IJDC | *Peer-Reviewed Paper*

The Use of PDF/A in Digital Archives: A Case Study from Archaeology

Tim N. L. Evans
Digital Archivist
Archaeology Data Service

Ray H. Moore
Digital Archivist
Archaeology Data Service

Abstract

In recent years the Portable Document Format (PDF) has become a ubiquitous format in the exchange of documents; in 2005 the PDF/A profile was defined in order to meet long term accessibility needs, and has accordingly come to be regarded as a long-term archiving strategy for PDF files. In the field of archaeology, a growing number of PDF files – containing the detailed results of fieldwork and research – are beginning to be deposited with digital archives such as the Archaeology Data Service (ADS). In the ADS' experience, the use of PDF/A has had benefits as well as drawbacks: the majority of PDF reports are now in a standard format better suited to longer-term access, however migrating to PDF/A and managing and ensuring reuse of these files is intensive, and fraught with potential pitfalls. Of these, perhaps the most serious has been an unreliability in PDF/A conformance by the wide range of tools and software now available. There are also practical and more theoretical implications for reuse which, as our discipline of archaeology alongside so many others rapidly becomes digitized, presents us with a large corpus of 'data' that is human readable, but may not be amenable to machine-based technologies such as NLP. It may be argued that these factors effectively undermine some of the perceived cost benefit of moving from paper to digital, as well as the longer-term sustainability of PDF/A within digital archives.

Received 17 July 2013 | *Revision received* 30 September 2014 | *Accepted* 1 October 2014

Correspondence should be addressed to Tim Evans, Archaeology Data Service, Department of Archaeology, The King's Manor, University of York. Email: tim.evans@york.ac.uk

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v9i2.267

Introduction

As the shift from paper-based to digital methods of distributing textual information gains momentum (Maron et al., 2013; Padova, 2011, Chapter 16; Shafait et al., 2012), the PDF has arguably become the *de facto* format for the exchange and sharing of documents. This is especially true for archaeology, a discipline which – across the globe – produces digital documents from fieldwork and research in ever-increasing numbers (Gibbs and Colley, 2012; Moore and Evans, 2013). These reports, often colloquially known as 'grey literature', are often the main archival output of an archaeological investigation; combining descriptive text with rich digital content, such as raster and vector images as supporting documentation, and are arguably *the* key component to understanding the results of the fieldwork, as well as the supporting archive (Moore and Evans, 2013). The appeal of the PDF format for archaeological sector is that it can combine the text, image and tabular elements of a typical report, along with the structural information that defines the appearance of the document, in a single file. Furthermore, this file consistently maintains the visual appearance of the original document across different platforms and can be created and viewed in a range of freely available software by the large number of stakeholders and researchers that may wish to use it. Accordingly, as the quantity of PDF files has increased it has become imperative that stable versions are created that preserve the significant properties of the reports, namely formatting, such as pagination, searching and extraction of text and quality of images, and provide access to these in the long term. Archaeology is of course not alone in this regard, and accordingly librarians and archivists have recognised the need for the development of a standardised archival version of the PDF (Bernardinello, 2010; Fanning, 2008). The resultant ISO standard (ISO, 2005),

> 'is based on Adobe's PDF Reference Version 1.4 and specifies how to use a subset of PDF components to develop software that creates, renders and otherwise processes a flavour [sic] of PDF that is more suitable for archival preservation than traditional PDF. PDF/A-1 aims to preserve the static visual appearance of electronic documents over time and also aims to support future access and future migration needs by providing frameworks for embedding metadata about electronic documents, and defining the logical structure and semantic properties of electronic documents.' (Sullivan, 2006).

The PDF/A-1 profile allows for two levels of compliance: PDF/A-1a conformance preserves the semantic and structural content of the document allowing searching and reuse, whilst the more minimal PDF/A-1b maintains the visual appearance and text searching/extraction capabilities. However, it is interesting to note that the ISO document itself acknowledged that the PDF/A 'should be used as *one* component of an organization's electronic archival environment' (ISO, 2005, our emphasis), while the Association for Information and Image Management (AIIM) has similarly cautioned that PDF/A-1 alone does not guarantee preservation (AIIM, 2006). Furthermore, as one of the standard development team reported:

> 'Our intent was *not* to claim that PDF-based solutions are the best way to preserve electronic documents. We simply defined PDF/A-1 as an archival profile of PDF that is more amenable to long-term preservation than traditional PDF.' (Sullivan, 2006, our emphasis).

In recent years, the release of PDF/A-2 (ISO, 2011), based around the PDF 1.7 specification, has extended the capabilities of PDF/A by supporting facets missing from

the original PDF/A profile and addressed many of the shortcomings of the original PDF/A standard (ISO, 2011; Drümmer, 2010). The development of the PDF/A-3 standard has broadened the scope further by allowing source formats (XML, DOCX, CSV, CAD, images and others) to be embedded as attachments to document along with any data they contain (ISO, 2012).

Clearly, and with the stated caveats, the family of PDF/A profiles offer the archivist a potential solution to the growing numbers of documents deposited in digital archives (Library of Congress, 2011) and are undoubtedly important tools to be incorporated within the working practices of such organizations. In order to feed into the wider discourse of PDF/A and digital preservation the following paper reports on the various technical, management and thematic issues encountered over the last five years in the use of the PDF/A-1 profile as a preservation and dissemination format for reports deposited in PDF (i.e. non-PDF/A) format with the Archaeology Data Service (ADS), a digital archive for primarily UK based archaeological research and commercially funded fieldwork.

