PDF/A – TOWARDS A TRUE DIGITAL ARCHIVAL SURROGATE (DAS) FOR DIGITAL MANUSCRIPT COLLECTIONS

Rodney Obien and Jeffrey Monseau
Wallace E. Mason Library, Keene State College, 229 Main Street, Keene, New Hampshire, U.S.A.
College Archives, Springfield College, 263 Alden Street, Springfield, Massachusetts, U.S.A.

Keywords: Digital Surrogates, Digital Libraries, Digital Preservation, Manuscript Collections, PDF/A.

Abstract: Digital surrogates provide a non-invasive means to study old manuscript documents that are often too fragile and valuable for wide public access. These surrogates are generated from web-accessible derivatives made from high-resolution archival masters; these masters serve as long-term digital preservation copies. What if there was a file format that combined the functions of digital surrogate, web-accessible derivate, and archival master? This paper considers the notion of the archival file format PDF/A (ISO: 19005-1) as digital archival surrogate or DAS that combines the functions of surrogate, derivative, and master. The paper discusses, furthermore, the versatility of PDF/A in dealing with the complex nature of old manuscripts, and the possible implications of adapting PDF/A as a DAS standard.

1 INTRODUCTION

Digital surrogacy unlocks many options for studying rare and fragile cultural artifacts. Often, surrogates provide the only feasible access to the originals. Surrogates allow for the "virtual viewing" of an artifact -- enabling detailed analysis (Nichols, 2007). They also allow cultural institutions to virtually bring together materials, "re-construct wholes from parts and make resources both visible and convenient for researchers to use" (Butler, 2009).

Nowhere are digital surrogates more valuable than in the study of old manuscripts. Manuscripts, like medieval illuminated text, musical scores, letters, and diaries, offer researchers and scholars a rich source of historical documentary material. These handwritten documents are complex, multi-page objects, bound and unbound, and often too fragile and valuable for wide public access. Digital surrogates afford a non-invasive means of providing access to these unique complex materials.

Several digitization projects have focused on creating high-quality digital surrogates for manuscript collections. The British Library (http://www.bl.uk/onlinegallery/ttp/ttpbooks.html), for example, is making its manuscript collections available through the "elegant" *Turning the Pages*™ system, developed by the Armadillo Corporation. The Internet Archive project (http://www.archive.org) is committed to digitizing manuscripts, making them accessible through document viewers like Djvu (AT&T) and PDF (Adobe).

The standard process of creating a surrogate is by first digitizing the "original" manuscript and creating archival masters, such as TIFF or JPEG2000 files, and then high-quality web-deliverable derivates, whether JPEGs, GIFFs, or PDFs, are made from the masters which then constitutes the surrogate. This process also follows the standard two-tier digital preservation model of master and derivate practiced in the library and archival professions. Creating, storing and retrieving these derivatives, though, is a complex and cumbersome process that often requires expansive resources and a high level of expertise that can be daunting to small or large institutions a like.

What if there was a way to simplify this process? What if there was a means to combine the functions of an archival master, derivative, and digital surrogate? What if there was only need for one file format. And what if this file format existed already as a digital archival standard, recognized by the International Standards Organization (ISO)? Such a file format does indeed exist in PDF/A.

The PDF/A file Standard (ISO: 19005-1) could

Obien R. and Monseau J.
PDF/A âĂŞ TOWARDS A TRUE DIGITAL ARCHIVAL SURROGATE (DAS) FOR DIGITAL MANUSCRIPT COLLECTIONS.
DOI: 10.5220/0002843702340239
In *Proceedings of the 6th International Conference on Web Information Systems and Technology (WEBIST 2010)*, page
ISBN: 978-989-674-025-2

be the solution to simplifying this multi-tiered process. PDF/A has the potential for electronically preserving manuscripts and providing high-resolution, web-deliverable digital replicas. Effectively, PDF/A could be used as a "true" digital archival surrogate (DAS) for long-term digital preservation and high-quality access to digital manuscript collections.

The PDF/A file standard (ISO: 19005-1) was developed to archive and preserve Portable Document Files (PDFs) and digitally-born documents, such as word-processing and publication files. PDF/A has grown quickly into an industry standard for preserving textual documents and is used in many companies, libraries and archives.

