THE BENEFITS AND RISKS OF THE PDF/A-3 FILE FORMAT FOR ARCHIVAL INSTITUTIONS

AN NDSA REPORT

February 2014

Authors

- Caroline Arms, Library of Congress contractor
- Don Chalfant, National Archives and Records Administration
- Kevin DeVorsey, National Archives and Records Administration
- Chris Dietrich, National Park Service
- Carl Fleischhauer, Library of Congress
- Butch Lazorchak, Library of Congress
- Sheila Morrissey, ITHAKA
- Kate Murray, Library of Congress

Representing the NDSA Standards and Practices Working Group

CONTENTS

Persistent URL: http://hdl.loc.gov/loc.gdc/lcpub.2013655115.1

EXECUTIVE SUMMARY

As the most recent iteration of the PDF/A series of specifications, PDF/A-3 adds a single and highly significant feature to its predecessor PDF/A-2 (ISO 19005-2:2011). The PDF/A-2 specification permitted the embedding of other files as long as the embedded files were valid PDF/A files. PDF/A-3 (ISO 19005-3:2012) permits the embedding of files of any format (including XML, CSV, CAD, images, binary executables, etc.), within a PDF/A file. This new feature is intended to expand the functionality of PDF/A from "electronic paper" (albeit suitable for use over the very long term) to an archival format for a page-oriented document that can be bundled with related files. While a PDF/A file's primary document is intended to be robust against preservation risks over the very long term, PDF/A-3 does not require that the embedded files be considered archival content. A PDF/A-3 conformant reader is responsible for presenting only the primary document, but permits extraction of embedded files for use with other tools.

The NDSA PDF/A-3 Working Group recognizes that there are scenarios in which such an "archival bundling" use of PDF/A-3 might make sense. The concept of "hybrid archiving" has been proposed, where a source document in a less preservation-robust format is edited and then embedded in a "save-as" PDF/A-3 file, thereby having an always-current version of the document in a preservation-robust format. PDF/A-3 could also be employed in transactional workflows, where the "classic" PDF document would contain human-readable content, and presumably-equivalent content in embedded files of another format (typically XML), which would be processed by machine. Both scenarios would require PDF/A-3 processors that have been extended with custom capabilities tailored to the scenario.

For memory institutions, any such use of embedded files in PDF/A documents would depend on very specific protocols between depositors and archival repositories, clarifying the formats acceptable as embedded files, and defining a workflow that guarantees that the relationship between the PDF document and any embedded files is fully understood by the archival institution. Of greater concern is the possibility for PDF/A-3 to be used in some scenarios as a general-purpose bundling format, with the visible primary document of less long-term importance than the embedded files.

Further, the scenarios presented above are complicated by the fact that their articulation is not a part of the standard itself. Nor do the use cases articulated by proponents of the new standard address the complexities implied by retaining PDF/A's character as the digital equivalent of acid-free paper, while at the same time permitting its use as a general-purpose archival bundling format. In the latter case especially, the complexity of the PDF format and the potential and actual resulting faultiness of PDF rendering implementations and creating applications suggest that PDF/A-3 may be appropriate for use in controlled workflows, but may not be an appropriate choice as a general-purpose bundling format.

The working group strongly recommends that tools that create PDF/A-compliant documents be engineered to identify (through the pdfaid:part element) files that have no embedded files, or whose embedded files are all in PDF/A format, as compliant with PDF/A-2 rather than PDF/A-3.

The proposed creation by the PDF Association of a free and open source PDF validation tool might mitigate the long-term preservation risks constituted by the complexity of the PDF/A format as a bundling format. Absent

such robust validation tools, conversion of PDF files to PDF/A in preservation workflows remains a somewhat problematic preservation tactic.

Further, should the consensus of the preservation community be that PDF/A-3 is inappropriate as a general-purpose archival bundling format, then the community must identify and/or create tools that make it possible to bundle together complex digital objects with sufficient manifest information (i.e. metadata) to establish the relationship amongst the components within the bundle. Such a bundling format might usefully be based on the BagIt File Packaging Format and/or a constrained form of the ZIP format, such as the proposed ISO/IEC 21320-1 "Document Container File" specification (as of December 2013, this standard is still under development). ISO/IEC 21320-1 is intended to be a formally-specified interoperable and royalty-free subset of the proprietary ZIP format developed and maintained by PKWARE. BagIt is a hierarchical file packaging format for storage and transfer of arbitrary digital content. A typical serialization as a single file is based on ZIP or TAR. Operational elements of archival strategies often include (a) a list of preferred and acceptable formats with associated levels of commitment or confidence with respect to long-term preservation and access and (b) action plans for each format that govern procedures on ingest. Many institutions have given PDF/A a high priority on such lists and assumed that files that comply with the PDF/A-1 and PDF/A-2 standards require no normalization on ingest. PDF/A-3 must be treated separately from the other PDF/A versions in preference lists and for action plans and with more caution.

The introduction of such a problematic new feature in the latest version of the PDF/A family suggests that perhaps the community of memory institutions need to take a more strategic, active, and vocal role in the standards development process.

ABOUT THE NATIONAL DIGITAL STEWARDSHIP ALLIANCE

Founded in 2010, the National Digital Stewardship Alliance (NDSA) is a consortium of institutions that are committed to the long-term preservation of digital information. NDSA's mission is to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations. NDSA member institutions represent all sectors, and include universities, consortia, professional associations, commercial enterprises, and government agencies at the federal, state, and local levels.

More information about the NDSA is available from http://www.digitalpreservation.gov/ndsa/.

PDF/A-3 FORMAT BACKGROUND

In general, PDF/A-3 is a constrained form of ISO 32000-1,[1] also known as Adobe PDF 1.7.[2] PDF/A-3 is intended to be suitable for archiving page-oriented documents for which PDF is already being used in practice.

The specification, which goes by the full name of *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)* (or ISO 19005-3:2012 for short),[3] defines a file format based on PDF which provides a mechanism for representing electronic documents in a manner that preserves their static visual appearance over time, independent of the tools and systems used for creating, storing or rendering the files. The preservation of the "static visual appearance" is only possible if conforming PDF/A files are complete in themselves and require no external resources (for example, unembedded fonts) for them to render their pages properly.