PDF and the ADS

A brief introduction to the history of the ADS and the archiving of fieldwork reports is covered elsewhere, suffice to say that since 2008 and the advent of large-scale and semi-automated deposition of digitized and born-digital fieldwork reports from the OASIS system in England and Scotland (Moore and Evans, 2013), the organisation receives on average 4,000 digital documents for archive every year. At an early stage, it became evident that the majority of these documents were being supplied in PDF (i.e. non-PDF/A) format only. Despite requests for non-PDF versions (e.g. DOC or ODT), the ADS has, with some exceptions where specific agreements have been reached with Local Authority museums, no legal or financial remit to police the formats being submitted. Furthermore, upon consultation with depositors it rapidly became clear that organisational workflows were geared towards the creation of a PDF as an end product, often using a variety of component software including, but not limited to, packages such as Microsoft Office and OpenOffice, illustrating software such as Adobe Illustrator and CorelDraw, and CAD and GIS such as ESRI ArcGIS and AutoCAD. The final PDF report pulls all these parts together, and allows stakeholders, such as the creator, funder, consultant and Local Government Authority, to open and view the file without the need to purchase or install any of the aforementioned software. As detailed elsewhere, funding in UK archaeology, especially for Local Authorities has been severely impacted by the global economic crisis (ALGAO, 2013); thus a suite of proprietary software is often simply unavailable and/or the capability or knowledge of Open Source standards is not widespread. The PDF is thus the ideal vehicle for a professional looking report that can be read by anyone, on any platform, and with freely and most importantly readily available software.

Given the large-scale deposition of PDF reports and the – at the time – recent establishment of the PDF/A-1 profile as a long-term preservation solution, in 2009 the decision was taken to use the PDF/A-1 profile for the preservation version of the file stored within the Archival Information Package (AIP) created by the ADS[1]. Prior to this, the pragmatic decision was taken that PDF-only reports were either not accepted (i.e. the ADS received no file), or files had been batch converted to tiff format. The latter had always been used as an option of last resort, effectively preserving the 'look'

1 See http://archaeologydataservice.ac.uk/attach/preservation/ADS_Repository_Operations_V2.pdf for ADS Repository Operations.

 doi:10.2218/ijdc.v9i2.267

of a document but not the significant semantic properties. Although a third option was simply to do nothing, not only did this not fit the ADS implementation of the OAIS model, but contemporary reports highlighted serious long-term preservation risks of PDF files (van der Kniff, 2009). It is pertinent to note that the choice of PDF/A-1 – a format designed to provide a representation of a document but not necessarily for subsequent migration – was at odds with the ADS' traditional migration-based preservation strategy, focused on the practice of moving data to software-independent formats (normalisation) and subsequently migrating that data through successive technical infrastructures over time (refreshment). Thus, for example, non-PDF reports have historically been migrated to OpenDocument Text or Microsoft Office Open. However, it was hoped that by at least migrating all PDFs received for archive to the same standard that any use or conversion, even if this was not a batch-level migration to another standard, would at least be consistent.

It should be noted that since the adoption of PDF/A-1 by the ADS the profile has been extended to PDF/A-2 and PDF/A-3. Whilst acknowledging these developments, the ADS' working practice is to convert files to a single standard/version (for example all CAD files are currently migrated to DXF 2010 AC1024). Although a move to incorporate later profiles within practice is anticipated, any subsequent move to a later profile of PDF/A would have to be evaluated in light of potential benefits for reuse and preservation (cf. Arms et al., 2014; Oettler, 2013), as well as logistic considerations in software purchases.

Preservation: Conversion and Identification

Within its working practices, the ADS initially started using Adobe Acrobat 8 and subsequently Acrobat Pro 9 and 10 on single PDF to PDF/A-1 conversions. As has been noted elsewhere (Noonan et al., 2012; Walker et al., 2007), it has been extremely difficult to convert PDF to the PDF/A-1a standard, and thus conformance level 1b has become the default output. However, the single file migrations within Adobe software have required serious manual intervention to ensure a successful outcome, and in some cases creating a valid PDF/A-1 file has not proved possible. The most common problems encountered have been:

- XMP property for a page object not predefined and no extension schema present;
- Fonts have glyphs with zero character widths, reported by Acrobat as 'width information incomplete';
- Fonts have not been embedded.

Interestingly, other authors have identified these and other technical considerations when creating PDF/A files; such as issues with structural tagging, embedding fonts and incorrect sub-setting for TrueType (Drümmer, 2007). Drümmer also notes that whilst a PDF/A file can be formally correct it can still have incorrect glyphs and that only a careful visual check can uncover this problem. This issue has also been noted within PDF/A files created by the ADS, often where fonts have been used within pages inserted (or simply copied and pasted) from files created by packages such as Illustrator or CorelDraw. Although these can often be rectified, the identification of these errors in documents that can span hundreds of pages is often an intensive and time-consuming activity.

Since mid-2011 the ADS have also utilised PDF/A Manager[2] created by PDFTron to facilitate batch-level processing; although this has led to a reduction in processing time and the automation of previously manual fix-ups (such as the editing/stripping of XMP properties), the migrations are still not without problem cases. In the ADS experience, the success rate – i.e. the PDFTron conversion script returning a successful post-conversion validation – is on average around 80% of all files processed, a figure that bears a close similarity to a recent evaluation of several types of conversion software (Koo and Chou, 2013). However, in subsequent checks of the conformance level of these 'successful' files – created in PDFTron and apparently valid PDF/A 1b – using Adobe Preflight, between 5-10% have failed to conform to the profile. Aside from small issues with XMP properties, the most common inconsistencies in PDFTron-converted files are errors generated by rendering of glyphs and subsets of embedded fonts.

