PDF/A: ISO Standard for Long-Term Archiving

Klaus Jung and Thomas Zellmann, LuraTech Imaging GmbH, Berlin, Germany

Abstract

The authors review the requirements of long-term archiving that led into the initiative to create a standard for electronic document archiving. The current status of the PDF/A standardization activity is given. The standard itself, its goals, and its usability are described in detail. The differences between traditional PDF and PDF/A are explained. Existing solutions and a potential validation process are discussed. Special emphasis is placed on how PDF/A can be used to create full-text searchable PDF files from scanned documents, and on how such documents can be created in a compact and easily accessible way (e.g. via internet). This includes the embedding of OCR results, and the compression of scanned page images using MRC (Mixed Raster Content) techniques. Certain application areas are listed, ranging from processing scanned paper documents to workflows that deal with digital born documents. An outlook is given on future standards based on PDF, like ISO 32000 and PDF/A-2.

Introduction

Our world is getting more digitally every day. A lot of information and documents only exist in digital form today, but will they still be legible "tomorrow"? That was the theme of an interesting TV show appropriately called "The Digital Disaster". It began with cave drawings from the Stone Age and papyrus rolls from ancient Egypt, both of which have survived as documents for thousands of years. What documents from the 21st century will future generations be able to find and still read?

But it's happening much quicker than you may realize. Just think of a 5¼ inch floppy disk and it demonstrates a lot of the problems of long-term archiving. It begins with the hardware: Where can you buy a 5¼ inch floppy disk drive today? And even if you find one, there's a good chance that the disk is physically damaged. If these two hardware hurdles are successfully cleared, then what kind of software or document will you find on the floppy disk? Are the appropriate viewing and processing programs still available? And this example is a mere 15 years old!

Requirements

This short introduction leads us to the demand the long-term archiving of documents. Electronic archiving is critical for businesses and organizations, because documents today often only exist in digital format. The length of time that documents have to be archived varies from sector to sector and country to country, but some examples can help us to get an idea. Federal laws often require an archiving period of around 10 years. Banks and insurances demand that customer dossiers are retained for more than 50 years. In the engineering branch, archival periods of 100 years are common for aircraft; bridges hopefully hold a whole lot longer. In the area of archives and libraries long-term typically translates into “forever”!

And saving documents in proprietary formats for this length of time is really not a good idea. This leads to the second problem with the digital document world - that many users already have a real “format zoo”, which can quickly become unmanageable (if it isn't already so). Proprietary document formats have to be migrated on a regular basis, in order that newer versions of the processing software can still read them.

Employees working with documents aren't really impressed when 10 different viewing programs are opened up at the same time. In some of the programs they might not even know how to navigate around in a document. In order to solve this problem, a document and archiving format is needed that guarantees the required long-term archiving period and offers the option of a single format type. This is where PDF/A as an ISO standard for long-term archiving enters the stage.

Based on the popular and widely distributed PDF format, more modern and providing more functionality than the TIFF format that is commonly used for archiving purposes, PDF/A has a lot to offer.

Different organizations and archives often define their own policies for long-term archiving. Most commonly they include the following fundamental requirements:

- Authenticity: The quality of being genuine, not a counterfeit, and free from tampering.
- Integrity: The quality of being whole and unaltered through loss, tampering, or corruption.
- Correctness of the links to descriptions and the origin of the archived material.
- Assurance that the chain of custody can be tracked, and that the information content remains unchanged.
- Independence from infrastructure: Not to use any proprietary processes or format that could prevent migration of the archived material to another choice of technology.

More requirements on the archival environment for long-term retention of documents can be found in ISO standards ISO 15498-1 [1], ISO/TR 15801 [2], and ISO/TR 18492 [3].

In view of the formats used to store digital objects, the following requirements ideally apply:

- Accessibility: Not using encryption or proprietary formats.
- Independent from platform, operating system, and device: The object should be readable, understood, displayed, and printed on many different hardware platforms, ideally using different alternative software implementations.
- Published specification: Using open, accepted specifications controlled by standards organizations.
- Self-contained: All resources required for display (e.g. fonts) must be included in the object, not using external references.
- Widely distributed: the format should be accepted by both, industry and government.