PDF/A-3 adds a single and highly significant feature to its predecessor PDF/A-2 (ISO 19005-2). PDF/A-3 permits the embedding within a PDF/A of files in arbitrary formats (including XML, CSV, CAD, images, binary executables, etc.), not just other PDF/A files (as previously permitted in PDF/A-2[4]).

While allowing the embedding of files of any type, PDF/A-3 imposes requirements and adds capabilities that do not apply to ISO 32000-1 documents as a whole. For example, an explicit association must be made between each embedded file and the containing PDF or object or structure (e.g., image, page, or logical section) within the PDF/A-3 file. Embedded files that comply with PDF/A-3's requirements are termed "associated" files. Each associated file is required to include an "AFRelationship" key that must contain a value that represents the relationship of the embedded object to the "primary document" in the PDF/A-3 file.

The values for the AFRelationship keys defined in the PDF/A-3 specification are `Source`, `Data`, `Alternative`, `Supplement`, and `Unspecified`. Additionally, PDF/A-3 requires that MIME types be provided for associated files; the application/octet-stream MIME type is required if a more specific MIME type for the embedded file is not available. While explicit associations are made, the semantics of the associations may be too coarse-grained for some archival institutions. Human-readable descriptions for the associated files can be provided and are recommended.

---

[1] ISO 32000-1:2008 Document management -- Portable document format -- Part 1: PDF 1.7 Available at http://www.iso.org/iso/catalogue_detail?csnumber=51502

[2] Adobe Systems Incorporated. Document Management – Portable Document Format – Part 1: PDF 1.7, First Edition. Jul 2008. Available at http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/PDF32000_2008.pdf

[3] 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). Available at http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57229

[4] ISO 19005-2:2011 Document management -- Electronic document file format for long-term preservation -- Part 2: Use of ISO 32000-1 (PDF/A-2). Available at http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50655

The PDF/A-3 specification recommends that a conforming reader not render any non-PDF/A-3-conformant file embedded in a PDF/A-3 file, but does not forbid such rendition. The specification also recommends, but does not require, that a conforming reader enable the extraction of any embedded file, and, in addition, recommends but does not require that such an extraction be initiated by an explicit user action. Extraction is defined as "copying the raw byte stream of the embedded file data (after any decoding of filters that might be applied) from inside the PDF to some external byte storage system (e.g. disk or memory)." The standard notes that "these recommendations are to aid users in avoiding potential security risks inherent in opening unknown file types."

The changes introduced by the PDF/A-3 specification introduce both benefits and risks that archival institutions must consider. These issues are detailed in the "PDF/A-3 Analysis" section below.

PDF/A-3 STANDARDS WORKING GROUP

The PDF/A-3 Working Group was given a charter to investigate the PDF/A-3 standard by the NDSA Standards and Practices Working Group at its December 17, 2012 meeting.

In its charter, the NDSA PDF/A-3 Working Group stated that its members would research the pros and cons of using the PDF/A-3 standard in different preservation scenarios, including use as an extension to PDF/A-1[5] and PDF/A-2 in circumstances for which those formats have been adopted or recommended, and use as a wrapping or bundling format for various digital asset/media types, such as textual, audio, video, photo and GIS data.

The goal of the Working Group is to develop guidelines for the appropriate use of PDF/A-3 with respect to different scenarios, including both detailed technical information and a practical quick reference guide for end-users. In particular, the Working Group has worked to determine whether or not PDF/A-3 is appropriate as a de facto wrapping or bundling format for some or all media types or in particular circumstances, and, for circumstances in which PDF/A-2 has already been deemed an appropriate preservation format (primarily for textual documents), identify and explore the risks and opportunities are offered by the ability to embed non-PDF/A content in PDF files.

Working Group Members

The following are the members of the PDF/A-3 Standards Working Group:

- Caroline Arms, Library of Congress contractor
- Don Chalfant, National Archives and Records Administration
- Kevin DeVorsey, National Archives and Records Administration
- Chris Dietrich, National Park Service
- Carl Fleischhauer, Library of Congress
- Butch Lazorchak, Library of Congress
- Sheila Morrissey, ITHAKA

[5] ISO 19005-1:2005 Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1) Available at http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920

- Kate Murray, Library of Congress

Outreach

In addition to, and as part of, the Working Group's monthly phone meetings, the Working Group sought information from others engaged in the PDF/A-3 community. This outreach included:

- Attending the Digital Preservation Coalition (DPC) Workshop on PDF/A-3 in Leeds, UK, in March 2013.[6] This workshop included presentations by members of the PDF Association and the PDF/A Competence Centre. Many of the members of both of these organizations are engaged in producing commercial tools for the use of PDF/A-3. In addition, there were presentations, questions, and discussions by representatives of UK, US, and European memory institutions.

- A presentation on PDF/A-3 to the working group members at a monthly phone meeting by Stephen Levenson, an IT Specialist for Policy and Planning at the Office of the US Courts and the chair of the PDF/A working group. Mr. Levenson discussed both industry and government workflows (in Germany, USA, and Brazil) that looked to PDF/A-3 as a means of packaging together a "human-readable" document with a "source file" of some type, typically XML, that contained the same intellectual content, and that was intended for automated processing.

- A presentation on PDF/A-3 to the working group members at a monthly phone meeting by Duff Johnson, International Project Co-Leader for ISO 32000 (PDF), US Committee Chair for ISO 14289 (PDF/UA) since 2005, AIIM's Standards Committee Chair, a member of AIIM's Board of Directors, and Vice Chairman of the PDF Association.[7] In the course of that presentation, Mr. Johnson spoke of an effort to have the PDF Association commit resources to developing a free and open-source (F/OSS) PDF validator.

- A presentation by the NDSA PDA/A-3 working group at the PDF Association Technical Conference North America 2013 PDA/A day in Seattle in August 2013. This presentation, at the invitation of Duff Johnson after his presentation mentioned above, discussed the working group's understanding of the PDF/A-3 specification, its intended benefits, and its possible perceived risks. It was an occasion both to express the working group's understanding of, and concerns about, the PDF/A-3 specification, and to listen to members of the larger PDF community, including makers of PDF/A-3 tools. One of the observations distilled from the conversations was the very great need for PDF/A-3 advocates and toolmakers to educate users about the limitations and risks of the new feature of PDF/A-3, as understood from a long-term preservation perspective. It was suggested that it would perhaps be more appropriate to speak of the PDF/A family as "reliable" rather than "archival" formats.