This inconsistency in cross-software conformance has prompted ADS staff to revisit archives with PDF/A-1b files created in Acrobat 8 and they have discovered that significant numbers (c. 25%) of these were not validating as such in the Preflight tool for either Pro 9 or 10, and even a very small number created in Acrobat 9.5.5 that were not validating in Pro 10. On closer investigation it seems these issues have not been confined to the ADS experience. The Bavaria report (PDFlib, 2009) identifies the same key problems regarding validation, and attribute them to the fine-tuning of the PDF/A-1 standard throughout versions 7-9 of Adobe Software, as well as third party tools. The following statement perhaps best summarised the situation in 2009, and that is undoubtedly influencing some of the historic/legacy problems encountered by the ADS:

> 'This emphasis on providing more and stricter test criteria made users wonder whether PDF/A validation will ever stabilize or whether they'll be forced to constantly adjust their documents to the latest validation technology… the status of PDF/A validation is considerably better than in the early days.' (PDFlib, 2009).

If from this we may expect the PDF/A standard to have stabilized within certain software, it has been noted that this 'historic' problem is still having a knock-on effect in files being deposited with the ADS to the present day. For example, in recent years there have been a substantial number of PDF files being deposited with the ADS that initially appeared to be PDF/A-1b, but on running the Preflight compliance tool in Adobe Professional 9 or 10 did not comply with the 1b standard. Typically, these files appear to have been produced by a range of non-Adobe software such as 'Nitro PDF', 'PDFCreator' and 'Solid PDF Creator', with the most common error reports being:

- CIDset in subset font is incomplete;
- Encoding entry prohibited for symbolic TrueType font;
- Data entry in in document information uses Unicode heading.

Although more mainstream tools embedded in Adobe and Microsoft software are, in the authors experience, more reliable it has also been noted that PDF/A-1b files created in Acrobat PDFMaker for Word (v.8) in particular have subsequently failed validation in Adobe Preflight. A recent survey by the Florida Digital Archive has highlighted similar uncertainties in the validation/conformance in 3-Heights and (the aforementioned) PDF/A manager, as well as the difficulty in detecting conformance problems files that claim to be PDF/A compliant – what the authors' term *false positives* – at a batch level (Koo and Choo, 2013). The cited case study is almost identical to the ADS' experience,

2 PDF/A Manager: http://www.pdftron.com/pdfamanager/

 doi:10.2218/ijdc.v9i2.267

whereby issues on a simple file-by-file-basis the situation can be identifiable (through cross-checking) and redeemed. However, the ADS hold an archive of over 85,000 PDF files that are managed through an internal database that identifies the files using batch-level identification software. At the time of writing this is the DROID[3] toolkit. In this scenario, the accuracy of the type of PDF file the ADS holds is of paramount importance, as resources and practicality do not permit checks on all individual files using a fleet of software. In the limited checking and validation that has been undertaken, it has been noted that some files identified by DROID as PDF/A-1, do not validate as such. On closer inspection, DROID (and indeed the initial recognition of a 'PDF/A' file in Adobe software) identification is based on the <pdfaid:part> namespace declaration in the PDF/A Identification.

However, it has been noted that *some* software embeds this declaration upon an attempted conversion to PDF/A-1, and it is only upon checking the file against an Adobe PDF/A-1 conformance tool that the error is noted. In the latter instance, the ADS is continuing to work with the DROID team, based at the National Archives[4], and other partners, such as the SPRUCE project[5], to find ways to correct these misidentifications, and as with the overall issue regarding consistent validation, it is hoped that, as the standard becomes firmly established, the amount of erroneous PDF/A files will decrease. Nevertheless, we would argue that the situation for those attempting to create or identify compliant files is still not straightforward. In recent years there has been a proliferation of different sites and services (often freeware) offering to create PDF/A-1 compliant files. In our field – archaeology – commercial organizations without access to the latest Adobe software will often look, in good faith, towards a free alternative, which in turn can lead to these highlighted inconsistencies within the files being produced. Thus digital archives, such as the ADS, could be left dealing with a significant legacy issue as the PDF/A-1 standard is firmly established within the archaeological community.

Dissemination: Use and Reuse

As detailed above, the PDF format is the most common type of file deposited with the ADS, typically as reports but increasingly as a wrapper for other data types, including raster and vector images. Indeed, part of the ADS' ongoing roles is to attempt to educate the community about best practice for digital data through joint initiatives, such as the Guides to Good Practice[6]. However, and perhaps due to the aforementioned limitations in software and technical skills available to large sections of the community, there is not only a prevalence for reports and data to be deposited in PDF, but also a *requirement* by large parts of the community for reports to be disseminated in this format. Thus despite the advantages offered by dissemination routes such as XML (Falkingham, 2005), the overarching trend both within the UK and further abroad for is for a paginated (i.e. citeable) PDF report online[7]. Of course, outside archaeology, the limitation of PDF-only access to data-rich reports has been reported elsewhere:

3 DROID (Digital Record and Object Identification): http://digital-preservation.github.com/droid/
4 National Archives: http://www.nationalarchives.gov.uk/default.htm
5 SPRUCE: http://wiki.opf-labs.org/display/SPR/PDFA+Validation+tools+give+different+results
6 Guides to Good Practice: http://guides.archaeologydataservice.ac.uk/
7 For examples of non-ADS archaeological reports see: http://www.worcestershire.gov.uk/cms/archive-and-archaeology/search-our-records/online-archaeology-library.aspx or http://nswaol.library.usyd.edu.au/index.jsp;jsessionid=B298CAFE616AF5A8284B802EC1C9CBF1?page=home

> 'The status quo in development reporting practices is built on the foundation of the PDF report. This is understandable. There are often numerous different documents used to make a single project report, including excel models, GIS shapefiles, and Photoshop charts. The ease of taking screenshots and putting it all into a PDF report, and sending it along via email is completely understandable. But this is like funding James Cameron to make Avatar, and then releasing it in a black and white flipbook. We are missing all the good stuff.' (Manning, 2013).

