Beyond TIFF and JPEG2000: PDF/A as an OAIS submission information package container

Yan Han
The University of Arizona Libraries, The University of Arizona, Tucson, Arizona, USA

Received 26 June 2015
Revised 13 July 2015
Accepted 15 July 2015

Abstract

Purpose – The purpose of this paper is to introduce PDF/A to replace TIFF as the preferred file format for digitization of textual documents. In addition, PDF/A can be used as an open archival information system (OAIS) submission information package (SIP) container to reduce digitization and digital preservation costs.

Design/methodology/approach – The author first reviewed the current digitization guidelines, the OAIS model and provides on an overview of the development PDF and PDF/A as international standards. Then literature review of the uses of PDF/A is presented. The author analyzed pitfalls of TIFFs as the preferred format for digitization, and showed how to use PDF/A to code digitization SIP.

Findings – TIFF file format has been the preferred master file format by Federal Agency Digitization Guidelines Initiative digitization guidelines for the past 20 years. However, there are drawbacks of TIFF format. Literature reviews show that PDF/A has been the preferred standard for coding born-digital documents in court, government and business sectors. PDF/A-2 and PDF/A-3 are relatively new standards released after 2010. However, few understood the standards and have utilized the full potentials in digitization. The author shows that PDF/A can be used as an OAIS SIP container.

Practical implications – In order to delivery OAIS SIPs, current practices require a combination of files, directories and various types of metadata. The author shows that PDF/A (PDF/A-2 and/or PDF/A-3) can be a better file format for textual document digitization with coding various types of metadata in extensible metadata platform and arbitrary file/data can be coded in PDF/A-3. These features in PDF/A provide much better ways to deliver SIPs in a cost-efficient manner.

Originality/value – PDF/A has been recognized as the preferred standard for born-digital documents, but it has not been used as the preferred file format for digitized materials. The author recommends that: PDF/A with lossless JPX compressions as the preferred file format; and PDF/A with lossless JPX compressions along with metadata/data as the preferred OAIS SIP container. As a result, the uses reduce costs in digitization and digital preservation and also increase productivity. The author recommends to update the national and international digitization practices using PDF/A.

Keywords Digital documents, Digitization, Standards, Digital preservation, PDF/A

Paper type Research paper

1. Background

1.1 Overview of current digitization guidelines

Libraries, museums and archives have been digitizing materials for preservation and access since the 1990s. Over the past 20 years, Federal agencies such as the National Archives and the Digital Library Federation (DLF) have published several critical digitization guidelines and best practices, which have been the de facto standards for digitization projects in libraries, archives and museums. These guidelines were written by experts and specify in great details in every aspects of digitization including file format and various metadata considerations. These guidelines greatly influence almost

The author would like to thank Leonard Rosenthol, Project leader for ISO PDF/A and Adobe PDF Architect for his comments on TIFF, PDF and PDF/A file formats.

Library Hi Tech
Vol. 33 No. 3, 2015
pp. 409-423

0737-8831
DOI 10.1108/LHT-06-2015-0068

all institutions' digitization projects, create standardized digitization practices in the communities and contribute to access and preservation of scholarship and culture heritage resources in reaching various audiences. These guidelines are:

- 2002: The DLF published *Benchmark for Faithful Digital Reproductions of Monographs and Serials* (Version 1), available at: http://purl.oclc.org/DLF/benchrepro0212
- US National Archives and Records Administration (2004): *Technical Guidelines for Digitizing Archival Records for Electronic Access: Creation of Production Master Files – Raster Image*s.
- 2008: Bibliographic Center for Research (BCR) published its *BCR's CDP Digital Imaging Best Practices Version 2.0*, available at: http://books.google.com/books/about/BCR_s_CDP_Digital_Imaging_Best_Practices.html?id=vjeEXwAACAAJ
- Federal Agency Digitization Guidelines Initiative (2010) released digitization guidelines related to audio, video and digital imaging. These federal agencies include the National Archives and Records Administration, Library of Congress, the Government Printing Office and other federal libraries. The set of guideline includes *Technical Guidelines for Digitizing Cultural Heritage Materials* and *Embedded Metadata in TIFF Images*. The *Technical Guidelines for Digitizing Cultural Heritage Materials* draws substantially on the National Archives' *Technical Guidelines for Digitizing Archival Records* listed above.
- 2013: National Archive of Australia has scanning specifications, available at: www.naa.gov.au/Images/ScanSpecsAmended22082013_tcm16-70095.pdf. All the requirements are similar or the same as the US National Archives guidelines, except that it also recommends using PDF/A as a preferred file format.

1.2 Analysis of file formats in digitization guidelines

Currently all US digitization guidelines including the Federal agencies' technical guidelines for digitization favor TIFF 6.0 as "Preferred format for production master file" and JPEG2000 as "Increasingly considered as a viable format for master image files, but not yet widely adopted." PDF is listed as "Not recommended for production master files," while PDF/A section is empty as it was considered along with PDF (Federal Agency Digitization Guidelines Initiative (FADGI), 2010). FAGI adopted quite a lot from the US National Archives and Records Administration (2004) guidelines. Several PDF standards have been published since 2010, which provide new and better ways to digitize documents. Moreover, in the Federal Agency Digitization Guidelines Initiative (FADGI) format document there are a few technical discrepancies referring to PDF and PDF/A. For example, PDF 1.4 before does not support JPEG2000 compression, and PDF/A has multiple versions (Appendix).