[6] Digital Preservation Coalition, "Digital Preservation with Portable Documents: a workshop to introduce and discuss the PDF/A version." http://www.dpconline.org/events/details/55-DPC_PDFA3_briefing?xref=58

[7] PDF Association http://www.pdfa.org/pdf-association/

PDF/A-3 ANALYSIS

The Working Group has looked both at general pros and cons of PDF/A-3 and at scenarios where it might or might not be appropriate. This section discusses possibilities and risks associated with PDF/A-3 in general terms. The following section introduces more specific scenarios.

Pros/Possibilities

The Working Group recognizes that there are scenarios where the use of PDF/A-3 might make sense. Several of these scenarios are not archival in nature, but build on the benefit recognized in PDF/A as a constrained form of PDF intended to guarantee a reliable visual presentation of a page-based document. This benefit applies not only over time but also across computing environments.

One scenario is "hybrid archiving." In this widely applicable scenario, a document originates and is edited by a word-processor, but is managed in a document management system as a PDF/A-3 file, with the source word-processor file embedded within the PDF/A-3 document file. In this scenario, the document is considered archive-ready throughout its lifecycle. However, the embedded source file is considered as a non-archival artifact.

Another scenario where use of PDF/A-3 seems appropriate is one in which a document is essentially a record of a transaction between stages in a controlled workflow, in which the transaction record's format requires a consistent visual representation across devices and operating systems (a primary objective of PDF/A), but also requires or benefits from a machine-processable representation of part or all of the content or metadata. This machine-processable representation could be in XML and employ a standard or well-known schema. The Working Group learned of several examples where such use of PDF/A-3 is in use or active development. Examples with embedded machine-readable data structures include: the ZugFERD initiative[8] for exchange of invoices in Germany; the use of PDF/A-3 in US Bankruptcy Court exchanges with major creditors to reduce the need for manual data entry in records management systems; and in the legislative process of the Brazilian Senate.

Another use of embedded files in PDFs that the group learned about involved PDF as a wrapper for the audio from a court proceeding. The visible "primary document" provides context and metadata for the embedded sound file, intended for retention only for a specified period after the proceeding.

Possible uses or benefits of PDF/A-3 that might be relevant to archival institutions include:

- Associating machine-processable data extracts with particular charts or diagrams in a document.
- Embedding rich metadata in a "native" or standard format that has no RDF/XMP equivalent. This metadata would not be subject to changes by PDF processing tools that expect to update XMP metadata.

[8] German ZUGFeRD Format for electronic Invoices. http://www.pdflib.com/knowledge-base/pdfa/zugferd-invoices/

- Making explicit the function or relation of embedded files within PDF documents (i.e., non-PDF/A files). Files (or chunks of non-visible data) can already be embedded in PDF files in ways that are obscured from all but vendor-specific or application-specific reader/viewers (although not in PDF/A-1 or PDF/A-2 files). The requirement for file specification dictionaries with required relationship values in those dictionaries will make associated files embedded in a PDF/A-3 more obvious and discoverable through generic PDF/A-3 viewers. Ingestion into a preservation system could extract all associated files (and MIME types) and submit them to identification/characterization/validation procedures. (NOTE: the file-association feature is included in the draft ISO 32000 Part 2 specification and will be available to non-archival profiles of PDF once ISO 32000-2 is published.)

Cons/Risks

The PDF/A-3 specification allows an unlimited variety of files to be embedded in a PDF/A-3 document. Although the specification stipulates that a MIME type must be provided, MIME types do not necessarily provide the granular information necessary for proper file management. For example, the MIME type of `application/octet-stream`, by itself, is nearly useless. The PDF/A-3 specification does provide a required mechanism, the AFRelationship key, for expressing a relationship between the embedded file and the primary document, but the specification suggests no methods of verifying the stated relationship, nor is there any means in principle of doing so.

Even with the AFRelationship key mechanism, the embedding of files in any format creates concerns for memory institutions receiving such content. The PDF/A-3 specification makes it clear that the primary document is constrained by rules that will preserve its static visual appearance over time, but makes no requirements for long-term usability for embedded files. This might not be sufficient to free institutions from other custodial responsibilities in relation to the embedded files.

Many institutions are responsible for reviewing content for classification, copyright, or privacy concerns prior to public release and presumably these responsibilities would extend to attachments, in effect forcing them to maintain access to files in formats that they never intended to preserve. Similarly, archival institutions will have to take steps to assure that embedded files are not infected with computer viruses. In comparison with PDF/A-2 (which allows embedded files, as long as they are compliant PDF/A-1 or A-2 files), PDF/A-3 suggests long-term preservation challenges for cultural heritage institutions that may only fully be addressed through the development of detailed technical acquisition policies and/or prior agreements between the depositor and the archival repository. Such an agreement must clarify the formats acceptable as embedded files and how the depositor's workflow guarantees that the relationships between the primary PDF document and any embedded files is fully understood by the archival institution, so that appropriate rules are applied on ingest. The agreement might be based on a community best practice or a formal, legal regulation.

The primary focus of the PDF/A family of specifications, to this point, has been to define a format that supports the preservation of the content of page-based electronic documents. As the ISO 19005-3 specification states:

> The primary purpose of ISO 19005 is to define a file format based on PDF, known as PDF/A, which provides a mechanism for representing electronic documents in a manner that preserves their static visual appearance over time, independent of the tools and systems used for creating, storing or

> rendering the files. A secondary purpose of ISO 19005 is to define a framework for representing the logical structure and other semantic information of electronic documents within conforming files.
>
> Another purpose of ISO 19005 is to provide a framework for recording the context and history of electronic documents in metadata within conforming files.
>
> These goals are accomplished by identifying the set of PDF components that can be used, and restrictions on the form of their use, within conforming PDF/A files.