To readers of this paper this may seem something of a truism. However, the key points of digital preservation are still unheeded within parts of the UK archaeological community, and, as the above quote suggests, other sectors in general. Indeed, defining the split *within the archive* between the 'report' and the 'data' is still very much mixed within the archaeological sector[8].

Despite these caveats, the ADS have begun to use PDF/A for dissemination of all reports by the ADS originally deposited as PDF. As stated, this is a reaction to the requirement to maintain the text element that is not possible in other formats such as TIFF, as well as keeping the formatting, especially pagination for citation purposes, required by users but not retained in plain text or XML. However, it has been noted by the authors that conversion of PDF to PDF/A does have some practical knock on effects on the final output. For example, in a small number of cases there is a lack of available encoding information, either in the PDF/A, or in the embedded font data, to derive the meaning of the characters/glyphs that are displayed on the pages in the document. As these reports are often composites of multiple source files, this error is often restricted to small sections of the report, resulting in a document that initially looks useable, but is in fact restricted in certain places.

These issues are, with careful checking, often fixable but have an impact on the time it takes for staff to prepare a dissemination ready file. However, whilst undoubtedly serving a large proportion of our community's needs – namely downloading a referencing a report in their own research – some recent work on text mining has illustrated some problematic issues with using PDF (and PDF/A) for more advanced uses. For example, in a recent attempt to index archaeological 'grey literature' based on natural language processing (NLP) techniques, the authors found that PDF files within the dataset did not contain uniform characteristics such as spacing or line breaks, which, in turn, had ramifications for the application of the NLP (Vlachidis et al., 2010). Another machine-based survey of PDF chemistry e-theses encountered other problems, such as lack of a tagging structure in non-PDF/A files that made them unsuitable for text-mining purposes (Downing et al., 2010). The authors of this study did address the role of PDF/A as providing a solution via tagging and Unicode mapping, but – and as has been highlighted in this article – this is not always successfully implemented in PDF to PDF/A conversions (ibid). The authors then conclude that:

> 'any continuous text could still be broken by images, tables, or embedded objects (crucial components of chemistry e-theses). Although it is likely that PDF/A will in the future be considered by many institutions as the primary document archival format, it has been noted that its use is not a straightforward automatic process.' (Downing et al., 2010).

8 For example, compare a complete digital archive from archaeological fieldwork (http://dx.doi.org/10.5284/1000180) and the results of a completely scanned PDF archive (https://library.thehumanjourney.net/665/).

 doi:10.2218/ijdc.v9i2.267

Another issue to consider is that of personal/organizational ownership of data within PDF files. In recent years there has been a growing awareness of the commercial, or indeed academic, value of data. In this respect, the PDF medium offers an option of control over the dissemination of information. For example, files can be password secured, or images can be flattened and reduced in quality. In this way the information is arguably available but perhaps no more reusable than a traditional paper version. There is an irony that just as grey literature – and access to grey literature – is becoming acknowledged as important by the archaeological community (Ford, 2010; Seymour, 2010), its longevity and reuse potential is perhaps being undermined by the same file media that brought it to prominence. However, the digital archive is in a Catch-22 situation: if they accept and disseminate PDF then elements of the community (including the archivists themselves) will point out the restrictions in use of this format; if they refuse to accept PDF, or other parts of the community cannot or will not engage in non-PDF methods of reporting, then the archive often does not receive them at all.

Impact and Cost Benefit

Assessing the value and impact of digital data and repositories has become increasingly important in a climate of recession and reduced economic activity. Writing at the outset of the economic crisis it was observed that:

> 'sustainable economics for digital preservation is not just about finding more funds. It is about building an economic activity firmly rooted in a compelling value proposition, clear incentives to act, and well defined preservation roles and responsibilities.' (Blue Ribbon Task Force, 2010).

As a profession we have made significant advances in assessing the true financial cost of the digital preservation lifecycle, even if it can still seem like something of a 'dark art' (ADS, 2012; Lavoie, 2008; Wheatley, 2008). Of course, measuring the actual value of digital resources is notoriously difficult and complex. Indeed, Meyer observes that "[n]o single measure reflects 'the impact' of a digital resource" (2011). Tanner and Deegan have identified five "modes of value" associated with digital resources: the possibility of enjoying resources (option), utility from knowing a resource is cherished (prestige), the contribution it makes to education (education), the benefit from knowing a resource has been preserved (existence), and knowledge that resources are preserved for future generations (bequest) (2011). It is these final two 'modes of value' that should concern us here, but it is worth remembering that assessing such impact remains woefully under researched. The recent work carried out by the ADS, for example, in association with Charles Beagrie Ltd and the Centre for Strategic Economic Studies at Victoria University, stands as an exemplar (ADS, 2013; see also Beagrie, Lavoie and Woollard, 2010).