The scanning guidelines by the National Archive of Australia recommend using PDF/A besides recommend using TIFF and JPEG2000 as the preferred format. However, several important things are missing in this guideline. First, no specific guidance on PDF/A standards: PDF/A has three standards. Each one has its own features. There are multiple ways of coding raster images. In other words, raster images can be coded with no compression, lossless compression or lossy compression. As a result, technically raster images in PDF/A can be stored in no compression, lossless or lossy mode. To ensure high-quality images, an institution shall adopt a policy regarding how to handle raster

images in PDF/A. Second, no metadata coding instructions. The guideline does not have recommendations on how to code various metadata information.

The current preferred file formats (the latest FADGI, 2010 guideline and the latest US National Archive and Records Administration's, 2004 guideline) results in more management overhead and higher costs in operation, file management and long-term preservation. In the past several years, there has been new developments in the international standardization of file format and metadata. In this paper, the author proposes a different file format PDF/A over the current preferred TIFF 6.0 or JPEG2000 for textual document digitization. In addition, PDF/A can be used for digitization of other materials such as graphic illustrations, maps and aerial photographs. Furthermore, beyond simply as a file format, PDF/A can be used as open archival information system (OAIS) submission information package (SIP) containers. This international open standards of PDF/A simplify digitization process, reduce digitization cost, improve production substantially and build more confidence for preservation and access. The next section will discuss why and how.

1.3 Overview of OAIS model and digitization

OAIS, ISO 14721:2012 (previously ISO 14721:2003), is well known in library and archive communities as it is the de facto standard for producing, processing, archiving and delivering information from producers to consumers. An OAIS is "an archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community" (CCSDS, 2012). It provides frameworks to ingest, preserve and provide access to facilitate information flow from producers to consumers. In addition, the OAIS defines an information model in which an OAIS information package (IP) shall include: content information; preservation description information; packaging information; and descriptive information. Three types of IPs are defined: SIP, archival information package (AIP) and dissemination information package (DIP). The following figure is adopted from the OAIS Magenta book, illustrating the OAIS IP and data flow (Figure 1).

A producer produces SIPs through digitization process. After that, one or more SIPs are transformed into AIPs during internal management and finally one or more DIPs will be delivered to consumer. Section 3 of this paper discusses a real example of SIPs in digitization and how PDF/A can be a SIP container.

2. PDF standards and uses of PDF/A

2.1 PDF/A standard

First, let's review the development of PDF as international open standards. The PDF 1.7 specification was released as "the full-function PDF" was released under ISO 32000-1 in 2008. Other subset standards were released as ISO standards for "more specialized uses" (Adobe Systems, 2008). For example, PDF/X (ISO 15930) for electronic printing; PDF/A (ISO 19005) for archiving of digital documents. The primary three goals of PDF/A are to:

- "provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems used for creating, storing or rendering the files."
- "provide a framework for recording the context and history of electronic documents in metadata within conforming files."

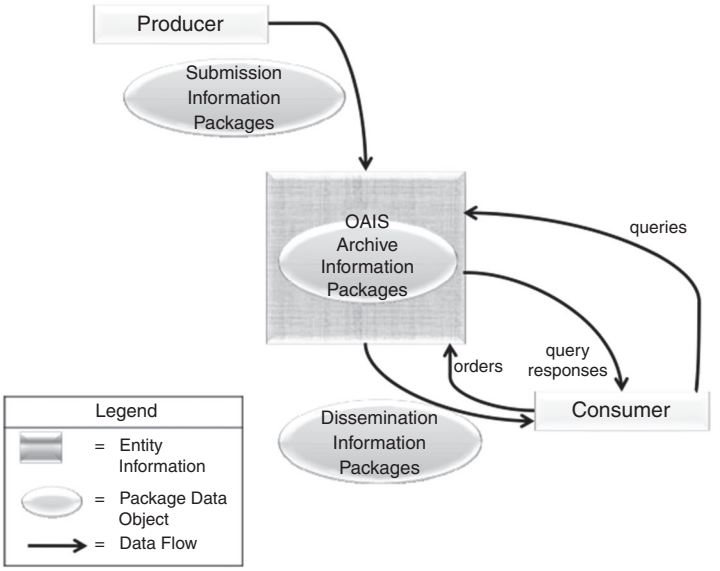

Figure 1.
OAIS information packages and data flow

Source: Adopted from CCSDS (2012)

- "define a framework for representing the logical structure and other semantic information of electronic documents within conforming files" (ISO, 2005, 2011, 2012).