By enabling the embedding of arbitrary files within a PDF/A-3 file, the PDF/A-3 specification burdens the PDF/A family with an additional requirement: that of a packaging or bundling format. This additional requirement adds significant qualifications to the warranties implicit in terming the PDF/A family of formats as "archival."

The support for embedded files in PDF/A-3 threatens a significant principle that informed the development of PDF/A-1 and PDF/A-2: that the PDF/A document instance contain within itself everything necessary (given a conforming reader) to extract the complete semantic value of the document.

The PDF/A-3 specification itself neither addresses use-cases for embedding non-PDF/A files within a PDF/A-3 file instance, nor motivates the addition of this capability. Use cases offered by proponents of the new specification attempt the delicate balance of retaining PDF/A's character as the electronic equivalent of acid-free paper while suggesting that the format provide a means for bundling files related to the primary document – a function that might well constitute a contradiction of the first purpose even as it supports real-world use-cases.

These proponents make a distinction not made in the specification itself between "archival content" and "non-archival content" in the PDF/A-3 file. This distinction is drawn in a use case termed "hybrid archiving." The "archival content" is the constrained, reliable "digital paper" of PDF/A-1 and PDF/A-2, with all the warranties of those versions of the PDF/A specification. Embedded source files, permitted in PDF/A-3, comprise "non-archival" content, which no one is to consider having any archival warrant whatsoever. In this scenario, the embedded files are like "junk DNA" – artifacts that travelled along with the PDF/A-3 document in the process of its creation, and are present merely as artifacts of its process of production, but which have no standing of their own, nor any necessary semantic significance. The specification, indeed, declares that embedded files "should not be rendered by a conforming reader" (although, if an explicit user action initiates the process, it should enable the extraction of the byte stream comprising the embedded file). From the perspective of memory institutions, it's not clear in this scenario why these "non-archival" artifacts should continue as parts of the PDF/A-3 file, or, if they do continue, why a conforming reader should supply the capability to display information about them, or include the ability to extract their constituent byte streams.

The hybrid archiving use case is in contrast to a different use case being actively implemented in Germany. This is the "document communication" use case, where PDF/A-3 is employed for electronic invoicing. The "primary document" component of the PDF/A-3 file presents a "human-readable" form of an invoice. The embedded XML file is a version of the invoice intended for machine-processing. This pair of bytestreams is intended to complement each other in a workflow, with the Associated File (AF) dictionary specifying the relationship between these two "versions" of the document as "`Data`."

The difficulty, of course, is that there is no warrant, short of individual viewing of both the document and the embedded XML (which would of course involve an application in addition to a PDF viewer), that both "versions" of the invoice represent the same information. Given the assumed bulk machine processing of such documents, archival institutions may be obligated to understand the nature of the semantic content of embedded files and their relationship to the primary document in more complex ways than are indicated in the PDF/A-3 file in order to fulfill their mission. The availability and development of tools and services to assist archival institutions in understanding the semantic relationships amongst bytestreams in PDF/A-3 documents will assist them in making decisions on whether embedded files should be stripped out because they are not considered the official archival record and not guaranteed to be equivalent to the primary document or whether they can be retained.

As the example of wrapping audio of a court proceeding in a contextual document illustrates, the PDF/A-3 specification can be followed even when the primary content is an embedded file and the visible document merely a cover note rather than a transcript. An advantage, for the producer of such a PDF document, may be the ability to manage audio files within an existing document management system that is not designed to handle and deliver audio. Archival institutions will need to understand when PDF/A-3 might have been used in this way for files they receive.

The complexity and technical transparency challenges of the PDF format for the non-expert motivated the restrictions placed on PDF/A instance features. *Transparency* is used here as defined as a sustainability factor[9] by the Library of Congress in analyses of digital formats in their resource on Sustainability of Digital Formats. The technical transparency challenges and the complexity of the format itself (both potential and actual resulting faultiness of implementation in PDF writers and readers), and the absence of a reference implementation of any of the PDF format profiles, suggest that PDF may be inappropriate for use as a general-purpose bundling format or submission information package.

PDFs have proven useful for a variety of purposes, but they are not purpose-built as a bundling format. In this regard, the PDF is like a Swiss army knife, with the format continually expanding to meet new user needs. It is illuminating to compare this design approach with the UNIX philosophy of writing programs that do one thing, and do it well. This is not to suggest that either design approach has exclusive merit, only to suggest that for the long-term archival community, a purpose-built bundling format may be more appropriate.

Because of the potential complications for the long-term preservation of PDF/A-3 files, the working group strongly recommends that tools that create PDF/A-compliant documents be engineered to identify (through the pdfaid:part element) files that have no embedded files, or whose embedded files are all in PDF/A format, as compliant with PDF/A-2 rather than PDF/A-3.

There is currently no robust vendor-independent mechanism for assessing that a PDF/A file does in fact comply fully with the standard and the conformance level it claims in its internal metadata. Experience has shown that attempts to create a PDF/A can easily result in an invalid file. The file produced may identify itself to be a PDF/A although an error message indicates a problem with its generation. Because the Adobe Reader is very forgiving,

9Caroline R. Arms and Carl Fleischhauer. "Sustainability of Digital Formats: Planning for Library of Congress Collections". http://www.digitalpreservation.gov/formats/sustain/sustain.shtml#transparency

the document may look fine to its creator. However, other PDF processors may reject it because it is technically invalid.

SCENARIOS FOR PDF/A-3 USE

The Working Group has reached the conclusion that whether PDF/A-3 is an appropriate format for the long-term preservation of content depends heavily on the type of content, the nature of the workflow that created it, and whether the archival submission process allows for detailed negotiation on allowable formats for embedded files. Several scenarios have been developed to present risks and benefits in particular situations. In many cases, these scenarios are hypothetical but they have been chosen as representative of real-world examples based on institutional experience. A general scenario is first, followed by some scenarios that are specific to particular contexts or institutions.

General Scenario

Scenario for embedding supporting data in PDFs for scholarly documents.