PDF/A is a versatile format. Like its cousin PDF, PDF/A can preserve the structure, layout, and visual appearance of a document, has the ability to embed metadata, and is designed to work with multi-page documents. The difference is that, unlike normal PDF, it has the added benefit of backward compatibility and is completely self-contained. These attributes gives PDF/A the capacity to deal with the complex nature of manuscripts, and hence, the potentially serve as a true digital archival surrogate.

This paper considers the notion of PDF/A as a digital archival surrogate (DAS) for digital manuscript collections. The paper will discuss (1) complexities associated with old manuscripts; (2) provide an overview of the PDF/A standard; and (3) discuss the adaptation of PDF/A as a DAS and its implications.

2 MANUSCRIPTS – COMPLEX OBJECTS

Manuscripts vary in form and appearance. They can include letters, diaries, illuminated text, musical scores, scientific notations, financial ledgers, or legal documents, such as this 1727 "Oath to King George II of England" (Figure 1). They can be bound, unbound, or rolled like a scroll. However, what they share in common is they are generally handwritten, manually produced, and page-oriented documents. When digitizing manuscripts, one must consider their inherent complexities.

First, the value in a manuscript is not only informational but also visual. Manuscripts are considered textual documents, since their purpose is to be read. Yet, because of their handwritten nature, they can be classified equally as visual (see Figure 1). Therefore, manuscripts hold a "dual nature"; they are both "purely textual" and "purely graphical" (Nicolas, 2009).

Second, manuscripts require a means to preserve the context and meaning of a document. This requires a means of maintaining a visual relationship between pages. This is no different than a book or a newspaper. Context and meaning is lost if the sequential continuity is not maintained between the pages. An orchestra, for example, will be lost if a page in a musical score is misplaced or missing.

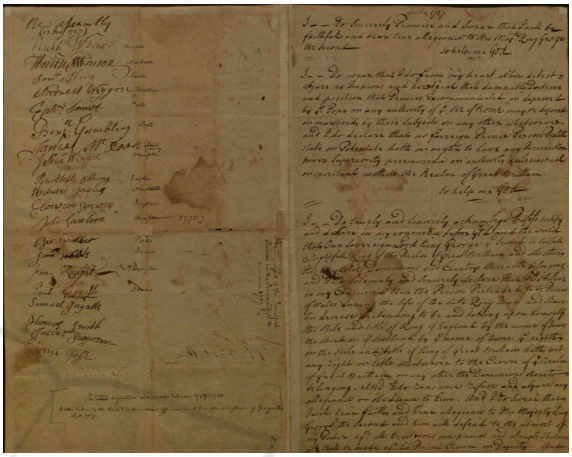

Figure 1: Handwritten Oath to King George II of England (1727) with multiple signatures. The figure represents a four-page document. It is simultaneously textual and visual (courtesy of Keene State College).

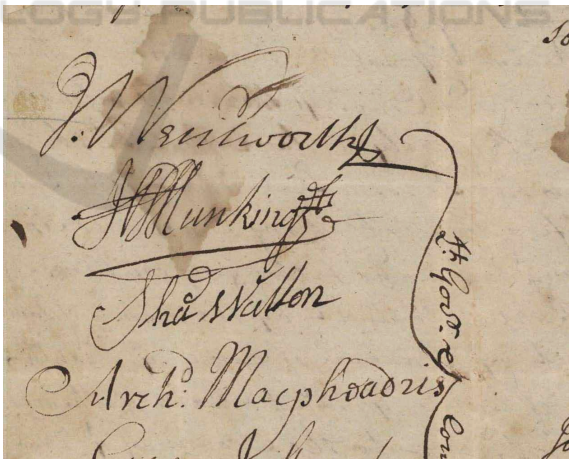

Figure 2: Detail of the Oath to King George II (1727) (courtesy of Keene State College).

Third, researchers need not only to be able to read the manuscript but also have a means to closely examine the handwriting, notations, and other aesthetic and tactile features of the manuscript. In effect, the manuscript is not merely viewed as text but also artifact. It is not sufficient to solely provide

a transcription of the text. The visual appearance of the manuscript is imbued with meaning (see Figure 2). For example, the researcher may want to study the quality of the signature to determine attributions or authenticity.

