiable) means of fixing these issues in PDF files will be essential. That said, the severity and frequency of the risks identified in the full report remain relatively poorly understood. Existing published research has only begun to scratch the surface in revealing how these risks may affect an archive collection of PDF files (or not, as the case may be!). Research to apply validation tools to collections in order to more clearly identify genuinely problematic PDFs, or indeed discount identified risks whose frequency or impact is not significant, would help considerably to inform handling guidelines and potentially avoid overly prescriptive and potentially costly PDF fixing that has been adopted by some organisations. Testing of this sort is expected to take place over the course of 2014/15 in a Tool Assessment workstream, using collections identified as part of a Collection Profiling exercise (the subject of another paper submitted to iPRES 2014). The nature of the restrictions in PDF/A preclude preservation of some functionality and therefore its application will not necessarily suit every use case. For example, wholesale migration of a PDF collection to one of the PDF/A versions is unwise as functionality such as audio and video will be discarded. However, receipt of deposit of a PDF/A-1 may not raise significant preservation concerns as the PDF/A restrictions prohibit functionality associated with the preservation risks identified in the assessment - assuming of course that the PDF/A-1 files do indeed conform to the restrictions described in the PDF/A-1 standard. This is nonetheless a potentially dangerous assumption and one that may be difficult to test given concerns about PDF/A validation.

7. CONCLUSIONS

It is clear from reviews of earlier work that proxy measures of preservation risks are insufficient to capture the subtleties involved in practical digital collection management and long term preservation. Format assessments should be informed by thorough practical considerations and, insofar as is possible with long term investigations without a crystal ball, empirical evidence. This will only be possible at scale in a global community if we share not only our findings but also our aims, our context, and our underlying data. Otherwise we are doomed to repeat our failings. The conclusions of this work concerning the JP2 format are, we hope, an alarm bell for institutions choosing to preserve in this

format primarily on the basis that others are doing so. Preservation Masters are the files from which future iterations of a digital collection item will be generated, and it is essential that their selection is fully informed.

Considering the applicability of the assessments to date to a much bigger and heterogeneous digital collection, as is the case at the British Library, it is further noted that assessments based around file formats *alone* reveal only some of the critical preservation issues that need to be addressed. Many digital collection items are compound in nature and may consist of a number of files, each possibly of a different format. Consideration must be given to all formats, their inter-relationships, and the compound object, for an assessment to be valid. The potential for a format to store different types of content must also be accounted for, as formats for digitised still images may likely have different requirements to formats for digitised manuscripts or born-digital images. Assessments of this sort are, however, the first step along that road and remain essential for memory institutions to understand why a given format is preferred over another, particularly those institutions with a mandate to preserve for the very long term. Transparency of the process is key to that understanding.

Finally, we observe the importance of the action taken as a result of an assessment. This work suggests a new and more nuanced approach is necessary to avoid the comparative scoring of format against format and the focus on format obsolescence without consideration for more subtle and pressing preservation risks. Assessments can provide an invaluable steer to essential preservation activities. This could take the form of specific handling guidance to mitigate clearly identified preservation risks, identification of preferred deposit formats for different types of content, further research and practical testing to fill gaps in existing understanding, or engagement with the responsible owner of a format to provide feedback on file format specification errors or ambiguities.

Ultimately a Preservation Master, with respect to a particular collection, can only be established through an effective preservation planning activity in which file format assessments provide only one of many essential information inputs.

8. ACKNOWLEDGMENTS

Thanks go to Johan van der Knijff of the National Library of the Netherlands for his tireless investigative work on which many aspects of this format assessment work were based. Thanks also go to Sheila Morrissey of Portico for her PDF expertise.

9. REFERENCES

[1] Pennock, M. 2012. *British Library Digital Preservation Strategy, 2013 – 2016*. British Library (May 2012). URL= http://www.bl.uk/aboutus/stratpolprog/collectioncare/discovermore/digitalpreservation/strategy/BL_DigitalPreservationStrategy_2013-16-external.pdf.

[2] The GDFR Ontology defined a format as "a byte-serialized encoding of an information model". The document is no longer available but is referenced here: http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml.

[3] Morrissey, S. M. 2012. The Network is the Format: PDF and the Long-term Use of Digital Content. In *Proceedings of IS&T Conference Archiving 2012* (Copenhagen, Denmark, June 12 - 15, 2012). 200-203. Online ISSN: 2168-3204. URL= http://www.ingentaconnect.com/content/ist/ac/2012/00002012/00000001/art00044.

[4] Jackson, A. N. 2012. Tweet. URL= https://twitter.com/anjacks0n/status/167279401057255425.

[5] Spolsky, J. 2008. Why are the Microsoft Office file formats so complicated? (And some workarounds). *Joel On Software Blog* (2000 – 2014). URL= http://www.joelonsoftware.com/items/2008/02/19.html.

[6] Rothenberg, J. 1995. Ensuring the Longevity of Digital Documents. *Scientific American*. 272, 1 (1995). ISSN: 0036-8733.

[7] Rosenthal, D. 2012. Formats through time. *DSHR Blog* (2007 – 2014). URL= http://blog.dshr.org/2012/10/formats-through-time.html.

[8] Rusbridge, C. 2012. The PowerPoint 4.0 adventure: what did I learn? *Unsustainable Ideas Blog (2011 – 2013).* URL= http://unsustainableideas.wordpress.com/2012/10/15/ppt-4-adventure-learning/.