The issue of preservation is certainly a complex one in which a seemingly simple decision such as file format can have serious ramifications for preservation and reuse. Many will well remember the zeal for digitisation from the mid-nineties onwards, motivated by the perceived substantial cost savings that digital storage offered over its physical equivalent (Fanning, 2008). Whilst organisations increasingly recognised the potential cost savings and value that could be added to resources through the enhanced access that digital resources provided, file format was considered to be inconsequential. Within this milieu, PDF was widely regarded as the "*de facto* standard for distributing and exchanging documents in an unchangeable manner", providing a stable, cross platform format that allowed the exchange of page-orientated textual data and was

widely adopted as an acceptable format (Chapman et al., 2010; Fanning 2008). For many, the PDF's ability to preserve the layout of the original made it a particularly effective within many of these digitisation projects, where its ability to provide a single solution for both preservation and dissemination was much lauded. The simplicity and cross-platform support for the PDF format made it popular and it quickly became ingrained within workflows and this existing support for PDF made progression to a preservation standard relatively straightforward for many organisations and repositories. Certainly an ability to transform the wide variety of data formats employed in the workflows of organisations into a single, ISO-recognised preservation format offered significant advantages in terms of data management, negating conversion and/or migration, and perhaps most significantly, was economically sustainable (Drümmer, 2007). As a recent survey has reported:

> 'Digital assets can be extremely fragile and ephemeral, and the need to make preservation decisions can arise as early as the time of the assets creation, particularly since *studies to date indicate that the total cost of preserving materials can be reduced by steps taken early in the life of the asset.*' (Blue Ribbon Task Force, 2008, our emphasis).

The "widespread adoption of PDF/A-1" as a preservation format can also be "closely linked to the availability of software" that supports it (Hodge and Anderson, 2007), perhaps leading some to consider it a simple and inexpensive solution to the archiving of files. At the same time, within an environment where dispute exists over appropriate alternative formats for the long-term preservation of textual and other data types, it is understandable how the resultant confusion has led many to rely on PDF/A (Park and Oh, 2012). Yet as data creators and repositories increasingly employ PDF/A as their preservation solution, fundamental decisions are made over whether data is held or archived in a way that preserves access to content, or whether it is conserved in a manner that facilitates reuse (Hodge and Anderson, 2007; Park and Oh, 2012). Such conceptual debates are difficult to reconcile in an environment where a 'traditional' notion of the role of the archive is still pervasive, and where an emphasis is placed on the physical appearance of a document (Morgan et al., 2008; Holmes and Romary, 2011). Yet, as Ross observes,

> 'reuse over time of digital materials *will produce opportunities for the growth of creative and knowledge economies*. We know that, as the transition from in vitro to in silico science gathers pace, the longer-term viability of this new scientific paradigm requires that we curate digital materials in ways that ensure their reusability.' (Ross, 2012).

A concern for reuse is certainly warranted in an age of economic austerity where the generation of digital data and the digitisation of physical data has become so costly, but where increasing concerns have highlighted the importance of an ability to present and reconstitute constituent data in forms and formats that are different from the original. Attempts have been made to assess the reusability of the PDF/A, but despite its popularity as a preservation format such studies are rare and often somewhat rudimentary (e.g. Lundell and Lings, 2010). More significant to this debate is the fact that PDF is intended to be read by the human eye, preserving the layout of the original document, with no or limited conceptual or hierarchical structure, making programmatic or automated analyses difficult (Morgan et al., 2008; Reggi, 2011). For some within the preservation community the reuse issue is more conceptually than technically problematic, particularly when issues of ownership and copyright are considered (Ure and Hanley, 2011). We would argue that these apprehensions should be considered as

 doi:10.2218/ijdc.v9i2.267

data management issues rather than being a specific preservation problem. Whilst PDF/A does provide significant advantages in terms of preservation, any notion of reuse is "fundamentally flawed because of the way they 'freeze' the digital content" (Cayless, 2010). Whilst we may hope future developments will provide us with the tools to extract digital content from a PDF/A, or at least lead the development of workflows and tools which will allow us to convert them to alternative formats, we should not assume that this will be the case.

Conclusions

As PDF has become widely used within contemporary academic and commercial workflows, institutional repositories and data archives have been required to respond to the increasing delivery of digital content in the format. A more progressive approach from the digital archiving sector, one which has involved the wider community of software manufacturers and developers, has seen the development and implementation of PDF/A as a more technically dynamic response to the issue of digital preservation of the PDF format. In the author's experience within the ADS, PDF/A-1 has indeed provided a robust if not perfect solution to archiving PDF. While we may ask ourselves if this is necessarily the *best* response in terms of limitations in future migration, issues with conversions and identification and the limitations of reuse, it is pragmatically the only current response we can make. At the time of writing the key concern is ensuring that the PDFs accessioned and migrated by the ADS conform to the PDF/A-1 standard, and that we are not mistakenly creating a backlog of documents that retrospectively fail to meet this profile. Indeed, it is anticipated that going forward our organization will move towards the later versions of the standard such as PDF/A-3, which address issues encountered with the earlier format and aspects of PDF which have become available since PDF1.4 (Library of Congress, 2012). The decision to do so will be informed not only by an assessment of reliability and adoption by the rest of the digital preservation community, but also an evaluation of the technical ramifications of any embedded files: whether these objects will be simply viewable, or whether they can be extracted, edited and effectively preserved and reused.