Fourth, non-linear navigation between pages is essential. We assume that documents are read linearly, left to right (or vice versa), and page-by-page. Consider if the document's author intends for the reader to "jump" forward or backward in a manuscript. A common example is a coda or repeat symbol in a musical score, directing a musician to move forward or backward in a piece.

These complexities necessitate different techniques to allow a digital surrogate to honestly portray a manuscript. Text must be able to be read, but the quality of that text must be able to be evaluated. This requires high resolution and the ability to effectively view these documents. Page sequence must be able to be maintained while being able to be navigated in non-linear ways. This requires complex methods to ensure that the document can be navigated.

3 PDF/A – AN OVERVIEW

PDF/A-1 or part 1 was approved as an open standard in May 2005 and published by ISO in September 2005. Developed by the working group ISO/TC 171 SC2, Document Imaging Application, Application issues, for which Association for Information and Image Management (AIIM), or The Enterprise Content Management Association as it has been renamed, acts as secretariat, the official standard is listed as "ISO 19005-1:2005. Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)" (pdf-tools, 3).

PDF/A is based on PDF Reference, Third Edition, version 1.4. This means its files can be read with Adobe Reader Version 5.0 or higher. Not all the parameters outlined in version 1.4 are used in the PDF/A standard. Some things have been excluded, such as transparency, sound, and movie actions.

The purpose of the exclusion is to make sure that the digital material can be preserved. Most of the exclusions are simply unable to be included due to their nature. The use of multi-media: Audio and Video or JavaScript or other executable files that are external are examples of this. Anything that lies outside of the document and is dependent on existing technological standards would limit the preservation life of the document. Encryption is also prohibited since it depends on a key that may be lost or forgotten. LZW Compression (a lossless data compression) is excluded because it is proprietary and therefore dependent on the whims of a company.

PDF/A-1 also mandates that the file follow certain rules. Again, these are to ensure that preservation of the digital material is ensured. Foremost, all fonts are to be embedded within the document, and that they must be legally (some fonts are not in the common domain) able to be embedded for universal rendering. Color spaces are also to be specified in a device-independent manner and PDF/A requires the use of Metadata about the document structure, creation and provenance. This includes basic metadata entries like title, author, creation date, modification date, subject, etc. For this purpose PDF/A uses Extensible Metadata Platform (XMP), though other metadata standards such as Dublin Core can be used if they are embedded within the XMP framework. In fact, Dublin Core elements are automatically created in Acrobat, Adobe's PDF software, when a PDF is made. XMP is a subset of Resource Description Framework (RFD), the language created by the World Wide Web Consortium for the development of the semantic web. The metadata can then be used by content management systems to tie documents together and to provide search capabilities.

3.1 PDF/A-1a & 1b

The PDF/A standard comes in two different compliance types, PDF/A-1a & 1b. PDF/A –1a adheres to all requirements as outlined in the standard. It is also known as a "tagged" PDF/A. It ensures preservation of a document's logical structure and content in natural reading order. Advantage to this level of compliance is that it can be read with basic text editing tools such as MS Notepad. PDF/A – 1b is for minimal compliance to the standard. It ensures that the text can be correctly displayed, but does not guarantee that extracted text will be legible or comprehensible.

3.2 PDF/A-2

The development of the PDF/A standard is not done. There is a new version of the PDF/A standard under development based on PDF Reference versions 1.5 – 1.7. It will be called PDF/A-2. All PDF/A-1 documents will be compatible with PDF/A-2, though PDF/A-2 documents may not be completely compatible with PDF/A-1 due to the added features. PDF/A-2 plans on the inclusion of JPEG 2000 image compression, more sophisticated digital signature

support, OpenType fonts, 3D graphics, audio/video content, and consistency with other PDF file formats like PDF/X (printing), PDF/E (engineering) and PDF/UA (universal accessibility). PDF/A-2 is set to be released in late 2009 or early 2010 (Fluckinger).

4 PDF/A = DAS

A digital archival surrogate (or DAS) is a hybrid of archival master and high-quality digital surrogate. PDF/A has the potential of acting as a true DAS that can deal with the inherent complexities of manuscript material. PDF/A can preserve the structure, layout, and visual appearance of a manuscript. The manuscript can be read as a textual document or viewed as a visual object.

PDF/A, via Adobe Acrobat Reader, provides for a variety of ways to view the document – single, multi-page, and/or side-by-side. This eliminates or greatly reduces any potential loss of context or meaning.