Background and Assumptions:

Many scholarly journals are now expecting authors to supply supporting machine-readable data extracts used as the basis for charts and summary tables in articles. Publishers differ as to data formats accepted and as to how they support access for users to the data. One common form for distributing scholarly articles is as PDF documents. An assumption behind this scenario is that a journal publishing system supporting the generation of compliant PDF/A documents is adapted to support generation of PDF/A-3 documents with data extracts embedded and linked to the appropriate tables and graphics.

Scenario:

The publisher of a scholarly journal that already uses PDF/A as one of the formats disseminated to users, extends its peer review, editorial, and production workflows to support generation of PDF/A-3 files for articles that have supporting data.

Desirable outcomes and benefits:

A peer reviewer using a compliant reader to consult the PDF/A-3 version of an article can easily extract the data extract that supports a table (or chart) of interest and evaluate whether the data does indeed support the table as presented. The reviewer can also evaluate whether the data extract is consistent with the journal's policies.

A scholar using a compliant reader to consult the PDF/A-3 version of an article can easily extract the data extract that supports a table (or chart) of interest and perform his own analysis on the data.

A publisher may be able to simplify its content management system by storing supporting data for an article within the article document.

Risks and disadvantages:

A specific concern applies to numeric or tabular data. Numeric or other tabular data cannot be used effectively by anyone other than the authors without adequate description of the data collection context, the coding system employed, the accuracy of measurements, etc. As with any other bundling format, embedding a data extract to support a chart in a PDF should not be seen as a substitute for following practices appropriate to the data category or discipline to archive the underlying data.

Additionally, the resulting PDF/A-3 files may be too large for convenient dissemination to end users, given that most readers will not immediately want to consult the underlying data.

SPECIFIC SCENARIOS

U.S. NATIONAL ARCHIVES AND RECORDS ADMINISTRATION (NARA) SCENARIOS

U.S. Government agencies are required to deposit materials appraised as permanently valuable government records with the National Archives and Records Administration (NARA). Frequently, agencies store information in complex and often proprietary data systems with the permanent records forming only a small subset of the complete system. The determination of what digital materials from such systems are considered permanent records, and when and in what format they will be transferred is determined by NARA appraisal, accessioning, and processing staff and by agency records officers and IT staff.

The time between appraisal and transfer to NARA can be years or even decades. Hence, federal agencies need to plan well in advance of transfer to NARA. Ideally, each agency would consider choosing appropriate formats before a record is created and periodically review the format transfer guidance for revisions as part of an overall electronic records management plan.

Many agencies use records management systems built to conform to the Department of Defense Electronic Records Management Software Applications Design Criteria Standard (DoD 5015.02-STD).[10] As of late 2013, the current version of this standard is from April 2007. A specific set of content types is included in the standard; audio and video are not included.

In the scenarios presented below PDF/A-3 might be determined to be acceptable for legal transfer from an agency to NARA. It should be noted that these examples are hypothetical and do not represent current official policies. Since PDF/A3 is a recently developed standard, NARA, like most memory institutions, has not adequately evaluated the format for its suitability for use when transferring permanent records.

Scenario for PDF/A3 as an acceptable container for "as-built" engineering drawings originating from a CAD system.

[10] DoD 5015.02-STD. Electronic Records Management Software Applications Design Criteria Standard, April 25, 2007. http://jitc.fhu.disa.mil/cgi/rma/standards.aspx

Background and Assumptions:

One scenario in which NARA might allow for the transfer of permanent digital records in the PDF/A-3 format involves exporting a related subset of data and documents from a technical data system that includes architectural drawings and computer aided design (CAD) files. In this example, PDF/A-3 is a carrier for transferring both the permanent records required by NARA as well as associated non-archival non-permanent record content. It should be noted that an alternative scenario might include the agency outputting CAD data as a 3D PDF or some other CAD format. The choice of which scenario is "best" depends on what functionality is determined to be an inherent part of the record that must be preserved for future researchers.

Scenario:

The scenario involves several generalizable steps:

1. The creating agency contacts the archives regarding the transfer of scheduled permanent electronic records from a data system, in this case a CAD system.

2. NARA appraisal and accessioning staff defines the scope of the system's record material and identifies PDF/A-3 as an appropriate transfer format.

3. NARA appraisal, accessioning, and processing archivists identify appropriate files and documentation for inclusion in the PDF/A-3.

For this particular hypothetical situation, the NARA recommendation might be as follows. To serve as the official as-built drawing copies, a completed set "as-built" drawings resulting from an engineering project will be saved as TIFF images and watermarked "AS-BUILT". The TIFF images are incorporated as primary content in a PDF/A-3 document along with the embedded native CAD system working files used to create the drawings. The relationship of the native CAD files to the document is to be Source. An appropriate MIME type will be used for each embedded file at the time of file creation.

4. Agency records officer and/or IT staff produce compliant PDF/A-3 files for transfer.

As agreed, these files have embedded native CAD files and incorporate appropriate relationship codes and MIME types.

5. NARA accessions the PDF/A-3 files as official records and provides preservation actions as needed to the page-oriented content of the PDF/A-3.

After accessioning, NARA would only guarantee the preservation of the page-oriented content of the PDF/A-3, in this case the TIFF images of the as-built drawings. In this scenario, the native CAD files would exist only as artifacts of the accessioned document. They would not be accorded preservation actions such as migration but would be maintained as data.

Desirable outcomes and benefits:

The PDF/A-3 provides a stable container for encapsulating both the official copy as well as files from the original system. Encapsulating the original working files along with the official records within a PDF/A-3 provides the creating agency a tidy package for successful management of their related files prior to transfer. The transfer of the PDF/A-3 meets the records creator's obligation to NARA that a permanent record of the as-built drawings is archived for future reference.

One additional advantage is realized. The records creator, or other qualified patrons, would be able to re-use the technical work for a future design if they have the capability and software to render the native system files. The PDF/A-3 file would be a designated record copy in an accepted, unalterable format along with an embedded non-official copy in a format primarily used by the record creator during the record's active use and which supports the dynamic manipulation of record's data.

With the resulting document NARA is able to preserve, manage, and to provide access to official records. Staff and researchers are able to access permanent records from a complex data system in an accessible format.

Risks and disadvantages:

The associated embedded data files are in formats that may be not supported in the future and this information is vulnerable to loss.