The PDF/A ISO Standard

The PDF/A standard [4] was, and is continuing to be, developed by the ISO committee ISO/TC 171/SC 2/WG 5. It was specifically created to meet the requirements for a document format used in long-term archiving; the letter "A" stands for "Archive". It envisions a single PDF/A-based archive for all documents in an organization, from input through to output, and includes all of the areas in between.

The goal of PDF/A is to provide a reliable storage format for electronic documents across multiple generations of technology. It is a multi-part standard with several levels of compliance. Businesses, governments, libraries, archives and other institutions and individuals around the world use PDF to represent considerable bodies of important information. The future use of and access to these objects depend upon maintaining their visual appearance as well as their higher-order properties for a substantial length of time.

The primary purpose of PDF/A is to define a file format based on PDF, which 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. A secondary purpose is to provide a framework for recording the context and history of electronic documents in metadata within conforming files, and to define a framework for representing the logical structure and other semantic information in electronic documents.

These goals are accomplished by identifying the set of PDF components that must be used, and another set that is prohibited to be used.

Required use: To ensure full access to all elements belonging to a document. This leads e.g. to the requirement that fonts must be embedded. A link to a font file or just the name of a font is not sufficient, since it cannot be ensured that such a font is available on future computer systems, nor it is ensured that the font's outline will be exactly the same on different systems.

Prohibited use: Some PDF features must be avoided, since they can alter the visual appearance of the document. Such elements include interactive features, optional content (PDF layer switches), and others.

Structure and Status of the PDF/A Standard

PDF/A was published in September 2005 with the number ISO 19005-1:2005 [4]. It is based on the PDF Specification 1.4 [5] published by Adobe Systems in 2001. We refer to it as PDF/A-1.

An additional part, ISO 19005-2 or PDF/A-2, is currently under development. More details are given at the end of this article.

PDF/A-1 defines two levels of conformity. Level A ensures the preservation of a document's logical structure, including the natural reading order of the text content. Text extraction is of particularly importance for devices that need to re-flow the text content due to limited screen size (such devices are not conformant PDF/A viewers), or that need accordance with Section 508 of the US Rehabilitation Act. In addition, Level A demands all the requirements of Level B.

Level B ensures the reproduction of the document without any visual ambiguity. This includes the correctness of colors and the rendering of text. So the human readability is ensured, but the structure and semantics are not described in an independent manner, needed to make it accessible to machine readers.

Comparison between PDF and PDF/A

The following table lists PDF features that are required or forbidden in PDF/A-1 compliant documents.

Restrictions and requirements defined in PDF/A-1

PDF Version	PDF 1.4
Identification	Required PDF/A identifier specifying version and conformance level.
Metadata	Must be XMP-compliant, XMP extensions must embed their descriptions.
Logical structure	Level A: Must be tagged PDF describing structure and semantics. Level B: No requirements.
Encryption	Prohibited. Must be possible to open and render without password.
Color	All colors must be identified, e.g. by ICC color profiles.
Transparency	Prohibited. The visual appearance of transparent objects stacked on top of each other is not always clearly defined.
Layers	Optional content layers are prohibited.
Compression	LZW compression is prohibited. JPEG 2000 compression is prohibited (not part of PDF 1.4).
Fonts	All fonts must be embedded, except they only appear in rendering mode "invisible". The mapping of character codes to glyphs must be unambiguous. Only fonts that are legally embeddable for unlimited, universal rendering shall be used. Level A only: Requires a map to Unicode.
Annotations	Static text/label-style annotations allowed, others e.g. movies or sound are prohibited.
References	References to external images or external page content are prohibited.
Alternate images	Alternates for lower resolution display are prohibited.
Programming	Embedded JavaScript is prohibited.
Actions	Certain actions like opening movies or sending forms are prohibited.
Forms	Restricted. See [4] for details.
3D objects	Prohibited.
Postscript	The embedding of postscript is prohibited.
Digital signatures	A PDF/A-1 conforming way to add digital signatures is possible.