The Acrobat Reader browser allows for non-linear navigation. The Reader makes it simple to page forward or back, or jump to a selected page. The thumbnail function also allows for a simple means to navigate quickly between several pages.

The magnification function of Acrobat Reader allows for close, detailed examination of the document. The high-resolution of the images offer a high-quality replication.

PDF/A offers a means to embed documentary information regarding the original manuscript using Extensible Metadata Platform (XMP) functionality. Essentially any metadata standard, such as Dublin-Core, could be used to create a record containing information on the creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, and title or any other data that was needed.

4.1 Concerns Regarding Adaptation

The potential obstacles to PDF/A acceptance as a DAS are file size and efficiency as a web-deliverable surrogate, and questions of conversion and migration. Another concern, albeit minor, is the limitations caused by the complexities of manuscript material in what sub-class of PDF/A can be used for DAS.

4.1.1 Web Deliverability

PDF/A files are web-deliverable and can be read with Adobe Acrobat Reader or a plug-in version of the reader for a web browser. Since PDF/A is backward compatible, it provides an added advantage if a user is without the latest version of Adobe Acrobat Reader. The problem for the use of PDF/A, though, is that some of the files can be quite large, which can make it hard for their transfer over the web.

With the proliferation of high-speed, broadband Internet access, file size is less of an issue. Once a 100MB+ document would be a daunting task to deliver via the Web, this, though, has changed over the past few years. The substantial increases in electronic transfer rates, sending megabytes per second, have made large document delivery, although not instantaneous, quite tolerable.

The next generation of PDF/A, PDF/A-2, will go further to change attitudes towards web delivery with its use of JPEG2000 technology. JPEG 2000 is an ISO standard that has been published as ISO/IEC 15444. JPEG2000 allows for LZW (lossless) compression of large, high-resolution images. Compression will reduce file size significantly, which in turn will exponentially increase the efficiency of web delivery for PDF/A files.

4.1.2 Future Conversion & Migration

When adapting a new file standard, questions of future conversion and migration need to be asked. PDF/A is an ISO standard with a requirement for backward compatibility. It will be readable and accessible by future versions of Acrobat Reader or other PDF viewers.

PDF/A files can be converted to other formats, like TIFF, using commercial software programs like Adobe Acrobat or freeware such as MyMorph [http://docmorph.nlm.nih.gov/docmorph/] developed by the United States National Library of Medicine. Pages from a multi-page PDF/A can also be abstracted and converted using available software.

Future conversion and migration of PDF/A, if needed, will not be an issue as long as PDF/A continues to be an ISO standard, and as long as long as PDF technology continues to be an industry-wide standard.

4.1.3 Use of PDF/A-1b as DAS

The strong visual nature of manuscript material would make the sub-class PDF/A-1b the only current option for DAS since text attraction (via OCR) would be made difficult due to the handwritten characters. Handwritten text, currently, is very difficult to convert using any standard OCR capture program. PDF/A-1b, of course, meets all the minimal but important standards for the format.

4.2 New Digital Preservation Model

The adaptation of PDF/A as a DAS standard has the potential of shifting the current digital preservation model. The current model, as espoused by many leading institutions, calls for a two-tier system. The top tier directs for the creation of archival masters. The widely accepted standard is TIFF or JPEG2000, and some instances for PDF/A for text-based documents. Reference versions, the lower tier, include JPEG or GIFF for visual material, plain text in ASCII or UTF-8 character encoding, XML, PDF, or PDF/A for text or textual documents (CSU Libraries).

	Tier 1 Preservation	Tier 2 Web Derivate
Visual Materials	TIFF JPEG2000	JPEG GIFF
Textual Documents	PDF/A	PDF PDF/A TXT

Figure 3: Two-Tier Digital Preservation Model.

PDF/A can eliminate the two-tier system by creating a single tier. PDF/A, as DAS, serves as both preservation and derivative. This has great implications for file management requirements and would eliminate complicated file-naming conventions.

A single manuscript document may consist of several pages. Each page requires an individual archival master and a corresponding web-access derivative. Each file will also require a unique file name and a means to document the structure of the original manuscript.