To achieve the above goals, PDF/A has additional requirements and also prohibits some PDF features such as encryption. For more details, please consult the PDF association (www.pdfa.org/) publications on PDF/A topics. Currently PDF/A standards have three parts: ISO 19005-1 (PDF/A-1), ISO 19005-2 (PDF/A-2) and ISO 19005-3 (PDF/A-3). Any of them can be used for long-term archival purpose. PDF/A-1 is based on PDF 1.4; while PDF/A-2 and PDF/A-3 are based on ISO 32000 (PDF 1.7). The naming of PDF/A may confuse people, as it does not mean that PDF/A-2 is better than PDF/A-1 or PDF/A-1 is obsolete. Specially for digitization purpose, one of the most important features in PDF/A-2 and PDF/A-3 is that JPEG2000 compression is supported. This means that there will be 40-60 percent space saving raster images using lossless JPEG2000 compression with PDF/A-2 and/or PDF/A-3 comparing to that of PDF/A-1. The three standards have different features and they co-exist. PDF/A-1was the first PDF/A standard published in 2005, while PDF/A-2 and PDF/A-3 were published in 2010 and in 2012, respectively. The major difference in PDF/A-2 and PDF/A-3 is that PDF/A-3 can embed any arbitrary file or data, while PDF/A-2 can only embed any PDF/A files. This feature of PDF/A-3 makes it be the universal file container like BagIt. The British Digital Preservation Coalition reported that PDF/A is one of the best file formats to preserve electronic documents and suggested using PDF/A as the standard format for archiving electronic documents (Fanning, 2008).

2.2 Uses of PDF/A in born-digital documents

PDF/A has been widely accepted as a preferred master file format for born-digital documents as a SIP container. In some cases, PDF/A is even used as an AIP container. Governments and courts such as European Union and the US Federal Courts (2015) (Borstein, 2010), the National Information Standards Organization (NISO, 2007),

national libraries, private sectors such as banks and hospitals, library centers like California Digital Library (2011) and Florida Center for Library Automation (Chou, 2006) have endorsed PDF/A as the required or preferred file format for born-digital documents over other formats such as Word and PDF. Some case studies have been published regarding the file format. For example, Florida Digital Archive conducted a study to evaluate software conversion from PDF to PDF/A-1 (Koo and Chou, 2013). Archaeology data service uses PDF/A as AIP and analyzed the benefits and drawbacks (Evans and Moore, 2014).

2.3 Use of PDF/A in digitization

PDF/A has been acknowledged to be the preferred master file format for born-digital materials, but it has not been recognized as a preferred master file format for digitization of textual documents. This paper is intended to present the benefit of using PDF/A for digitization.

Only few articles have been published about using PDF/A in digitization since the release of PDF/A-1 in 2005. It appears that all the cases used PDF/A-1 as the file format, and none have discussed using PDF/A-2 or PDF/A-3. The Ohio State University Libraries mentioned to investigating of using PDF/A-1 for digitization (Noonan *et al.*, 2010). They did not report if there is a policy on how to handle the raster image. Most likely, the raster images inside the PDF/A-1 files were saved in lossy compression mode. If this is the case, the digitization quality of digitization files is lower. In comparison, PDF/A-2 provides a better way to handle raster images due to its JPEG2000 compression feature. In addition, the paper did not report how to code metadata info such as technical and descriptive metadata. Optical character recognition (OCR) can be handled differently to achieve better results. Following common digitization practice, South New Hampshire University digitized papers as TIFFs, and made PDFs for access files. Then the PDFs were saved as PDF/A for improving access files longevity and accessibility (Platt, 2010). The use of PDF/A in this case is questionable, as raster images in PDFs do not rely on embedded-font and there is no benefit needed for longevity of access file. India National Agricultural Research System digitized documents using PDF/A. Unfortunately the paper did not discuss which PDF/A standard was used, and did not mention its choice of compression policy and metadata information (Veeranjaneyulu, 2014). In summary, all the reported case studies did not utilize the full potentials of PDF/A.

3. Digitization and SIP

In the OAIS model the producer creates SIPs, which may be in any format that the producer and the archives agree to. An OAIS IP shall include: content information; preservation description information; packaging information; and descriptive information. In the current digitization process, a typical SIP consists of a corresponding directory containing the following information:

- Content: preservation master files – raster images files saved in preferred file format as TIFFs/JPEG2000s) as each page of textual documents is scanned as a raster image. Access files – a compressed PDF and/or the same number of JPEGs/JPEG2000s. Other content such as OCR data.
- Preservation description: preservation metadata saved in TIFF header and other metadata such as structural and technical metadata; checksum files.
- Packaging information: directory and File naming, structural metadata.

- Descriptive information: descriptive metadata saved in digitization management system, catalog or textual/XML files.

For example, A book of 100 pages typically consists of the following files: 100 TIFFs/JPEG2000s as master files, one PDF file consisting of compressed images for access, one checksum file consisting of all files' checksums, one structural metadata file consisting of structural metadata information for this book, OCR data either saved in the PDF file or a separate text file or an ALTO XML file (technical metadata for OCR). The SIP information is spread out from different parts of a file or in multiple files. A TIFF image contains the content (the visual appearance of a physical page from a book), it also contains some of the preservation information in its header tags. However, other content information such as OCRed text from this TIFF image is most likely saved in a separate file. As a result, to gather all the SIP information, ingestion must interact with multiple files and even database. It is error-prone when dealing with a mix of multiple files in a variety of file format. Here are the examples.