PDF/A Creation

PDF files are created from many different sources. Here we focus on application areas that create PDF/A documents from scanned documents (analog to digital conversion) and from digital document formats (digital born PDFs).

PDF/A from Scanned Documents

For the preservation of cultural heritage the digitalization of paper documents becomes more and more important. But also commercial organizations tend to digitalize their paper documents, which can be older documents, no longer needed to be stored in original paper form, or just all incoming paper mail to implement a completely digitalized workflow.

Here we do not focus on the scanning process itself, what scanners or cameras to use, how to preserve the original colors, etc. Once the scanning is done, we end up with a large amount of image data, one image per page.

Typically a large number of pages need to be processed and stored. Moreover scanned page images require much more memory than pages that just typeset its text using a font. So file size becomes an issue. The size of an uncompressed letter sized page in 300 dpi is 25 MB.

Various compression techniques can be applied to reduce the size of a page image. For the given example, Fax Group 4 compression can reduce the page size to approx. 160 KB, but it loses all the color information. Fax Group 4 can only store bi-tonal (black and white) content. The well-known JPEG compression can reduce the given page to 160 KB as well, and it allows reconstructing a color image. The newer JPEG 2000 compression can even do a better compression on color images [8-9]. But both, JPEG and JPEG 2000 are performing best on photo realistic images. A document page typically consists of color images and text. Applying JPEG or JPEG 2000 at higher compression rates results in a degradation of the image quality in text areas. Text typeset in small font sizes may become hard to read.

Mixed Raster Content (MRC) compression techniques can be used to obtain both, high compression rates with good image quality, and an excellent readability of text [10-13]. For the given example page we receive a size of 55 KB. The MRC technique is based on the segmentation of the page content to distinguish text from image-like areas, and the application of different compression techniques, JBIG2 [14-15] for text and JPEG or JPEG 2000 for image-like areas.

Within PDF/A the use of JBIG2 and JPEG compression, as well as the masking technique shown in Figure 1 are allowed. Once PDF/A-2 is rolled out, JPEG 2000 offering higher compression performance will be available for PDF/A documents as well.

PDF files created from scanned documents per se are not text searchable. The page content takes simply the form of an image. A character recognition (OCR) process is needed to create searchable text. This is an important step besides the compression using MRC or other techniques. Full-text searchable PDF files offer a new feature to archives, not available for paper documents.

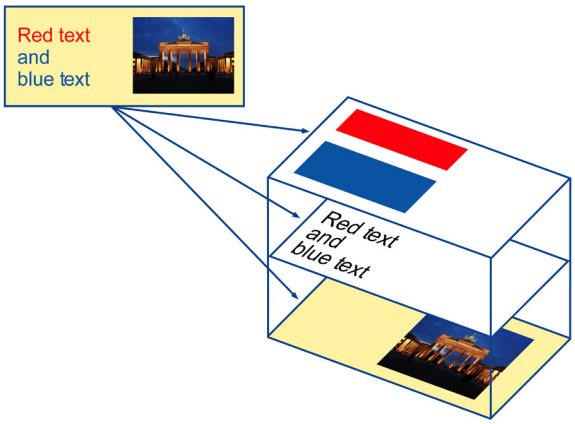

***Figure 1**. Text and image segmentation scheme*

The OCR result is typically placed as invisible text behind the page image. So the original content and visual appearance of the document is preserved, and in addition searching the document and copying text becomes possible, as it is with digital born documents. Since the OCR text is set invisible, the corresponding fonts need not to be embedded. The PDF/A specification explicitly excludes fonts used in invisible mode only from the need to be embedded.

Scanned documents do not provide any structure information or semantics. There is no description, defining certain text to be title, sub-titles, author names, etc. or giving explanations for certain figures. Thus an automated process that consists of scanning, compression, OCR, and PDF/A assembly can only create PDF/A-1 files at Level B. The PDF/A standard explicitly notes, that it is inadvisable for PDF/A writers to generate such structural and semantic information using automated processes without appropriate verification.

Digital Born PDF/A Documents

Electronic documents can exist in many different formats ranging from simple ASCII text, various office formats, CAD and other drawing formats, up to PDF itself. In most cases the structure and semantic information to create PDF/A-1 Level A is available. The difficulty here is to cope with all that different formats, most of them being proprietary, not including required elements like used fonts, or referencing inaccessible external resources.

