PDF/A-2: The New Part of PDF/A

Thomas Zellmann, Member of the Board, PDF/A Competence Center and Shareholder, LuraTech Imaging GmbH, Berlin, Germany
Mark McKinney, President, LuraTech Inc., San Jose, California, United States

Abstract

During IS&T 2008 in Bern, LuraTech discussed PDF/A-1, which is the current ISO standard for long-term archiving. Since then, the ISO committee and its Class A liaison member, the PDF/A Competence Center, has been working on the new part of the standard, PDF/A-2 (ISO 19005-2). PDF/A-2 is now in draft international standard (DIS) status and is expected to become an international standard in Fall 2010. This paper presents an overview of PDF/A and describes the new functions of PDF/A-2 with respect to user applications. In addition, we will discuss the current usage of PDF/A in archives and libraries and outline the advantages of PDF/A-2 for these markets.

Introduction

PDF/A – the ISO 19005 standard [1] – defines requirements for creating documents suitable for archiving using the widely available PDF format. The standard specifies in detail what content is allowed and what is not. These specifications and others are intended to ensure long-term readability of the documents, regardless of the application software and operating system in which they were originally produced. The advantages of PDF/A, such as full-text search capabilities, have resulted in it replacing TIFF as the preferred archiving format for numerous international governmental organizations, as well as in the private sector.

As a standard for long-term archiving, PDF/A-1 (ISO 19005-1) will, by definition, never become obsolete or replaced by a new version. As a result, per the ISO standards-making process, PDF/A-2 (ISO 19005-2) will become a new part of the PDF/A standard [2]. With DIS complete, only formal standardization steps remain and therefore Part 2 is expected to become an international standard in Fall 2010.

PDF/A-2 will be based on PDF 1.7, which became an ISO standard (ISO 32000-1) [3] since the standardization of PDF/A-1. Therefore, the archiving subset PDF/A-2 will be based on a solid ISO standard, adding the functionality of PDF 1.7 [4], which will help with long-term archiving but does not endanger the goal of long-term reproducibility.

PDF/A-2 functions that will be described in the paper include JPEG2000, layers, transparency, Open Type fonts, ICC profiles and collections. Furthermore, some technical limits have been extended. A special emphasis will be put on the usage of PDF/A in archives and libraries and how those functions may be deployed in such applications.

PDF/A-2 Improvements

The PDF/A-2 standard will include new functions that are not in the current PDF/A-1, which will extend the benefit of PDF/A to further complete the “e-paper” and long-term archiving capabilities. Below we discuss the new functions of PDF/A-2, and illustrate their use in examples. These new functions include:

JPEG2000

Users frequently ask why JPEG2000 was not initially permitted in PDF/A-1. The explanation is simple. At the time of the PDF/A-1 standardization, PDF 1.4 [5] was valid and fully-developed, and therefore it was chosen as a basis for the archiving standard. JPEG2000 (ISO/IEC 15444) [6] only became part of the PDF specification when PDF 1.5 was released. PDF/A-2 therefore takes on JPEG2000 as a natural and sensible enhancement. JPEG2000 is particularly of interest in practical applications if the document in question is a scanned document. In this case, procedures that work according to the mixed raster content (MRC) [7,8,9] principle can achieve higher compression rates and higher levels of quality.

In contrast to the old JPEG, JPEG2000 also supports lossless compression, which can prove important in special applications. For example, libraries require that valuable, historic books and documents are first digitalized in the highest quality. The resulting “digital master” is compressed without loss using JPEG2000, but this file is still very large. The process of embedding the file in PDF/A has the advantage that the metadata also can be integrated into this file in a standardized way. In addition, the lossy JPEG2000 compression can be used to create a manageable PDF/A of manageable size, which then can be used for presentation and delivery downloads in the Web.

Levels

PDF/A-2 supports levels or “optional content” in documents. This feature can be used in cases with multilingual documents. For example, the user can use the level function to easily switch between German and English versions. Also, this function is frequently used by the construction industry, which likes to hide or display specific aspects of detailed design drawings so that they can gain a clearer view.

Transparency

This PDF function is often used in printing, for shading, borders or smooth transition between page objects. Highlighting flags are often used in comments to clarify passages of text and, in PDF/A-2, these can be transferred to the archive without any changes.

OpenType Fonts

Because of their high performance and far-reaching Unicode support, OpenType fonts (cf. ISO/IEC 14496-22) are more frequently used today. Now, they can be embedded in PDF/A-2 without being recoded to PostScript Type 1 or TrueType fonts.