(1) Directory listing

Directory Identifier	Page #s
azu_acku	
	00000001.tif
	00000002.tif
	00000003.tif
	00000004.tif
	00000005.tif
	…
	azu_acku_1.pdf
	stru_meta.txt
	checkmd5.txt

(2) Checksum file
The checksum file consists of checksums of all the master image files, whose main purpose is to ensure data integrity for error-detection during the process of data transmission and/or storage.

(3) Structural metadata
Structural metadata describes the logical structure and components of content. This type of metadata can be used for both page-to-page and semantics navigation in delivering digitization materials to enhance users' access experience. Yale University Library (2008) published a detailed policy on how to use structural metadata in multiple levels. A simple example is a table of content for a book. More examples of structural metadata can be found in Yale University Library's best practices for structure metadata. The University of Arizona Libraries has a similar practice with the following options:

- No structural metadata.
- Structural metadata defining file sequence: Yale University Library, Cornell and University of Michigan have used this approach. The UAL uses a modified file naming convention to code file sequence as part of the file name. TIFF files are specifically named with leading zeros to facilitate sorting.

- Structural metadata defining logical components: This level is typically for books and manuscripts. Some examples include title pages, chapters and indices. The UAL has been using coding of this type for some of the digitized materials. The information can be saved in a text/XML or database or in METS.

(4) Technical metadata
Technical metadata describes technical attributes of digital objects during digital capture and other processes. It typically comes from digitization equipment such as scanners and digital cameras. Examples are hardware, software to produce the digital object, resolutions, file formats and color profiles. Library common practice is to code this metadata in a separate text/XML file or in METS.

(5) Descriptive metadata
Descriptive metadata is the most commonly used to describe a resource for identification and discovery. This type of metadata is the most widely used for resource description and discovery via search engines and local search functions. Typical elements are title, author and keywords. Typically the descriptive metadata is saved as a separate MARC/MODS/METS/Dublin Core file.

(6) Other metadata such as preservation metadata
Other type of metadata such as preservation and rights metadata might be added in the digitization process, or might be updated at a later stage.

(7) Other data
Other data can be embedded in PDF/A files. For example, OCR data are delivered separately in another file such as text file or ALTO technical metadata for OCR or within a PDF.

4. PDF/A as an OAIS SIP container

The key requirement of PDF/A is that it is self-described and self-contained so that it can be reproduced exactly the same way with different software in various platforms. All of the information necessary for displaying the document is embedded in the PDF/A file. This includes any content such as text, raster images and vector graphics, fonts and color profiles. For digitization, PDF/A can be used as a structured, self-contained and self-described data container, which codes raster images in uncompressed, lossless and/or lossy compressed mode depending on the users' preference. The PDF/A file format can achieve structured, self-contained and self-described status by doing the following:

(1) tagged PDF: embed structural metadata via pre-defined PDF tags or create your own tags;

(2) self-contained: embed required color profiles, fonts and other related information; and

(3) self-described using extensible metadata platform (XMP) metadata: PDF/A can code all the required information from an OAIS SIP through the standard and XMP.

All the files and data from the above digitization SIP can be coded in PDF/A. How to do so:

(1) All the raster images (page 1, page 2, etc.) from a book can be compressed and saved in desired sequence in one PDF/A file with JPEG2000 lossless compression.

(2) All the metadata can be coded with another ISO metadata standard XMP (ISO 16684):

- Structural metadata can be coded with XMP inside of the PDF/A file metadata stream with pre-defined tags. If users need to have their own customized tags, they can do so.
- Descriptive metadata can be coded with XMP standardized Dublin Core elements such as dc:contributor and dc:title. If users need to code other descriptive metadata elements such as MODS, they can achieve this via XMP extension.
- Rights and media management metadata can use XMP standard namespaces. XMP also have camera raw metadata namespace if needed.
- OCR data in ALTO format can be coded within XMP using extension. or OCR text can be stored in PDF.
- Other metadata such as METS and MODS can be coded using XMP extension.

(3) Any arbitrary files can be saved within a PDF/A-3 file.

4.1 Criteria for master file formats

When choosing a file format for digitization, future viability of the master file format is the primary factor. Therefore, common considerations include non-proprietary, open and documented international standards, commonly used, unencrypted, uncompressed or lossless compression. Digital preservation is running on a stack of hardware and software. Rendering any file relies on appropriate hardware, operating systems, libraries and application software. While some experts may argue the compression is the preferred choice, the author believe that lossless compressed file using non-proprietary and/or open documented algorithms is equivalently as good as uncompressed file.

4.2 TIFF issues as the preferred master file format

For the past 20 years TIFF 6.0 has been the preferred master file format for digitization due to a few factors such as availability of the technical specification and easy-to-understand file structure. TIFF does have certain advantages as a file format for raster image only materials, as the baseline TIFF is very simple, easy to repair and migrate, as it cannot include layers and JPEG or LZW compressions. However, there are several significant issues with TIFF 6.0 file format and implementation when dealing with textual materials:

- Just a raster image format: in the OAIS model, the SIP is generated by information producers, and the IP is ready for ingestion for management. Due to limitations of the TIFF file format, other ways of managing and handling SIP are required. Although TIFF supports coding multiple images and XMP metadata, current practices limit the uses. One common way is to have all the related data (e.g. image files, structural information) saved in a directory. Another way is to maintain a digitization management system which capture the SIP information. Unfortunately all these ways result in huge increase in costs along with inefficiency.
- Proprietary standard: many people perceive that TIFF (aka TIFF 6.0) is an open standard. They are partially correct. Adobe still holds the copyright on the TIFF 6.0

specification, although Adobe does not require a license to implement TIFF software. A license was required at one time to implement the LZW compression algorithm, but all patents on that have been now expired. TIFF-EP (electronic photography), as a subset of TIFF 6.0, is an ISO standard. Although there are no major difference from TIFF 6.0, many of TIFF 6.0 tags are ignored in TIFF/EP.

