PDF/A-3, the newest part of the ISO standard 19005

Thomas Zellmann, PDF Association, Berlin, Germany and Hans-Joachim Hübner, SRZ, Berlin, Germany

Abstract

This paper discusses five topics concerning PDF/A-3, the ISO standard 19005-3:

It starts with a short overview about PDF/A as the ISO standard for long-term archiving and reviews its developments since the release of PDF/A Part 1 in 2005.

Secondly, PDF/A-3 as the newest part of the PDF/A family is presented with the reasons as to why this additional part was standardized by ISO.

PDF/A-3 allows new use cases of PDF/A, and sample applications are then discussed.

Fourthly there were several discussions about the pros and cons of embedding arbitrary files into PDF/A, and the arguments are reviewed.

The paper finishes with addressing new format recommendations from NARA and NSDA-Library of Congress.

Overview PDF/A

PDF/A is a multi-part ISO standard developed over many years of committee work by industry associations, businesses and public authorities around the world. The result is ***"a file format based on PDF, known as PDF/A, 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."*** (ISO 19005-1, quoted from the introduction [1]).

The first part of the standard, PDF/A-1, has been available since the 1st of October 2005. Its official designation is "ISO 19005-1:2005. Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)".

Since then, two further parts have been made available to users: PDF/A-2 (since 2011) and PDF/A-3 (since 2012 [3]). These parts exist in parallel and are optimized to meet particular needs.

The PDF/A standards family regulates how to create electronic documents to ensure they can be reliably reproduced for decades to come. The standard does not describe how to build a revision-safe archive, nor the theory behind one.

The decisive advantages of PDF/A

- A PDF/A file contains everything needed to display it and nothing which could negatively impact the display.
- PDF/A files can be used on any platform.
- Free programs exist for displaying PDF/A files.
- The multi-part PDF/A standard offers great flexibility to users.

Widespread acceptance of PDF/A

PDF/A is becoming more and more common, be it in industry, public administration, financial services or academia. A large number of authorities and institutions worldwide recommend PDF/A or specifically require the use of the standard.

PDF/A-3, the newest part

ISO realized from the beginning of the PDF/A standardization process that the future would bring new possibilities, and therefore planned for multiple parts from the start.

On one hand, it is clearly defined by ISO that PDF/A-1 (and all other parts) will never become invalid or superseded.

On the other hand, technological developments are integrated through releasing new parts of the PDF/A standard. PDF itself became the ISO standard 32000 in 2008 [3] and PDF/A-2, released as the second part in 2011, was based on ISO 32000, integrating some useful features of PDF 1.7.

ISO and its experts are always open for suggestions from the user community, and multiple users from different industries proposed to extend PDF/A with the so-called container option which PDF had already offered for some years.

ISO processed this user input and published PDF/A-3 in 2012 [2].

Looking from a features viewpoint, PDF/A-3 only adds one: the possibility to embed multiple other files into a PDF/A-3 container. This opens numerous new application areas which are discussed below.

The embedded files have a 'type' which indicates how the files relate to the PDF/A part of the PDF file. The types are SOURCE, SUPPLEMENT and ALTERNATIVE.

New use cases with PDF/A-3

This section describes sample use case of PDF/A-3 and first experiences on how users deploy PDF/A-3 for their applications.

For the type SOURCE, this paper itself provides a basic example. The paper was written as a Microsoft Word file. It has always been a good idea to convert such papers to PDF or PDF/A for distribution and archiving. Now it is possible to convert the source Microsoft Word file into a long-term safe PDF/A file, and by using PDF/A-3 the Word file can be easily embedded in it. The result is one file or archive object which contains the complete paper. Independent of unknown future computer environments, there will always be only one file which contains the original source as well as a long-term reproduction with the PDF/A part.

SOURCES can be any office documents like Microsoft Office, Open Office, LibreOffice documents, and others in the classic document world.

SOURCES can also be files from industry specific applications.

For example, a CAD program creates a DWG file that can be embedded as well. This helps in cases within an organization where typically only the construction department has CAD applications and can open and read the DWG format. If e.g. the marketing department needs information about the CAD object, they can easily work and display the PDF/A part with a simple PDF viewer.

Another interesting use case is when signed files or emails arrive from an external source. If for example a signed PDF arrives and the organization employs a good archiving policy using PDF/A, it has been an issue that simply converting the PDF file to PDF/A already breaks the signature, which is the correct approach from the signature point of view. Now, a conversion to PDF/A is possible with the original signed file being embedded into the PDF/A-3 file.

There are countless other cases where embedding the original source file into a PDF/A-3 file makes sense.