The native system, non-record files might contain personal information or other sensitive data but would be difficult to review by archives staff.

The relationship between the embedded files may be unclear or erroneous.

Scenario for PDF/A3 as an acceptable container to circumvent 5015.2 records management application restrictions.

Background and Assumptions:

An agency using a records management system (RMS) that strictly enforces outdated requirements of the DoD 5015.2 specification is unable to manage long-term temporary records in formats that do not comply with that standard. To circumvent this restriction they embed non-conformant audio and video files in PDF/A-3 wrappers so that they can be managed in the RMS.

Scenario:

The scenario involves several generalizable steps:

1. The agency has a DoD 5015.2 compliant RMS that restricts the types of formats that can be ingested into it based on that standard.

2. The agency identifies the audio and video records as having long-term temporary value.

3. The agency does not consider NARA's transfer requirements regarding sustainable formats.

Desirable outcomes and benefits:

Agency staff embeds audio and video files into PDF/A-3 wrappers that can be ingested into their DoD 5015.2 compliant records management system.

The agency uses the primary document of the PDF/A-3 to provide narrative context for the embedded audio or video.

Risks and disadvantages:

The associated embedded data files are in formats that may be not supported in the future and this information is vulnerable to loss. Another issue is that the PDF container will make any preservation actions on the embedded files (e.g. migration/emulation) more challenging or complex.

The agency is unable to adequately search for and provide access to the audio and video records embedded in the PDF/A-3 files.

NARA refuses to accept transfer of PDF/A-3 formatted records for audio and video subsequently re-appraised as permanent.

The PDF container adds additional complexity to any required preservation actions for embedded long-term-temporary files.

LIBRARY OF CONGRESS SCENARIO

The Library of Congress acquires content for its collections in a number of ways. In only occasional circumstances is it able to negotiate actively with a depositing or donating party with respect to digital format. The bulk of the general collection is acquired through copyright deposit, for which preferred or acceptable digital formats are listed in a document published by the U.S. Copyright Office as Copyright Circular 7b: "Best Edition" of Published Copyrighted Works for the Collections of the Library of Congress.[11] Collections of personal or organizational "papers" are usually donated as is. Content harvested from the Web for archiving is usually only available in one format.

Scenario for harvesting PDF documents from the web and converting to PDF/A-3

Background and Assumptions:

The Library of Congress has a pilot program harvesting and cataloging documents in PDF format from the web, when a recommending officer has identified the work as fitting LC collection policies for books. One typical use

[11] U. S. Copyright Office, Library of Congress, "Best Edition of Published Copyrighted Works for the Collections of the Library of Congress http://www.copyright.gov/circs/circ07b.pdf

is for documents published outside the U.S. and not subject to U.S. copyright deposit. For these, permission to harvest and archive is sought from the publishing website. Another use being explored is for PDFs that have embedded Creative Commons licenses, or equivalent, that permit non-commercial use with attribution. There is no opportunity for the Library of Congress to negotiate with the content owner over the characteristics of the PDF.

Scenario:

A possible workflow for ingestion includes conversion of source PDFs to PDF/A. This would allow fonts to be embedded (assuming fonts used were available and legally embeddable) and permit addition of metadata that supports Library of Congress services for access and preservation. Current experiments use PDF/A-1, but PDF/A-3 files could potentially be used with the source PDF embedded inside the modified PDF. A relationship type of `Source` (to the entire document) and a MIME type of `application/pdf` would be used for the embedded file.

Desirable outcomes and benefits:

The modified PDF/A file (without the embedded source) is in a form that has desirable characteristics (such as embedded fonts) for supporting both current access and for maintaining long-term usability of the content in the source PDF.

The PDF/A-3 resource (with the embedded source document) can be managed as a single master file and used to derive a smaller service PDF for dissemination to users.

Future users requiring assurance of the authenticity of the disseminated version can be given access to the source PDF as well.

Risks and disadvantages:

The PDF/A-3 with embedded source document will be larger than the original source file and may be inconvenient as a file for user access, for example on mobile devices. This leads to the need for storing multiple versions or for building dissemination systems that derive service PDFs on the fly.

Some source PDF documents will incorporate features prohibited in PDF/A, for example, a font subject to a license that prohibits embedding. This will prevent conversion to PDF/A and hence to PDF/A-3 for some source PDFs. This leads to a more complicated ingestion workflow with more opportunities for error.

Without more robust validation tools than available in 2013, the conversion to PDF/A and PDF/A-3 may be unreliable.[12]

[12] Jamin Koo and Carol C. H. Chou. "PDF to PDF/A: Evaluation of Converter Software for Implementation in Digital Repository Workflow." New Review of Information Networking Vol. 18, Iss. 1, 2013. Available at http://www.tandfonline.com/doi/full/10.1080/13614576.2013.771989

U.S. HOUSE OFFICE OF THE LEGISLATIVE COUNSEL (HOLC) SCENARIOS

The Office of the Legislative Counsel provides legislative drafting services to the committees and Members of the House of Representatives on a non-partisan, impartial, and confidential basis. The goal is to work with committees and Members to understand their policy preferences in order to implement those preferences through clear, concise, and legally effective legislative language.

The two scenarios that follow are based on discussions with Joe Carmel, Retired Chief, Legislative Computer Systems and contractor, U.S. House Office of the Legislative Counsel and Reynold Schweickhardt, Director of Technology Policy, Committee on House Administration. These scenarios concern the use of PDF/A-3 for internal workflows. They are not directly related to processes and policies that apply to public access to published legislative information.

Scenario for embedding source files as an aid to maintain attorney-client confidentiality

Background and Assumptions:

The House Office of the Legislative Counsel (HOLC) maintains attorney-client confidentiality. While the office routinely provides Members and their staff with PDF files, the House's process requires that some of the source files subsequently be made available to other House offices.

The legislative drafting and publication system is based on XML, from which PDF documents are generated for convenient reading.

Scenario:

HOLC prepares a draft amendment for an individual Member and generates a PDF document for convenient review. HOLC chooses to generate a PDF/A-3 file with the equivalent XML embedded in the PDF, using the relationship `Source` to associate the embedded file with the entire primary document. The Member reviews the PDF and judges that the amendment is ready for sharing with another office, which will need access to the source XML file for subsequent stages in the legislative workflow. The Member sends the file he received from HOLC.