[9] De Vorsey, K., and McKinney, P. 2010. Digital Preservation in Capable Hands. *Information Standards Quarterly.* 22, 2 (Spring 2010). ISSN 1041-0031. URL= http://www.niso.org/apps/group_public/download.php/4242/IP_DeVorsey_McKinney__Risk_Assessment_isqv22no2.pdf.

[10] Testbed Digitale Bewaring. 2010. *Migration: Context and Current Status*. White Paper. Digital Preservation Testbed project. (The Hague, December 5 2001). URL= http://en.nationaalarchief.nl/kennisbank/migration-context-and-current-status-2001.

[11] Cochrane, E. 2012. *Rendering Matters*. Technical Report. Archives New Zealand (Wellington, January 2012). URL= http://archives.govt.nz/rendering-matters-report-results-research-digital-object-rendering.

[12] Graf, R., and Gordea, S. 2013. A Risk Analysis of File Formats for Preservation Planning. In *Proceedings of the iPres 2013 Conference (*Lisbon, Portugal, September 2 – 6, 2013). PURL= http://purl.pt/24107.

[13] National Digital Stewardship Alliance (NDSA). 2013. *National Agenda for Digital Stewardship*. Report. NDSA (July 2013). URL= http://www.digitalpreservation.gov/ndsa/documents/2014NationalAgenda.pdf.

[14] Nilsson, L. 2014. File Format Action Plans in Theory & Practice. *The Signal Blog* from the Library of Congress (2011 – 2014). URL= http://blogs.loc.gov/digitalpreservation/2014/01/file-format-action-plans-in-theory-and-practice/.

[15] Florida Virtual Campus. 2012. *FCLA File Format Assessments*. Report from State University Library Services, Florida Virtual Campus. URL= http://fclaweb.fcla.edu/node/795.

[16] Library of Congress. *Sustainability of Digital Formats: Planning for Library of Congress Collections*. Digital Formats website. URL= http://www.digitalpreservation.gov/formats/sustain/sustain.shtml.

[17] Pearson, D., and Webb, C. 2008. Defining File Format Obsolescence: A Risky Journey. *International Journal of Digital Curation*. 3, 1 (July 2008). ISSN: 1746-8256. DOI = http://dx.doi.org/10.2218/ijdc.v3i1.44.

[18] Archivematica. 2014. Format Policies. *Archivematica wiki*. URL= https://www.archivematica.org/wiki/Format_policies.

[19] MIT Libraries. 2014. Data Management and Publishing. *MIT Libraries website*. URL= https://libraries.mit.edu/guides/subjects/data-management/formats.html.

[20] Van der Knijff, J. 2013. Assessing file format risks: searching for Bigfoot? *Open Planets Foundation Blog* (2010 - 2014). URL= http://www.openplanetsfoundation.org/blogs/2013-09-30-assessing-file-format-risks-searching-bigfoot.

[21] Leibniz Information Centre for Economics. 2013. Hunger for Automation – The first migration actions in our Rosetta Digital Archive (poster). *International Digital Curation Conference 2013* (Amsterdam, The Netherlands, January 14 – 17, 2013). URL= http://www.dcc.ac.uk/sites/default/files/documents/idcc13posters/Poster213.pdf.

[22] Rog, J. and van Wijk, C. 2008. *Evaluating File Formats for Long-term Preservation*. Technical Report. Koninklijke Bibliotheek, (Den Haag, The Netherlands, 2008). URL= http://www.kb.nl/sites/default/files/docs/KB_file_format_evaluation_method_27022008.pdf.

[23] Rimkus, K. et al. 2014. Digital Preservation File Format Policies of ARL Member Libraries: An Analysis. *D-Lib Magazine*, 20, 3⁄4, (March/April 2014). DOI= http://dx.doi.org/10.1045/march2014-rimkus.

[24] Todd, M. 2009. *File Formats for Preservation: A DPC Technology Watch*. Technical Report. Digital Preservation Coalition (London, October 2009). URL= http://www.dpconline.org/component/docman/doc_download/375-file-formats-for-preservation.

[25] Brown, A. 2008. *Selecting File Formats for Preservation*. Technical Report. The National Archives (London, August 2008). URL= http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf.

[26] Van der Knijff, J. 2010. Ensuring the suitability of JPEG 2000 for Preservation. *Wellcome Library Blog* (2010 - 2011). URL= http://jpeg2000wellcomelibrary.blogspot.co.uk/2010/12/guest-post-ensuring-suitability-of-jpeg.html.

[27] Open Planets Foundation. 2013. Lack of Performant Open Source Decoding Software for JP2. *Open Planets Wiki* (2010 – 2014). URL= http://wiki.opf-labs.org/display/TR/Lack+of+performant+open+source+decoding+software.

[28] Open Planets Foundation and Koninklijke Bibliotheek. 2014. Jpylyzer. *JP2 validator and feature extractor website*. URL= http://openplanets.github.io/jpylyzer/.

[29] Röthlisberger, C. 2014. Kost-Val entry in the COPTR directory. *Community-Owned digital Preservation Tool Registry (COPTR)*. URL= http://coptr.digipres.org/KOST-Val.

[30] Van der Knijff, J. 2013. ICC profiles and resolution in JP2: update on 2011 D-Lib paper. *Open Planets Foundation Blog* (2010 – 2014).URL= http://www.openplanetsfoundation.org/blogs/2013-07-01-icc-profiles-and-resolution-jp2-update-2011-d-lib-paper.