SUPPLEMENT is the second type for embedding files.
The concept with SUPPLEMENT is to add and embed every bit of information which is related to a specific document.
In an office environment, this could be for example a Mindmap graphic which belongs to a description or meeting protocol.
An example from the industry environment would be an analysis or report in PDF/A-3 format that has the raw data from measurement devices embedded in it.

Now we would like to discuss two specific application areas of PDF/A-3 which are already in use:

Electronic invoicing in Germany:

Driven by German government and industry, and in cooperation with the European Union, the German FeRD group defined a so-called ZUGFeRD XML schema which describes the content of an electronic invoice.

FeRD selected PDF/A-3 as the required file format for electronic invoices according to the ZUGFeRD schema.

This is a perfect example of the third type for embedding files, ALTERNATIVE. It reflects a combination of the invoice as a PDF/A-3 file with the machine readable XML file as an alternative format of the same content embedded in it.

Email archiving:

A lot of companies already use PDF/A in order to convert emails into a safe long-term archiving format for their document management system.

With the first parts of PDF/A, the email was basically flattened into one sequential file, which in itself is good but loses the relationship between the email body and attachments.

PDF/A-3 allows for a complete archiving of emails. The email body and attachments are converted to PDF/A for a guaranteed reproduction at all times. The attachments are also embedded into the PDF/A-3 file in their original format resulting in only one archive object which reflects and archives the complete email.

In our opinion, and based on the use cases above, we are convinced that PDF/A-3 is a perfect solution for all classic documents in the ECM world.

Pros and Cons of PDF/A-3

We would like to mention though that there have been several discussions concerning PDF/A-3, and some people have argued that PDF/A-3 opens a Pandora's Box.

The critics are of the opinion that PDF/A-3 breaks the aim of long-term guaranteed reproducibility with a PDF/A file, since PDF/A-3 allows the embedding of any file. The use cases above discuss practical cases, but the term 'any file' can also mean a virus, audio or video file, or any other material that could be deemed 'dangerous' from a technical point of view.

After expert discussions with the PDF Association, it was concluded that 'purists' who do not want to use PDF/A-3 can limit their document usage to PDF/A-1 or PDF/A-2, since all parts of the PDF/A standard are still valid.

The critics concerns are well taken, but knowledgeable users understand their processes and should be allowed to make meaningful use of the technical possibilities available to them.

Based on this discussion it must be emphasized that only the PDF/A part of a PDF/A-3 file is long-term reproducible! Embedding an audio, video or 3D file is possible, but since PDF/A is the long-term archiving standard for 2-dimensional documents the dynamic documents do not inherit the long-term reproducibility guarantee.

The embedded files can be very helpful as described in the use cases, but at the end of day they will always be dependent on an application that still can open the specific file type.
Looking from a document process and document live cycle point of view, the embedded files may be useful for ongoing work or new versions of documents, but in the long-term they may just end up as data garbage.

New Format Recommendations from NARA and NSDA

As the authors are, to be honest, not completely neutral, we also want to mention a detailed report of the NSDA PDF/A-3 Standards Working Group [6]. Interested readers can also find discussions concerning this paper on LinkedIn.

NARA's 2014 Transfer Guidance [7] for electronic documents also does not cover PDF/A-3 yet. This is a matter of timing as the paper was started before PDF/A-3 was published. We are curious as to how NARA and the whole archiving community will discuss PDF/A-3.

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)

[2] 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)

[3] ISO 32000-1:2008
Document management -- Portable document format --
Part 1: PDF 1.7

[4] PDF/A in a Nutshell, PDF Association 2013, www.pdfa.org

[5] FeRD and ZUGFeRD: www.ferd-net.de

[6] NDSA Report: The Benefits and Risks of the PDF/A-3 File Format for Archival Instititutions:
http://blogs.loc.gov/digitalpreservation/2014/02/new-ndsa-report-the-benefits-and-risks-of-the-pdfa-3-file-format-for-archival-institutions/

[7] NARA's 2014 Transfer Guidance – Introduction and link:
http://www.pdfa.org/2014/02/new-us-federal-government-transfer-guidance-released/

Author Biography

Thomas Zellmann *has been working in EDP for more than 20 years and has extensive experience with classic and modern IT solutions. He started his current position at* LuraTech *in 2001, working in the archives/libraries segment and is one of LuraTech's shareholders.*
As managing director of the PDF Association he coordinates and executes most of the organizations' activities. He is the main contact for all members, vendors and users.

Hans-Joachim Hübner is head of solutions and sales at Satz-Rechen-Zentrum Berlin, a German company specializing in the areas of ECM and document capture. He has made major contributions to the company's software development success in the areas of document capture, ECM and content management systems.

During more than 25 years of experience, he has managed successful large digitization projects and has developed a range of solutions and workflows in the ECM-, library and archive context.