[Note: the process of embedding XML source is in current use, but as of late 2013, the PDF file is not a PDF/A-3 file. Files with embedded attachments (primarily amendments) can be seen at the House Rules Committee website. An example is http://www.rules.house.gov/amendments/DELAMD_002_xml318131552385238.pdf.]

Desirable outcomes and benefits:

Emailing a single PDF file with an embedded source file enables HOLC to maintain attorney-client confidentiality since the client (a Member) can subsequently provide the PDF file to other House offices requiring access to the source XML for future action.

The potential for mismatching two files (PDF and XML source) due to human error is eliminated.

Risks and disadvantages:

None identified, since the process for creating the PDF and embedding the source XML is automated and guarantees equivalence.

Scenario for embedding XML files to align PDF page/line locations with XML when editing amendments

Background and Assumptions:

A project still in development known as the Amendment Impact Program (AIP) is being designed to show the impact of an amendment on a bill based on separate amendment document language.

Page-numbering for bills is applied when the human-readable versions are derived from the source XML. A sample amendment instruction might be: On page 6, line 5, insert “motivated” before “Congress”. This processing cannot be accomplished without having both the PDF and source files available as the PDF files contain the page and line numbers while the source documents do not incorporate page-numbering

Scenario:

A PDF/A-3 file with embedded XML is constructed for a bill. The proposed project would develop a system to take this file and amendment files and generate revised XML source.

Desirable outcomes and benefits:

The PDF/A-3 with embedded XML source can be generated and transmitted, assuring that the PDF rendition and the source XML are correctly matched.

The combination of the PDF and XML source file permits matching the page/line locations from the PDF file to the correct location within the source XML file.

Risks and disadvantages:

None identified, since the process for creating the PDF and embedding the source XML is automated and guarantees equivalence.

CONCLUSIONS

The most important characteristic of the PDF/A formats is that they are constrained in ways that are designed to preserve the "static visual appearance" of a page-oriented document. The differences between PDF/A-1 (2005) and PDF/A-2 (2011) relate primarily to the fact that they are based on different generations of the underlying PDF specification. The overarching intent did not change; the entire file was intended to be suitable for long-term preservation. PDF/A-1 and PDF/A-2 are both suitable for long-term preservation, although they would not necessarily be the preferred format for any particular content item. The introduction of arbitrary embedded files to PDF/A-3 introduces significant concerns for memory institutions. A PDF/A-3 file may have any other type

of file embedded within it and the only statement that the standard makes in relation to preservation intent is that a compliant PDF/A-3 reader should not render embedded files, but merely support their extraction. The standard is silent as to whether the embedded content may be considered essential to full understanding or use of the primary document whose visual appearance is preservable. The result is that, accepting a PDF/A-3 file without additional rules or active negotiation, may lead an archival institution to acquire embedded content in a format that it did not expect and cannot deal with and whose relationship to the primary document may be unclear.

The members of this working group recognize that the PDF/A-3 format serves important business needs unrelated to preservation for the long term. One general application is the inclusion of structured machine-readable information within a PDF that is intended for viewing by a human with visual appearance retained across computing systems. Many workflows or transactions that have no relationship to long-term preservation can benefit from this capability. Another general application is known as "hybrid archiving." The context is that an entity such as a corporation is required to "archive" certain categories of document for a certain period to satisfy government regulations. These documents start life in an editable form and at some stage are deemed to be final; if intermediate drafts are saved as PDF/A-3 with the editable form embedded, the document is archivable at any stage, simplifying records management. These business needs are driving the development of tools for creating PDF/A-3 files. Some will certainly be deposited with memory institutions.

The complexity of the PDF format and the potential and actual resulting faultiness of PDF rendering implementations and creating applications suggest that PDF/A-3 may be most appropriate for use in controlled workflows, but may not be an appropriate choice as a general-purpose bundling format.

However, the proposed creation by the PDF Association of a free and open source PDF validation tool might mitigate the long-term preservation risks constituted by the complexity of the PDF/A format as a bundling format. Absent such robust validation tools, conversion of PDF files to PDF/A in preservation workflows remains a somewhat problematic preservation tactic.

Further, should the consensus of the preservation community be that PDF/A-3 is inappropriate as a general-purpose archival bundling format, then the community must identify and/or create tools that make it possible to bundle together complex digital objects with sufficient manifest information (i.e. metadata) to establish the relationship amongst the components within the bundle.

Such a bundling format might usefully be based on the BagIt File Packaging Format[13] and/or a constrained form of the ZIP format, such as the proposed ISO/IEC 21320-1 “Document Container File” specification[14] (as of December 2013, this standard is in the Committee Draft stage). ISO/IEC 21320-1 is intended to be a formally-specified interoperable and royalty-free subset of the proprietary ZIP format developed and maintained by PKWARE. BagIt is a hierarchical file packaging format for storage and transfer of arbitrary digital content. A typical serialization as a single file is based on ZIP or TAR.

---

[13] The BagIt File Packaging Format (V0.97). Available at http://www.digitalpreservation.gov/documents/bagitspec.pdf

[14] ISO/IEC CD 21320-1.2, Information technology -- Document Container File -- Part 1: Core. Available (as committee draft) at http://isotc.iso.org/livelink/livelink?func=ll&objId=15629477&objAction=Open&viewType=1. Final version, when published, will be available at http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=60101

Since PDF/A creation tools may generate files that identify themselves as PDF/A-3 even when no files are embedded, one recommended triage step for PDF/A-3 during ingest is to check whether a PDF/A-3 file has any embedded files. If no files are embedded then a PDF/A-3 can be treated in the same way as PDF/A-1 and PDF/A-2 files in a preservation ingest workflow. The working group strongly recommends that tools that create PDF/A-compliant documents be engineered to identify (through the pdfaid:part element) files that have no embedded files, or whose embedded files are all in PDF/A format, as compliant with PDF/A-2 rather than PDF/A-3.