Certain products can be used to convert such documents into PDF/A. Basically there are three different approaches:

1. Use a generic mechanism to deal with all formats in a unique way. This can be a printer driver that outputs PDF/A files. So all applications that can open the given document format, and are able to print, can be used. Advantage: One unique solution. Disadvantage: All that applications need to be installed on the system that creates PDF/A.
2. Use applications that already have built-in PDF/A creation capabilities, or that can use plug-ins like Adobe PDFMaker for Microsoft Office 2003. Advantage: In most cases more information like the table of contents is available and can be

automatically embedded into the PDF file. Disadvantage: Not all applications offer this kind of built-in functionality.

3. Use applications that at least have built-in PDF creation capabilities, and convert the PDF files to PDF/A using appropriate software. This can only be an interim solution. More and more applications are adding native PDF/A output support.

PDF/A and Digital Signatures

Digital signatures are a topic of its own that goes beyond the scope of this article. We only want to mention, that PDF/A-compliant documents can be signed with an embedded signature without causing the files to lose its PDF/A validity. Either the digital signature is applied at the same time the PDF/A file is created, or it is applied later on. The later is possible since the signature does not actually constitute a document change. A valid signature ensures the document has exactly the same form compared to the time it was signed. Qualified digital signatures provide the highest level of security since they use qualified certificates signed by certified service providers and up to date cryptographic technologies.

Qualified digital signatures can be used to meet the integrity and authentication requirements of digital archives.

PDF/A Validation

In an archiving workflow there are potentially two places, where a PDF document should be validated for PDF/A standard compliance: When a document is received from external or created internally, and just before it is transferred to the digital archive's data storage.

Using professional PDF/A creation or conversion tools can guarantee PDF/A compliance at the corresponding step within the workflow. However, files could be edited later on, e.g. by applying password protection (which is prohibited in PDF/A), or they can be tampered or changed unintentionally. Such operations can result in a loss of the compliance. So a final validation often is required in an archiving workflow.

Certain software applications are available for such inspection. It is important to note that a compliance test is not only a simple check of certain strings in the PDF file that indicate PDF/A compliance. Validators must perform a deeper inspection, ranging from general syntax checks to specific feature tests. We only give one example here. Fonts must be embedded in PDF/A, except they are only used in invisible mode. So it is not sufficient to verify that all fonts are embedded. If a font is not embedded, all content referencing this font has to be analyzed to ensure that each glyph of the font is set invisible.

The PDF/A Competence Center [17] is currently developing a PDF/A test suite that consists of standardized documents that deal with different aspects of PDF/A. These can be used as examples on what is allowed and what is forbidden in PDF/A. Moreover they can be used to check the validation applications themselves, resulting in more reliable software and secure PDF/A files.

Acceptance

Since the second half of 2007 many organizations published recommendations or in some cases directives for the use of PDF/A which shows the acceptance and necessity for PDF/A.

The fact that PDF/A is advantageous for businesses and institutions who want (or have) to be able to reproduce their files decades from now, is therefore becoming more and more internationally accepted.

Master studies are a good example. E.g. the British Library, the Technical University Library Vienna and the University of Potsdam store and publish master studies already in PDF/A format.

The German National Library prefers and recommends for submissions of documents to the Library, the PDF/A format [16].

A lot of other archives, libraries and government agencies already recommend or even require PDF/A as preferred format [17]. This is an ongoing process. And it would also be good practice for the valuable IS&T papers to be stored and published in the secure PDF/A long term format.

Outlook

As mentioned earlier, an additional part, ISO 19005-2 or PDF/A-2, is under development. Its current status is Working Draft and it is supposed to be published at the earliest in late 2009. PDF/A-2 is based on ISO 32000-1 [6] which will be the standardized version of the PDF Specification 1.7 [7]. PDF/A-2 takes into account new technologies that have been added to PDF since PDF 1.4, for example JPEG 2000 image compression [8-9]. It is important to note that PDF/A-2 will not invalidate PDF/A-1. Part 1 of the standard will stay valid and reliable for archiving. So there is no need for migration to PDF/A-2 once this part is published. Under certain condition, using PDF/A-2 will make sense in future, e.g. if digital master objects already exist in JPEG 2000 file format and should be converted to PDF/A documents without any recoding of the original page images.