ICC Profiles

Introduced in PDF 1.6, v4 of ICC profiles (ISO 15076-1:2005) also can be used in PDF/A-2. Therefore, in certain cases, supports more detailed color definitions than the version that was previously supported for ICC profiles. This enhancement is particularly important for digital photos and color scanning.

Collections

Collections – also referred to as “portfolios” in Acrobat - enables users to logically merge several PDF files into one “container PDF.” This function has now been enhanced to include the process of creating PDF/A collections of several PDF/A files.

In this case, users have identified two practical uses of this feature: First, within the social security sector, state scan operators are legally required to use a qualified digital signature to confirm that the scan is a visual match to the original. The paper version can then be destroyed. In practice, a single-page signature is often used because users may need to change or split the documents when the contents are edited later.   A second use is in e-mail archiving. PDF/A is increasingly being used as a secure long-term archiving format. When using PDF/A-1, you can convert an e-mail with attachments into individual PDF/A files or all together into one single complete PDF/A file. However, if you use collections for this, it is even easier to retrace, for example, that the first attachment was originally a Word file. In accordance with this, the entire e-mail message is created as a collection that contains several PDF/A files.

No Limits

PDF/A-2 also eliminates technical limitations that existed in Adobe Acrobat 5. For example, a document may now have dimensions that are several “hundred kilometers” – page sizes on a scale 1:1 of up to 381 km feed size (previously 5.08 m) are possible.

In practical applications, this is definitely not noticeable for typical DIN-A4 business documents. However, in site plans in the land register and for water, gas or electricity plants, as wells as architectural drawings for buildings, it is possible to reach these limits. In certain circumstances, plans that have very large dimensions also arise in the digital construction sector.

As a result, in many cases, it is now possible to save such plans digitally on a scale of 1:1 and, for example, to measure ranges without conversion in Adobe Reader. Even if it still conforms to the standard, an older version of Adobe Reader finds it difficult to display oversized documents and is therefore slow. However, even though it does not directly adhere to the standard, the required performance is provided as of Adobe Reader 7 or higher, currently the most popular viewer. For problem-free work, it is important that the current versions of Adobe Reader or other PDF viewers are used.

PDF/A-2a, 2b and 2u

As with PDF/A-1, PDF/A-2 contains subtypes, also known as different compliance levels.

PDF/A-1 has only two compliance levels: PDF/A-1a and PDF/A-1b. Simply speaking users may consider “b” for “basic” because this sub-level contains all functions in order to guarantee long-term reproducibility, while “a” can be considered “advanced” because it adds Unicode support and accessibility through PDF tags.

This same concept of different conformance levels continues within PDF/A-2, however it is taken a step further. In the real world, it may be impractical to always create a large volume of PDF/A-1a compliant documents because of the additional manual effort that would be required to properly tag all of the content. However, if there is a requirement for Unicode support, then it becomes a difficult choice. Therefore, within PDF/A-2, the ISO standards committee created a new conformance level, the PDF/A-2u level in order to create documents that leverage the benefits of Unicode without having to tag all of the contents of the document. As a result, there are three conformance levels, starting with the simplest 2b, and increasing in complexity with 2u for Unicode and then 2a. However, users should keep in mind that all three conformance levels are equally valid. When deciding which conformance level, users should consider the actual requirements of the archive and eventual use cases of the PDF/A documents.

Impact on the User

Those organizations or individuals that already use PDF/A-1 today do not have to migrate to PDF/A-2. PDF/A-1 will continue to be a valid standard and the compatibility will, of course, continue to be ensured. A valid PDF/A-1 document also is automatically as a valid PDF/A-2 file.

If the new functions that are mentioned above are of interest for a project, you can convert to PDF/A-2 at a defined time that makes sense for the project, irrespective of the tools that you are using. If someone is planning to use PDF/A in their applications, they also can already incorporate PDF/A-2 into their plans. This enables organizations that have large projects that will extend into Fall 2010 and later to begin implementing PDF/A-2 now.

In a broader sense, PDF/A should be currently regarded as an obvious selection criterion when implementing new document solutions. In addition to the traditional advantages of PDF/A - including long-term reliability and format standardization – we believe PDF/A can simply be described as “good PDF”, in other words a high quality PDF.