Consider a manuscript collection of 1000 letters. Each letter consists of four pages. Each manuscript has a unique filename, individual file folders with separate sub-folders for archival masters and derivates. The collection would generate 4000 individual archival masters and 4000 derivates for a total of 8000 individual files. There would be a total of 1000 folders and 2000 sub-folders.

The DAS model would create a single four-page file. Each DAS would have one unique file name and no folder. Each DAS would contain descriptive data, unlike the previous model.

The model changes even more if PDF/A-2 is adapted. Not only will there be a reduction in file management requirements, one would also see a significant reduction in file storage requirements if compression is brought into the equation. This could have major impact on storage cost. JPEG2000 technology has a 1:50 compression ration. For example, in our 1000 manuscript digital library, each DAS averages 160MB in file size. The entire library would total 160GB. The use PDF/A-2 with LZW compression would reduce the library to 80GB. That is a significant reduction in storage and a potential savings in storage cost.

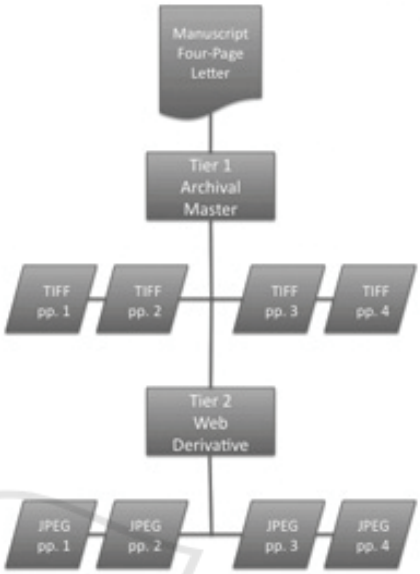

Figure 4: Two-Tier System.

Figure 5: DAS Single-Tier System.

5 CONCLUSIONS

PDF/A has the promise of serving as the standard for a digital archival surrogate (DAS) for old manuscript collections. It has the versatility to deal with the complexities of manuscript material, allowing for navigation within in a manuscript while easily maintaining page integrity. In addition, it has the

added advantage of allowing researchers to being to drill down into the image and study the manuscript's meaning as an artifact. PDF/A is able to accomplish this without the expertise and expensive software that causes large and small institutions to question the worth of digitizing manuscript material.

The next stage of research requires testing of PDF/A as a DAS through case studies using manuscript materials. The materials should consist of a wide-berth of document types, including, but not limited to, illuminated text, musical scores, handwritten letters, and personal diaries. Section 4 of this paper offers a useful model for PDF/A to be tested and compared against current digital preservation and access standards for efficiency and effectiveness in file management, organization, and storage, web deliverability, and usability, particularly in regards to researchers and scholars.

These tests should help verify PDF/A's viability as a true digital archival surrogate. Adaptation of PDF/A as DAS would have a major impact on digital preservation policies by simplifying the digital preservation model and creating a new standard that would ensure the tong-term preservation and digital access to the world's precious manuscript collections.

REFERENCES

Butler, S 2009. "Exploiting special collections: using digital methods to enhance their research and learning potential." *SCONFUL Focus*, 45, p. 81-86.

Colorado State University Libraries 2009. *CSU Digital Repository Preservation Format Support Policy*. Colorado State University Libraries, viewed 21 October, 2009, <http://lib.colostate.edu/repository/preservation. html>.

Fluckinger, D 2009. *PDF/A Takes Root as Digital Archiving Standard* (4 April 2009), PDFzone, viewed, 21 October 2009, <http://www.pdfzone.com>.

Nichols, SG 2007. "An Artifact by Any Other Name: Digital Surrogates of Medieval Manuscripts." In *Archives, Documentation and Institutions of Social Memory: Essays from the Sawyer Seminar*. edited by Francis X. Blouin, Jr. and William G. Rosenberg. University of Michigan Press.

Nicolas, S, T Pacquet and L Huttle 2003. "Digitizing Cultural Heritage Manuscripts: the Bovary Project." From ACM Symposium on Document Engineering, p. 55-57.

PDF-Tools.com, 2007. "White Paper: PDF/A – The Basics." *Understanding PDF White Papers*, v. 2.0 (January 22, 2007): p. 1-9, pdf-tools.com, viewed 21 October, 2009, <http://www.pdf-tools.com/public/downloads/whitepapers/whitepaper-pdfa.pdf>.