ISO 32000-1 is currently in the approval step for the Final Draft International Standard, so publication is expected in 2008. Its goals are to exactly reflect what is in the PDF 1.7 reference, but to put it in the more precise wording that is needed for an international standard. So implementers will benefit from an unambiguous description of the various PDF features, leading to more stable software implementations. A second part, ISO 32000-2, is currently in the planning stage and ideas for new technologies and features are collected. There is no schedule for this part so far.

References

[1] ISO 15489-1:2001, Information and documentation - Records management - Part 1: General, www.iso.org. (2001).

[2] ISO/TR 15801:2004, Electronic imaging - Information stored electronically - Recommendations for trustworthiness and reliability, www.iso.org. (2004).

[3] ISO/TR 18492:2005, Long-term preservation of electronic document-based information, www.iso.org. (2005).

[4] ISO 19005-1:2005, Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A-1), www.iso.org. (2005).

[5] PDF Reference, Third Edition, Adobe Portable Document Format Version 1.4, www.adobe.com. (2001).

[6] ISO/DIS 32000, Document management - Portable document format - PDF 1.7, www.iso.org. (2008).

[7] PDF Reference, Sixth Edition, Adobe Portable Document Format Version 1.7, www.adobe.com. (2007).

[8] ISO/IEC 15444-1:2004, Information technology - JPEG 2000 image coding system - Part 1: Core coding system, www.iso.org. (2004).

[9] D. Santa-Cruz, T. Ebrahimi, J. Askelof, M. Larsson, C. Christopoulos, JPEG2000 still image coding versus other standards, SPIE's 45th annual meeting, San Diego, August 2000, Vol. 4115, pp. 446-454. (1998).
[10] ISO/IEC 16485:2000, Information technology – Mixed Raster Content (MRC), www.iso.org. (2000).
[11] ISO/IEC 15444-6:2003, Information technology - JPEG 2000 image coding system - Part 6: Compound image file format, www.iso.org. (2003).
[12] Klaus Jung and Thomas Zellmann, JPEG2000/Part6 for Scanned Documents in Archiving Applications, IS&T Archiving Conference 2004, San Antonio, pp.281-285. (2004).
[13] Simon McPartlin and Carsten Heiermann, New File Formats in Archiving: JPEG2000, High Compressed PDF, JBIG2 with Real World Examples, IS&T Archiving Conference 2005, Washington, DC, pp. 233-236. (2005).
[14] ISO/IEC 14492:2001, Information technology – Lossy/Lossless Coding of Bi-level Images, www.iso.org. (2001).
[15] P. Howard, F. Kossentini, B. Martins, S. Forchhammer, W.J. Rucklidge, F. Ono, The Emerging JBIG2 Standard, IEEE Trans. on Circuit and Systems for Video Technology, September 1998, Vol. 8, No. 5, pp. 838-848. (1998).
[16] Deutsche Nationalbibliothek, File Formats used for Document Submissions, www.d-nb.de/eng/netzpub/ablief/np_dateiformate.htm. (2007).
[17] PDF/A Competence Center, Homepage, www.pdfa.org.

Author Biography

Klaus Jung studied Physics at the Technical University of Berlin (TUB) and received his degree in 1991. 1991 - 1997 he was a research assistant at the Mathematics Department of the TUB where he received his PhD in mathematics in 1997. Since 1998 he is the R&D Director of LuraTech. Within the JPEG group he is acting as the Head of the German delegation since 1997.

Thomas Zellmann has been working in EDP for more than 20 years now and therefore has extensive experience with classic and modern IT solutions. He started his job at LuraTech in 2001. Prior to joining LuraTech he worked for Softmatic AG, Software AG and Nixdorf among others. Thomas Zellmann is one of LuraTech's shareholders and Chairman of the PDFA/ Competence Center.