PDF/A in Archives and Libraries

Five years after PDF/A became an international standard, many organizations have recommended requirements for the usage of PDF/A, particularly in archives and libraries.

From the European perspective, there is growing momentum for PDF/A. Most importantly,at the start of 2010 in the Netherlands the National Archive stated, “Nationaal Archief kiest PDF/A”, which translates into the recommendation of PDF/A, “” Also, the Koninklijke Bibliotheek (KB) is working on PDF/A projects and Het Utrechts Archief is a long-term PDF/A user for scanned documents.

In Germany, the Bundesarchiv and the state archive of Baden-Württemberg are leading with PDF/A.

In the library community, there are recommendations for PDF/A by the Austrian and German national library [10]. The European Patent Office (EPO) also has heavily adopted PDF/A.

In North America, it is interesting because the concept of PDF/A was first created in the United States in a conversation between industry association AIIM and the U.S. Court systems (Betsy Fanning and Stephen Levenson). But the acceptance of PDF/A in North America has taken longer than in Europe. This is due to the fact that ISO standards are not always mandated by the government. Therefore, many organizations adopt standards after waiting for a consensus among many groups within their industry.

However, it is now starting to get close to a tipping point in the U.S. and Canada. There currently are 14 state governments that are considering implementing new laws that will suggest the use of AIIM and ANSI supported standards, which of course includes PDF/A. Some states have departments that have already adopted the PDF/A standard, such as in New York where the departments of Education and Health utilize it.

In addition, the National Archives and Records Administration (NARA) have posted guidelines for submitting electronic records on their website that include the ability to submit PDF/A documents as long as they meet their other electronic submittal guidelines [11]. In parallel, the U.S. Nuclear Regulatory Commission (NRC) follows the NARA guidelines for electronic submittal of PDF documents [12], and the U.S. Library of Congress (LOC) shares information about digital preservation using PDF/A on its website [13]

As awareness of the benefits around archiving documents, more industries and agencies have begun to investigate and implement PDF/A in North America. The PDF/A Competence Center [14] started the North American Chapter in 2009, and have already hosted two seminar series on PDF/A in partnership with AIIM. Moreover, there have been numerous talks on PDF/A at conference – both on local and national levels – hosted by ARMA, AIIM, NIRMA and IBM. Once North Americans realize that PDF/A is not an adoption of a new technology, rather a restricted use of PDF to mitigate risk, then decision to move to PDF/A becomes clear for organizations that are already working with PDF within their archives.

References

[1] 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).

[2] ISO/DIS 19005-2, Document management - Electronic document file format for long-term preservation - Part 2: Use of ISO 32000-1 (PDF/A-2), www.iso.org. (2010).

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

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

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

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

[7] ISO/IEC 15444-6:2003, Information technology - JPEG 2000 image coding system - Part 6: Compound image file format, www.iso.org. (2003).

[8] Klaus Jung and Thomas Zellmann, JPEG2000/Part6 for Scanned Documents in Archiving Applications, IS&T Archiving Conference 2004, San Antonio, pp.281-285. (2004).

[9] 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).

[10] Deutsche Nationalbibliothek, File Formats used for Document Submissions, www.d-nb.de/eng/netzpub/ablief/np_dateiformate.htm. (2007).

[11] NARA FAQs about Transferring Permanent Records in PDF/A-1 to NARA, www.archives.gov/records-mgmt/initiatives/pdf-faq.html.

[12] NRC: Guidance for Electronic Submissions, www.nrc.gov/site-help/e-submittals/guide-electronic-sub-r5.pdf. (2009).

[13] Library of Congress: Sustainability of Digital Formats, www.digitalpreservation.gov/formats/fdd/fdd000125.shtml

[14] PDF/A Competence Center, Homepage, www.pdfa.org.

Author Biography

Thomas Zellmann has been working in EDP for more than 25 years. Zellmann is a member of the board of the PDF/A Competence Center and one of LuraTech's shareholders. He started his job at LuraTech in 2001. Prior to joining LuraTech Zellmann worked for Software AG and Nixdorf among others.

Mark McKinney has more than 14 years of executive and product development/management experience. He is currently president of LuraTech, where he specializes in product management and corporate strategy. Prior to joining LuraTech, McKinney held various marketing and management roles at Wyse Technology, Axolotl Corp., Siemens and Uni-Data. McKinney was president of the San Jose Leadership Council in 2004-2005 and, from 2002 to 2005, he served as a board member and vice chair of non-profit InnVision.