Memory institutions and archival repositories should be prepared to treat PDF/A-3 files with care and develop procedures that fit their overall mission and policies. How embedded files should be treated depends on the context for creation, the expressed relationships that embedded files have to the primary document, the expectations of future users, and the policies of the archiving institution.

Operational elements of archival strategies often include (a) a list of preferred and acceptable formats with associated levels of commitment or confidence with respect to long-term preservation and access and (b) action plans for each format that govern procedures on ingest. Many institutions have given PDF/A a high priority on such lists and assumed that files that comply with the PDF/A-1 and PDF/A-2 standards require no normalization on ingest. PDF/A-3 must be treated separately from the other PDF/A versions in preference lists and for action plans.

The introduction of such a problematic new feature in the latest version of the PDF/A family suggests that perhaps the community of memory institutions need to take a more strategic, active, and vocal role in the standards development process.

APPENDIX A: RESOURCES AND REFERENCES

Adobe Systems Incorporated. PDF Reference, Third Edition, Version 1.4. Addison-Wesley, 2001. ISBN: 0-201-75839-3. Also available online at http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/pdf_reference_archives/PDFReference.pdf.

Adobe Systems Incorporated. *Document Management – Portable Document Format – Part 1: PDF 1.7, First Edition.* Jul 2008. Available at http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/PDF32000_2008.pdf

Caroline R. Arms and Carl Fleischhauer. "Sustainability of Digital Formats: Planning for Library of Congress Collections." Available at http://www.digitalpreservation.gov/formats/

The BagIt File Packaging Format (V0.97). Available at http://www.digitalpreservation.gov/documents/bagitspec.pdf

Digital Preservation Coalition, "Digital Preservation with Portable Documents: a workshop to introduce and discuss the PDF/A version." Available at http://www.dpconline.org/events/details/55-DPC_PDFA3_briefing?xref=58

DoD 5015.02-STD. Electronic Records Management Software Applications Design Criteria Standard, April 25, 2007. Available at http://jitc.fhu.disa.mil/cgi/rma/standards.aspx

German ZUGFeRD Format for electronic Invoices. Available at http://www.pdflib.com/knowledge-base/pdfa/zugferd-invoices/

ISO 32000-1:2008 Document management -- Portable document format -- Part 1: PDF 1.7 Available at http://www.iso.org/iso/catalogue_detail?csnumber=51502

ISO 19005-1:2005 Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1) Available at http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920

ISO 19005-2:2011 Document management -- Electronic document file format for long-term preservation -- Part 2: Use of ISO 32000-1 (PDF/A-2). Available at http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50655

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). Available at http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57229

ISO/IEC CD 21320-1.2, Information technology -- Document Container File -- Part 1: Core. Available (as committee draft) at

http://isotc.iso.org/livelink/livelink?func=ll&objId=15629477&objAction=Open&viewType=1. Final version, when published, will be available at http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=60101

Jamin Koo and Carol C. H. Chou. "PDF to PDF/A: Evaluation of Converter Software for Implementation in Digital Repository Workflow." *New Review of Information Networking* Vol. 18, Iss. 1, 2013. Available at http://www.tandfonline.com/doi/full/10.1080/13614576.2013.771989

Butch Lazorchak. "All In! Embedded Files in PDF/A" posted on The Signal Digital Preservation Blog, Nov.13, 2012. Available at http://blogs.loc.gov/digitalpreservation/2012/11/all-in-embedded-files-in-pdfa/

Sheila M. Morrissey, "The Network is the Format: PDF and the Long-term Use of Digital Content" in *Proceedings of Archiving 2012*, Volume 8 (June 2012). Available at http://www.portico.org/digital-preservation/wp-content/uploads/2012/12/Archiving2012TheNetworkIsTheFormat.pdf

PDF/A Competence Center. Available at http://www.pdfa.org/competence-centers/pdfa-competence-center/

PDF Association. Available at http://www.pdfa.org/pdf-association/

U. S. Copyright Office, Library of Congress, "Best Edition of Published Copyrighted Works for the Collections of the Library of Congress. Available at http://www.copyright.gov/circs/circ07b.pdf

GLOSSARY

Associated file

See Embedded file.

Bundling format

Digital format designed to aggregate a collection of related digital files into a single file for convenience of transmission or storage. [Note: the most widely used bundling format is ZIP. In contrast to wrapper formats, files encapsulated in bundling formats are often explicitly extracted as individual files before use.

CAD

Computer-aided design.

CSV

Comma-separated values (file format).

Embedded file

File stored inside a PDF/A-3 document in compliance with ISO 19005-3.

File Specification Dictionary

Data structure defined in the PDF specification, used in PDF/A-3 to relate embedded files to the document as a whole or specific sections of content.

GIS

Geospatial information system .

ISO

International Organization for Standardization.

ISO 19005-1

Specification of PDF/A-1 (2005).

ISO 19005-2

Specification of PDF/A-2 (2011).

ISO 19005-3

Specification of PDF/A-3 (2012).

MIME type

Value for a media type code, as assigned by IANA. [See http://www.iana.org/assignments/media-types].

PDF

Portable Document Format, file format defined in ISO 32000-1:2008 or Adobe PDF Reference 1.4.

PDF/A-1

Restricted subset of PDF, based on Adobe PDF Reference 1.4, as defined in ISO 19005-1:2005.

PDF/A-2

Restricted subset of PDF, as defined in ISO 19005-2:2011.

PDF/A-3

Restricted subset of PDF, as defined in ISO 19005-3:2013.

Primary Document

The content representing the page-oriented document as would be presented through a conforming interactive PDF/A-3 reader, not including any embedded files.

Transparency

The degree to which the digital representation is open to direct analysis with basic tools. [See http://www.digitalpreservation.gov/formats/sustain/sustain.shtml - transparency].

Wrapper format

Digital file format designed as an envelope for one or more streams of related digital content, where the stream format is specified independently. [Note: For example, the QuickTime format is a wrapper that typically contains audio or video bitstreams, which may conform to one of many compression encodings. In contrast to bundling formats, files compliant with a wrapper specification are typically used as is by a reader/player that understands both the wrapper and the encoding of encapsulated bitstreams.] .

XML

Extensible Markup Language.