- Big file size: many institutions choose uncompressed TIFF 6.0 as the master file format. Due to the nature of TIFF, file size is huge compared to lossless compressed one. For example, compressed TIFF with lossless LZW is 30 percent +smaller in file size. In comparison, lossless compressed JPEG2000 file is 40-60 percent+ smaller in file size.
- Inflexible for web and mobile delivery: in the web or mobile environment, access is critical, while TIFF cannot be viewed directly in browsers (except Safari) or mobile phones without a plug-in. Appropriate software is required to open the file. Along with huge size, delivery of digitized textual documents in TIFF has to be converted to some other file format such as PDF and JPEG for access and faster download.
- Indexing is difficult: indexing the content in a TIFF file generally cannot be saved with the file itself, and has to be achieved via other ways. For example, OCR data saved in textual file, PDF or, XML-format such as ALTO.
- Structural metadata not allowed: TIFF does not provide a way to capture structural metadata, which is critical for providing access to digitized manuscripts and journals. To achieving this feature, libraries have been using a database or an additional metadata wrapper such as METS to save structural metadata information.
- Inconsistent TIFF tag data: one of the oldest digitization projects from the Library of Congress, American Memory has used TIFF 5.0 and TIFF 6.0 as the master file format. It is also noted that "the Library's use of TIFF formats and headers has not always gone smoothly, perhaps the inevitable result of using a "multi-flavor" set of industry conventions rather than a true standard" (Fleischhauer, 1998). It is so true that we have seen various meta tags in TIFF header from different digitization vendors. In addition, in textual document digitization process, almost all of TIFF tags such as scanner and dimension are the same. The set of these TIFF tags are stored in each TIFF file header, resulting many duplicates.
- TIFF tags are difficult to work with: FADGI also points out that the proliferation of tags and tag sets complicates TIFF metadata extraction. Most TIFF programs only display TIFF's baseline, extension and a few private tags. In addition, the extracted data are difficult to use and store because of the different data types for the various tagged fields, and the lack of any systematic data structures and formats (Federal Agencies Digitization Guidelines Initiative (FADGI), 2009).

JPEG2000 file format has many advantages over TIFF 6.0, but also have a few drawbacks. A study conducted by Mr Knijff at the National Library of the Netherlands found out the limitations of JP2 handing ICC profiles and different handling of headers via major JPEG2000 software (van der Knijff, 2011). This finding led to a request to amend JPEG2000 standard. For raster images, JPEG2000 and/or TIFF will still play an important roles as a preferred master format. However, their roles as a preferred master format for textual documents are not justified.

4.3 PDF/A as an OAIS SIP container

The author recommends using PDF/A as the preferred file format for production master file for textual document digitization. The author further recommends to use:

(1) PDF/A with lossless compressions as the preferred file format for high-quality digitization. In addition, PDF/A can be used as an OAIS SIP container beyond merely used as a file format.

(2) PDF/A with lossless compressions along with metadata/data as the preferred OAIS SIP container.

A PDF/A file can be structured, self-contained and self-described for digitization by coding various metadata and roaster images within the file. PDF/A is a better file format than TIFF/JPEG200 for online access and delivery, as it requires no browser plug-in in web and/or mobile environment; can be tagged PDF with structural information. PDF/A offers the following advantages.

4.3.1 Open International Standards. PDF are now truly open documented international standards. ISO has been releasing multiple standards-related PDF since 2008. It was true that PDF was a proprietary format controlled by Adobe before 2008. ISO published PDF 1.7 as ISO 32000-1:2008, and since then the control of the PDF specification passed to an ISO committee of volunteer industry experts. In 2008, Adobe published a Public Patent License to ISO 32000-1 granting royalty-free rights for all patents owned by Adobe that are necessary to make, use, sell and distribute PDF compliant implementations. PDF/A-1, PDF/A-2 and PDF/A-3 are all ISO standards under ISO 19005. While many users are used to using Adobe products to handle PDFs, there are a few open source software and private companies providing ways to generate and update PDFs. In the worst case scenario, you can write your own software to handle PDFs based on ISO technical specifications.

4.3.2 Self-contained and self-described. PDF/A is suitable as an OAIS SIP container to code all the required digital objects, data and/or metadata. It can package all image objects along with ICC profiles into one PDF/A file instead of a directory of TIFFs/JPEG2000s. Limitations of TIFF 6.0 and JPEG2000 were identified above. For example, the PDF 1.7 specification section 4.5 "color spaces," 4.8 "Images" and 10.7 "Tagged PDF" explain how to handle ICC profiles, raster images and structural metadata. In comparison, in the current practice a digitization producer has to use a package of TIFFs/JPEG2000s and associated files to deliver data. As a result, several benefits will be achieved for both producer and receiving institutions. Using PDF/A simplifies ingestion and delivery process. Reduction in the number of files for AIPs results in less management overhead. The flexibility of XMP makes it easier to code standardized metadata such as Dublin Core and specialized metadata for AIPs and DIPs. As a result, the use of PDF/A file can eliminate current digitization inventory and/or management system.