In the authors' opinion, leaving aside 'preservation' issues, the facility for reuse is fundamental. As highlighted above, the reuse limitations of PDF/A-1 are evident; that is any PDF/A-1 file is designed for 'human consumption' such as reading, printing and copying of text and graphics. This is not a direct criticism of the standard itself, the 'readability' and freedom from reliance on libraries of fonts is, after all, the very essence of the 1-A standard. However, it is a point that needs to be re-enforced by practical experience, notably the difficulties in using 'text-based' reports for machine-based language processing and indexing. As the amount of reports online grows, and with the advance of cross-organisational sharing of documents and metadata the demand for more complex reuse may be nascent, but arguably will develop in the future as users begin to ask deeper questions of the corpus of documents available. For example, at the time of writing the use of UK 'grey literature' reports by large research projects is an emerging theme: however, what is noticeable is the time and effort it takes for researchers to physically process all of these documents. As the trial NLP project cited in this report shows, we are not far off a situation where a researcher may begin to ask if there are new ways of, for example, extracting the measurements or characteristics of a particular archaeological object or monument from a corpus of 20,000 reports without

having to download and read every single file. In being asked to facilitate complex reuse scenarios, such as NLP, the digital archive has to begin to think not only of simple preservation issues, but also how these files can be used again going into the short and longer-term future. Within this environment can a policy that relies solely on the visual appearance of digital data really be considered the most appropriate solution? Perhaps the future challenges are not just in ensuring that the PDF/A standard is used consistently and accurately, but that other avenues are explored to enable the information within files is not just limited to the human eye.

This issue also highlights the inherent weaknesses of extant reactive policies within some digital archiving practices that have a focus on issues of preservation, such as file format, with only incidental consideration of reuse. The OAIS reference model for digital archives (Allinson, 2006) certainly recommends equal importance be placed on both data preservation *and* reuse; unfortunately the difficulties of the former dominate, often at the expense of the latter. Arguably the most significant success of the development of PDF/A has not been the format *per se*, but rather that the approach advocated during its evolution, which illustrates the power of streams of dialogue between archiving specialists and software creators for emerging formats. This has enormous potential of this relationship to facilitate preservation strategies at the grassroots level. Unfortunately, in the case of PDF this comes too late in the process of software development, serving only to perpetuate a reactive policy to existing formats. Instead we would advocate a more proactive response where such engagement occurs during the development phase of new software and data types. With this more holistic approach, where planning for preservation *and* reuse occurs during the development phase of new formats and software, the issue of preservation should become less onerous and problematic.

Acknowledgements

The authors would like to thank colleagues at the Archaeology Data Service for informal contributions to the wider discussion of PDF/A, particularly Jenny O'Brien for feedback on conversion and validation errors, and to Catherine Hardman and Julian Richards for their comments on an earlier version of this paper. In addition, the thorough critique of the anonymous reviewers on an earlier version of this paper was instrumental in the final text. Any mistakes and opinions reside with the authors.

References

Abrams, S., Fanning, B., Helander, D., & Sullivan, S. (2005). PDF/A. The development of a digital preservation standard. Paper presented at the 69th Annual Meeting of the Society of American Archivists, New Orleans. Retrieved from http://www.aiim.org/documents/standards/PDFA69thSAA805.pdf

ADS. (2012). The dark art of costing for digital preservation [Web log post]. Retrieved from Archaeology Data Service SWORD-ARM web log: http://archaeologydataservice.ac.uk/blog/sword-arm/2012/03/the-dark-art-of-costing-for-digital-preservation/

ADS. (2013). Impact of the Archaeology Data Service: A study and methods for enhancing sustainability. Retrieved from http://archaeologydataservice.ac.uk/research/impact

AIIM. (2006). *Frequently asked questions (FAQs): ISO 19005-1:2005 (PDF/A-1).* Retrieved from http://www.aiim.org/documents/standards/19005-1_FAQ.PDF

ALGAO. (2013). *A fifth report on Local Authority Staff Resources.* Retrieved from http://www.ihbc.org.uk/skills/resources/5th-rep-LAStaff.pdf

Allinson, J. (2006). *OAIS as a reference model for repositories: An evaluation*. Retrieved from UKOLN website: http://www.ukoln.ac.uk/repositories/publications/oais-evaluation-200607/Drs-OAIS-evaluation-0.5.pdf

Arms, C., Chalfont, D., DeVorsey, K., Dietrich, C., Fleischhauer, C., Lazorchak, B., … Murray, K. (2014). *The benefits and risks of the PDF/A-3 file format for archival institutions.* Retrieved from Library of Congress, National Digital Stewardship Alliance website: http://hdl.loc.gov/loc.gdc/lcpub.2013655115.1

Beagrie, N., Lavoie, B., & Woollard, M. (2010). *Keeping research data safe 2*. Salisbury, UK: Charles Beagrie. Retrieved from http://repository.essex.ac.uk/2147/1/keepingresearchdatasafe2.pdf

Bernardinello, R. (2010). PDF/A 101 – Introduction to PDF/A. In *PDF/A Forever: Long-Term Archiving with PDF. 4th International PDF/A Conference* (pp. 22–23). Berlin, Germany: Association for Digital Document Standards. Retrieved from http://www.pdfa.org/wp-content/uploads/2011/08/PDFA-forever_1b.pdf

Blue Ribbon Task Force. (2008). *Sustaining the digital investment: Issues and challenges of economically sustainable digital preservation*. Retrieved from http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf

Blue Ribbon Task Force. (2010). *Sustainable economics for a digital planet: ensuring long term access to digital information.* Retrieved from http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Cayless, H.A. (2010). Ktêma es aiei: Digital permanence from an ancient perspective. In G. Bodard & S. Mahony (Eds.), *Digital Research in the Study of Classical Antiquity,* (pp. 139–150). Farnham: Ashgate.