4.3.3 Flexible. PDF/A offers options to encode raster images either in uncompressed mode or lossless or lossy way with royalty-free compression algorithms. It can use all the compressed methods offered in TIFF, and provide an improved compression offered in JPEG2000 for PDF 1.5 and above. These options are very flexible to enable users to handle master or access files in their preferred ways. The options are:

- uncompressed: one can code uncompressed images in PDF/A-1;

- lossless compression: one can code raster images with lossless compression (e.g. CCITTFaxDecode, FlateDecode) in PDF/A-1 (e.g. CCITTFaxDecode, FlateDecode, JPXDecode) in PDF/A-2 or PDF/A-3; and
- lossy compression: one can code raster images with lossy compression (e.g. JPXDecode) in PDF/A-2 or PDF/A-3.

4.3.4 Space saving. PDF/A-1 supports typical lossless compression algorithm in TIFF, including CCITTFaxDecode, JBIG2Decode and LZWDecode. In addition, PDF/A-2 and PDF/A-3 support improved compression using JPEG2000 (called JPXDecode). By doing so, the file size of a PDF/A containing a JPEG2000 image is almost the same as that of the JPEG2000 file, which has all the advantages of JPEG2000 over TIFF 6.0. These are:

- open standard without royalty or license fee;
- a single approach to lossless and lossy compression; and
- reduced cost for storage and maintenance with smaller file size.

PDF/A-2 can use lossless JPEG2000 compression, which results in almost the same file size as a congregation of the same JPEG2000s. In comparison to TIFFs, file size saving will be 40-60 percent.

4.3.5 Accessibility. PDF/A can be created to be accessible for disabled people. In this case, a logical structure can be established to associate a group of objects include graphic objects and others. The tagging is very similar in concept to markup languages such as HTML or XML. The ISO standards have pre-defined types of structure elements to enable organization of a document into common things such as chapters and tables (Adobe Systems, 2008). This feature is also extensible so that users can define their own structural information. In fact, Tagged PDF is required to be PDF/A-1a and/or PDF/A-2a.

4.3.6 Better metadata support with XMP. XMP was original created by Adobe, which is another ISO standard (ISO 16884-1) published in 2012. It is a data model in a serialization format for creation, processing and interchange of standardized and customized metadata for digital objects. Due to the open standard and its powerful data model, XMP has been widely used in the IT industry and can be coded in multiple file formats including TIFF, JPEG, JPEG2000, PDF, MP4, AVI and HTML. The details of use and analysis of XMP is beyond the scope of the paper. Readers shall consult ISO standards and other related literature for implementation.

FADGI's response to TIFF metadata issues is to use XMP to standardize TIFF metadata and store the entire metadata package in TIFF (FADGI, 2009). The XMP specification includes over a dozen pre-defined metadata schema, but it is also extensible by nature. In other words, industry-specific or user-specific metadata tags can be encoded in XMP. For example, typical library metadata standards such as METS and MODS can be encoded.

4.3.7 Other files or data. PDF/A-3 has the ability to have any file or data encoded. ISO 32000-1 specifically states "PDF provides means for applications to store their own private information in a PDF file. This information can be recovered when the file is imported by the same application, but it is ignored by other applications" (Adobe, PDF 1.7, p. 42). This raises some concerns about how/what to handle attached file or data. However, the author believes this feature provide many opportunities to code research data and digital publishing as long as the data are self-describing self-

containing. For example, institutions can embed other metadata in PDF/A-3 such as Marc Records and METS. Institutions can make their own XML data and schema and embed such data.

4.3.8 Emerging file formats and widely adoption. PDF/A is gaining widely used in US and European countries. The most notable uses have been discussed above. Due to its popularity, there are open source libraries and private software companies offering software to produce and manage PDFs.

4.3.9 Lower total cost of ownership (TCO). Using PDF/A as a SIP container will result in much lower TCO due to the following reasons:

- The use of lossless compression result in over 40-60 percent space saving in comparison with using the current preferred approach of TIFF without compression.
- The one-to-one relationship of a physical textual document and a PDF/A file eliminates the needs of managing hundreds of master image TIFFs/JPEG2000s and associated metadata. In current digitization workflow, each TIFF/JPEG2000 file will have to be managed and preserved. File management with TIFF/JPEG2000 includes: each file must have a unique file name; each file must be preserved (possible work includes generating checksums, verifying headers, saving multiple copies); and each file must be associated with some structural information to maintain its relations with others. All workflow can be eliminated using PDF/A. Institutions will no longer need to manage and preserve these files. Instead, they only manage and preserve one PDF/A file.
- A PDF/A file can be tagged to contain structural metadata, eliminating the current ways of storing, managing and displaying structural metadata.
- One file multiple uses: a master PDF/A file can be easily repurposed for other uses such as enhanced access via computer-assisted OCR and document analysis. For example, the master PDF/A file can be further compressed to be an access file as well. Various application programming interface can be designed in AIPs and DIPs to deliver customized access using PDF libraries.
- Better metadata with XMP: administrative, structural, copyright, technical or whatever metadata can be coded in XMP, and then stored inside of a PDF/A. This eliminates the needs for maintaining additional documents or database. In addition, these metadata can be easily integrated with any repository system or other preservation or access services.
- Data such as OCR and research data sets can be stored within a PDF/A file if needed with specific handling information.

In certain cases, derivatives of a master PDF/A file can be further served as AIPs and DIPs. For institutions who have legacy TIFFs and JPEG2000 from previous digitization projects, it is not difficult to migrate digitized titles from these directories of TIFFs/JPEG2000 along with associated data to PDF/A files. There are open source and proprietary software which can produce PDF/A files through a batch process. On the other hand, it is also easy to extract images, metadata and data from a PDF/A file. For archival management, a user needs to get raster images from a PDF/A file in original quality. The images can be saved in JPEG2000 or TIFF without losing any quality. Similar operations can be performed for encoded data/metadata in the PDF/A file. For delivery

purposes, the user can further compress the master PDF/A files to generate an access PDF as a DIP. The availability of open source or proprietary software make it straightforward to work on PDF/A as you wish.

4.4 Issues

There are several issues with PDF/A naming and implementation. One issue is that the standards and naming of PDF/A-1, PDF/A-2 and PDF/A-3 confused certain people, as they may think PDF/A-3 is a better and advanced than the other two. In addition, the conformance levels add more education needs for PDF/A, as there are three-level: a stands for accessible, b for basic, and u for unicode. That is, PDF/A-1a, PDF/A-1b, PDF/A-2a, PDF/A-2b, PDF/A-2u, PDF/A-3a, PDF/A-3b and PDF/A-3u.

The most critical need is reliable open source software for producing and validating PDF/A files. The author observed and some people also reported that some PDF/A files produced by scanners and software did not pass the PDF/A compliance test. Commonly used file validation tool JHOVE/JHOVE2 do not work well for validating PDF 1.7 and PDF/A. In fact, this issue is somehow true for TIFF and JPEG2000 as some software did not meet 100 percent of the original file specifications. The software issue is not related to format specification, but depends on quality of software developers and testers. The PDF association maintained a web page covering PDF/A validators (available at: www.pdfa.org/products/?c=1988&cn=645). A recent positive movement is that the PDF association formed the PDF validation technical working group to work with the Open Preservation Foundation for an open source project VeraPDF for validating PDF/A.

5. Summary

TIFF file format has been the preferred master file format by FADGI digitization guidelines for the past 20 years. However, there are many drawbacks of this practice. The author first reviews the current digitization guidelines, the OAIS model and provides on an overview of the development PDF and PDF/A as international standards since 2005. A literature review shows that PDF/A has been widely accepted as the preferred master file format for born-digital documents, but it has not been recommended for digitization. The author analyzes drawbacks of TIFFs and JPEG2000s as the preferred file format. The author shows how PDF/A is a better file format than current preferred TIFF and JPEG2000. Furthermore, the author shows that various metadata can be coded in XMP and arbitrary file/data can be coded in PDF/A-3. These features in PDF/A provide much better ways to deliver SIPs in a cost-efficient manner. As a result, the author recommends that PDF/A with lossless compressions as the preferred file format, and PDF/A with lossless compressions along with metadata/data as the preferred OAIS SIP container.

References

Adobe Systems (2008), "ISO 32000-1:2008, 14.8. PDF 32000-1:2008, 1st ed., document management – portable document format – Part 1: PDF 1.7", available at: www.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/PDF32000_2008.pdf (accessed May 28, 2015).

Borstein, R. (2010), "Federal Courts moving to requiring PDF/A for filings", available at: http://blogs.adobe.com/acrolaw/2010/10/federal-courts-moving-to-requiring-pdfa-for-filings/ (accessed June 20, 2015).

California Digital Library (2011), "CDL digital file format recommendations: master production files", available at: www.cdlib.org/gateways/docs/cdl_dffr.pdf (accessed June 20, 2015).

Chou, C. (2006), *Guidelines for Creating Archival Quality PDF Files*, Florida Center for Library Automation, available at: www.fcla.edu/digitalArchive/pdfs/PDFGuideline.pdf (accessed June 20, 2015).

Evans, T. and Moore, R. (2014), "The use of PDF/A in digital archives: a case study from archaeology", *International Journal of Digital Curation*, Vol. 9 No. 2, pp. 123-138.

Fanning, B.A. (2008), "Technology watch report preserving the data explosion: using PDF", Report No. 08-02, DPC Technology Watch Series, available at: www.dpconline.org/docs/reports/dpctw08-02.pdf (accessed June 20, 2015).

Federal Agencies Digitization Guidelines Initiative (FADGI) (2009), *Guidelines for TIFF Metadata Recommended Elements and Format Version 1.0.*, FADGI, available at: www.digitizationguidelines.gov/guidelines/TIFF_Metadata_Final.pdf (accessed June 20, 2015).