Chapman, A.D., Turland, N.J., & Watson, M. (2010). Report of the Special Committee on Electronic Publication. *Taxon 59*(6), 1853–1862.

CCSDS. (2012). *Reference model for an Open Archival Information System (OAIS)* (Magenta Book No. CCSDS 650.0-M-2). Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

Digital Preservation Coalition. (2008). *Digital preservation handbook*. Retrieved from http://www.dpconline.org/advice/preservationhandbook

Downing, J., Harvey, M.J., Morgan, P.B., Murray-Rust, P., Rzepa, H.S., Stewart, D.C., … Townsend, J.A. (2010). SPECTRa-T: Machine-based data extraction and semantic searching of chemistry e-theses. *Journal of Chemical Information and Modeling 50*(2), 251–261. doi:10.1021/ci9003688

Drümmer, O. (2007). *PDF/A – A look at the technical side*. Retrieved from the PDF Association website: http://www.pdfa.org/wp-content/uploads/2011/08/pdf-a_a_look_at_the_technical_side-2b.pdf

Drümmer, O. (2010) PDF/A-2 – Technical overview. In *PDF/A Forever: Long-Term Archiving with PDF. 4th International PDF/A Conference* (pp. 12–16). Berlin, Germany: Association for Digital Document Standards. Retrieved from http://www.pdfa.org/wp-content/uploads/2011/08/PDFA-forever_1b.pdf

Dryden, J. (2008). PDF/A-1: A ray of light in the digital dark age? *Journal of Archival Organization, 6*(1-2), 121-124. doi:10.1080/15332740802246841

Falkingham, G. (2005). A whiter shade of grey: A new approach to archaeological grey literature using the XML version of the TEI Guidelines. *Internet Archaeology 17*. doi:10.11141/ia.17.5

Fanning, B.A. (2008). *Preserving the data explosion: Using PDF* (Technology Watch Report No. 08-02). Retrieved from Digital Preservation Coalition website: http://www.dpconline.org/docs/reports/dpctw08-02.pdf

Ford, M. (2010). Hidden treasure. *Nature 464*, 826–827. doi:10.1038/464826a

Gibbs, M., & Colley, S. (2012). Digital preservation, online access and historical archaeology 'grey literature' from New South Wales, Australia. *Australian Archaeology, 75,* 95–103.

Hodge, G., & Anderson, N. (2007). Formats for digital preservation: A review of alternatives and issues. *Information Services and Use, 27*, 45–63. Retrieved from http://iospress.metapress.com/content/9v61373118049755/

Holmes, M., & Romary, L. (2011). Encoding models for scholarly literature: Does the TEI have a word to say? In I. Iglezakis, T.-E. Synodinou, & S. Kapidakis (Eds.), *E-publishing and Digital Libraries: Legal and Organizational Issues* (pp. 88-110). Hershey, PA: Information Science Reference. doi:10.4018/978-1-61692-834-6.ch005

ISO. (2005). *ISO 19005-1, Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1-1)*.

ISO. (2008). *ISO 32000-1:2008, Document management – Portable document format – Part 1: PDF 1.7.*

ISO. (2011). *ISO 19005-2:2011, Document management – Electronic document file format for long-term preservation – Part 2: Use of ISO 32000-1 (PDF/A-2)*.

ISO. (2012). *ISO 19005-3:2012, Document management – Electronic document file format for long-term preservation – Part 3: Use of ISO 32000-1 with support for embedded files. (PDF/A-3)*.

Koo. J., & Chou, C.C.H (2013). PDF to PDF/A: Evaluation of converter software for implementation in digital repository workflow. *New Review of Information Networking 18*(1), 1–15. doi:10.1080/13614576.2013.771989

Lavoie, B.F. (2008). The fifth blackbird: Some thoughts on economically sustainable digital preservation. *D-Lib Magazine 14*(3/4). doi:10.1045/march2008-lavoie

Library of Congress. (2011). PDF/A, PDF for long-term preservation. Retrieved from http://www.digitalpreservation.gov/formats/fdd/fdd000318.shtml

Library of Congress. (2012). PDF/A-3, PDF for long-term preservation, use of ISO 32000-1, with embedded files. Retrieved from http://www.digitalpreservation.gov/formats/fdd/fdd000360.shtml

Lundell, B., & Lings, B. (2010). How open are local government documents in Sweden? A case for open standards. In P. Ågerfalk, C. Boldyreff, J.M. González-Barahona, G.R. Madey, & J. Noll (Eds.), *IFIP Advances in Information and Communication Technology: Vol. 319. Open source software: New horizons* (pp. 177-187). Berlin, Germany: Springer. doi:10.1007/978-3-642-13244-5_14

Manning, N. (2013, October 21). Bad metrics and PDF graveyards: Why development needs open data. *The Guardian Global Development Professionals Network*. Retrieved from http://www.theguardian.com/global-development-professionals-network/2013/oct/21/development-open-data-action

Maron, N.L., Yun, J., & Pickle S. (2013). *Sustaining our digital future: Institutional strategies for digital content*. Retrieved from Strategic Content Alliance website: http://sca.jiscinvolve.org/wp/files/2013/01/Sustaining-our-digital-future-FINAL-31.pdf

Meyer, E.T. (2011). *Splashes and ripples: Synthesizing the evidence on the impact of digital resources*. JISC: London. Retrieved from http://ssrn.com/abstract=1846535

Moore, R.H., & Evans, T.N.L. (2013). Preserving the grey literature explosion: PDF/A and the digital archive. *Information Standards Quarterly 25*(3), 20–27. doi:10.3789/isqv25no3.2013.04