Federal Agencies Digitization Guidelines Initiative (FADGI) (2010), *Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files*, FADGI, available at: www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf (accessed June 20, 2015).

Fleischhauer, C. (1998), "Digital formats for content reproductions", available at: http://memory.loc.gov/ammem/formats.html (accessed May 20, 2015).

ISO (2005), "ISO 19005-1: 2005 Document management – electronic document file format for long-term preservation – Part 1: use of PDf 1.4 (PDF/A-1)".

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. and Chou, C.C.H. (2013), "PDF to PDF/A: evaluation of converter software for implementation in digital repository workflow", *New Review of Information Networking*, Vol. 18 No. 1, pp. 1-15, available at: doi:10.1080/13614576.2013.771989

NISO (2007), *A Framework of Guidance for Building Good Digital Collections*, 3rd ed., NISO, available at: www.niso.org/publications/rp/framework3.pdf (accessed June 20, 2015).

Noonan, D., Wl, A., McCrory, A. and Black, E. (2010), "PDF/A: a viable addition to the preservation toolkit", *D-Lib*, Vol. 16 Nos 11-12, doi:10.1045/november2010-noonan.

Platt, A. (2010), "Developing an institutional repository at southern new hampshire university: year one", in Ng, K.B. and Kucsma, J. (Eds), *Digitization in the Real World: Lessons Learned from Small and Medium-sized Digitization Projects*, Metropolitan New York Library Council, New York, NY, pp. 261-273, available at: http://metro.org/media/files/None/DITRW_16.pdf (ccessed June 20, 2015).

The Consultative Committee for Space Data Systems (2012), "Reference model for an open archival information system (OAIS) recommended practice CCSDS 650.0-M-2".

US Federal Courts (2015), "Case management and electronic case filing (CM/ECF) to transition to PDF/A", available at: www.pacer.gov/announcements/general/pdfa.html (accessed June 20, 2015).

US National Archives and Records Administration (2004), *Technical Guidelines for Digitizing Archival Records for Electronic Access: Creation of Production Master Files – Raster Images*, US National Archives and Records Administration, available at: www.archives.gov/preservation/technical/guidelines.pdf (accessed June 20, 2015).

Van der Knijff, J. (2011), "JPEG 2000 for long-term preservation: JP2 as a preservation format", *D-Lib*, Vol. 17 Nos 5/6, available at: www.dlib.org/dlib/may11/vanderknijff/05vanderknijff.html (accessed July 6, 2015).

Veeranjaneyulu, K. (2014), "KrishiKosh: an institutional repository of national agricultural research system in India", *Library Management*, Vol. 35 Nos 4/5, pp. 345-354, available at: doi:10.1108/LM-08-2013-0083

Yale University Library (2008), *Best Practices for Structural Metadata, Ver 1*, Yale University Library, available at: www.library.yale.edu/dpip/bestpractices/BestPracticesForStructuralMetadata.pdf (accessed June 20, 2015).

Further reading

Federal Agencies Digitization Guidelines Initiative (2014), *Raster Still Images for Digitization: A Comparison of File Formats: Part 2. Detailed Matrix (Multi-Page) (PDF)*, Federal Agencies Digitization Guidelines Initiative, September 2, available at: www.digitizationguidelines.gov/guidelines/FADGI_RasterFormatCompare_p2_20140902_r.pdf (accessed May 20, 2015).

Appendix. Corrections and comments on raster still images for digitization: a comparison of file formats: part 2. detailed matrix (multi-page) (PDF) September 2, 2014

(Note: The author appreciates Leonard Rosenthol, ISO PDF working groups leader and Adobe Architect comments)

- Page 11: Attribute: Sustainability Factors: Geo-referencing Metadata: PDF 1.7 supports Geospatial tagging (NOT "Not supported").
- Page 12: Sustainability Factors: Level of Effort to Embed Geo-referencing Metadata: PDF 1.1-1.7 shall be "Low" (NOT "N/A").
- Page 14: Sustainability Factors: Technical Protection Mechanisms: PDF/A-1, PDF/A-2 do not allow protection and security (NOT "No Impact (protection mechanisms are available but not required and not a deterrent from choosing this format)").
- Page 15: Cost Factors: Implementation Cost: Creation of PDF (1.1-1.7) and PDF/A shall be "Low-Medium", as there are a few open source libraries including iText and PDFBox (NOT "Medium-high (tools can be expensive").
- Page 16: Cost Factors: Cost of Software Tools: Creation of PDF (1.1-1.7) and PDF/A shall be "Low-Medium", as there are a few open source libraries including iText and PDFBox (NOT "Medium-high (tools can be expensive").
- Page 28: System Implementation Factors: Ease and accuracy of File validation: Unfortunately JHOVE/JHOVE2 do not work well for validating PDF 1.7 and PDF/A-2. veraPDF is an open source PDF/A validator built by the Open Preservation Foundation and The PDF association.
- Page 34: Settings and Capabilities: Notes on Maximum File Size: PDF 1.0 - 1.4 and PDF/A-1 is 2GB, while PDF 1.5-1.7, PDF/A-2 and PDF/A-3 support 10 GB. (NOT "Generally accepted practical limit is 2GB").

Corresponding author
Dr Yan Han can be contacted at: yhan@email.arizona.edu