Morgan, P., Downing, J., Murray-Rust, P., Stewart, D., Tonge, A., Townsend, J., … Rzepa, H. (2008, June). *Extracting and re-using research data from chemistry e-theses: The SPECTRa-T project.* Paper presented at the 11th International Symposium on Electronic Theses and Dissertations, Aberdeen, UK. Retrieved from http://www.dspace.cam.ac.uk/handle/1810/230116

Noonan, D.W., McCrory, A., & Black, E.L. (2010). PDF/A: A viable addition to the preservation toolkit. *D-Lib Magazine 16*(11/12). doi:10.1045/november2010-noonan

OASIS. (2013). *OASIS III: 21st monitoring report April 2013*. Retrieved from http://oasis.ac.uk/pages/attach/MonitoringReports_England/OASIS_III_21stMonitoringReport.doc

Oettler, A. (2013). *PDF/A in a nutshell 2.0: PDF for long-term archiving.* Berlin, Germany: Association for Digital Document Standards. Retrieved from PDF Association website: http://www.pdfa.org/publication/pdfa-in-a-nutshell-2-0/

Padova, T. (2011). *Adobe Acrobat X PDF Bible.* Indianapolis, IN: Wiley. doi:10.1002/9781118255728

Park, E.G., & Oh, S. (2012). Examining attributes of open standard file formats for long-term preservation and open access. *Information Technology and Libraries, 31*(4), 44-65. doi:10.6017/ital.v31i4.1946

PDFlib. (2009). *The Bavaria report on PDF/A validation accuracy.* Retrieved from http://www.pdflib.com/fileadmin/pdflib/pdf/pdfa/2009-05-04-Bavaria-report-on-PDFA-validation-accuracy.pdf

PDF Tools AG. (2009). *PDF/A – The standard for long-term archiving.* Retrieved from https://www.pdf-tools.com/public/downloads/whitepapers/whitepaper-pdfa.pdf

Reggi, L. (2011). Benchmarking open data availability across Europe: The case of EU structural funds. *European Journal of ePractice, 12*, 17–31. Retrieved from http://www.epractice.eu/en/document/5290093

Ross, S. (2012). Digital preservation, archival science and methodological foundations for digital libraries. *New Review of Information Networking, 17*, 43–68. doi:10.1080/13614576.2012.679446

Seymour, D.J. (2010). In the trenches around the ivory tower: Introduction to black-and-white issues about the grey literature. *Archaeologies: Journal of the World Archaeological Congress 6*(2), 226-232. doi:10.1007/s11759-010-9130-z

Shafait, F., Cutter, M.P, van Beusekom, J., Saqib Bukhari, S., & Breuel, T.M. (2012). Decapod: A flexible, low cost digitization solution for small and medium archives. In M. Iwamura & F. Shafait (Eds.), *Lecture Notes in Computer Science: Vol. 7139. Camera-based document analysis and recognition* (pp. 101–111). doi:10.1007/978-3-642-29364-1_8

Smith, A., Cairns, S., & Vokes, C. (2012) *Review of the development and implementation of OASIS in England.* Harrogate, UK: Pye-Tait Consulting. Retrieved from http://oasis.ac.uk/pages/attach/PROJECT%20HISTORY/OASIS-Review-Final-Report-270112-v2.pdf

Smith, R., & Tindall, A. (2012). *A survey of archaeological archives held by archaeological practices in England, Scotland and Wales.* Retrieved from Federation of Archaeological Managers and Employers website: http://www.famearchaeology.co.uk/2012/11/a-survey-of-archaeological-archives-held-by-uk-archaeological-practices/

Sullivan, S.J. (2006). An archival/records management perspective on PDF/A. *Records Management Journal, 16*(1), 51–56. doi:10.1108/09565690610654783

Tanner, S., & Deegan, M. (2011). *Inspiring research, inspiring scholarship: The value and benefits of digitised resources for learning, teaching, research and enjoyment*. London, UK: Higher Education Funding Council for England. Retrieved from http://www.kdcs.kcl.ac.uk/fileadmin/documents/Inspiring_Research_Inspiring_Scholarship_2011_SimonTanner.pdf

Ure, J., & Hanley, J. (2011). Curating complex, dynamic and distributed data: Telehealth as a laboratory for strategy. *International Journal of Digital Curation 6*(2), 128–145. doi:10.2218/ijdc.v6i2.207

van der Knijff, J. (2009). *Adobe Portable Document Format: Inventory of long-term preservation risks.* The Hague, the Netherlands: Koninklijke Bibliotheek. Retrieved from http://www.openplanetsfoundation.org/system/files/PDFInventoryPreservationRisks_0_2_0.pdf

Vlachidis, V., Binding, C., Tudhope, D., & May, K. (2010). Excavating grey literature: A case study on the rich indexing of archaeological documents via natural language-processing techniques and knowledge-based resources. *Aslib Proceedings 62*(4), 466-475. doi:10.1108/00012531011074708

Walker, F.L., Gallagher, M.E., & Thoma, G.R. (2007). PDF file migration to PDF/A: Technical considerations. In *Archiving 2007: Final program and proceedings* (pp. 6–11). Retrieved from http://lhncbc.nlm.nih.gov/files/archive/pub2007020.pdf

Wheatley, P. (2008). Costing the digital preservation lifecycle more effectively. In *Proceedings of The Fifth International Conference on Preservation of Digital Objects (iPRES 2008) – Joined up and working: Tools and methods for digital preservation.* (pp. 122–126). London, UK: British Library. Retrieved from http://www.bl.uk/ipres2008/ipres2008-proceedings.pdf