PDF/A – the standard for long-term archiving

- What are the advantages of the PDF/A Standard?
- What is the PDF/A Standard?
- What do PDF/A-1a, PDF/A-1b, PDF/A2 mean?
- How is the PDF/A Standard implemented?
- Is PDF/A the solution for long-term archiving?

Index

Background

PDF/A – the standard for long-term archiving

Introduction

On September 28, 2005 the International Standards Organization (ISO) formulated a new Standard governing archiving of electronic documents – the official formulation is: ISO-19005-1 - Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A-1).

The Standard is the result of more than 36 months of collaboration among companies and organizations around the world.

The initial impetus for this initiative occurred in May 2002 in the USA. The stated goal was to create a standardized format for electronically archived documents. The Association for Information and Image Management (AIIM), the National Printing Equipment Association (NPES) and the administrative body for the US courts were all involved. The kick-off meeting took place in October 2002. Renowned PDF manufacturers participated, including: Adobe Systems, the Library of Congress, Surety Inc., Quality Associates Inc., Appligent, Merck, EMC, PDF Sages, and the National Archives & Records Administration (NARA). These were joined at a later time by others, including Xerox, Honeywell, EDS and Glaxo Smith Kline, among others.

The founders of the project put together a first version and submitted their recommendation to the ISO in order to have it registered as an international Standard. The project was referred by the ISO to a Technical Committee designated TC 171 (Document Management Applications). This committee is composed of 15 member states and each has one vote, which is cast by their respective representatives. The committee is supplemented by an advisory commission representing another 21 countries. The Standard was improved over multiple stages until it was finally approved in September 2005.

PDF/A was elevated to the status of international Standard in 2005.

Renowned organisations and manufacturers, as well as professional users, were involved.

What are the advantages of the PDF/A Standard?

Almost every country uses its own format to archive documents. Traditional archiving methods such as paper, microfilm and microfiche can be archived for a long time and can be reproduced over the long-term. These formats, however, lack the advantages of digital technology: large documents cannot be quickly and easily sent around the world. Additionally, it is almost impossible to search documents archived in the traditional way for specific information. Many institutions chose the TIFF format for their first electronic archives. This format guarantees long-term reproducibility, and the image format TIFF is established and can easily and quickly be sent to networked institutions around the world. Freely searching for information in the TIFF format, however, remains problematic.

We were able to produce the electronic archive thanks to innovative technologies.

PDF offers more advantages than the TIFF format.

This was the reason why the PDF format began to be considered. There are many reasons why PDF is the most attractive alternative:

- PDF includes structured objects (text, vector graphics, rastered images). This allows efficient search queries to be performed through the entire data archive. TIFF is also a rastered format. In order to be able to perform a full-text search, the TIFF document must first be prepared by optical text recognition software (OCR).

- PDF can be efficiently and compactly compressed. In comparison to an equivalent TIFF file, a PDF file requires only a fraction of the memory space. Moreover, the quality is almost always better. The smaller file size is extremely advantageous for electronic file transfer in particular, such as email attachments, FTP, etc.

- Metadata (author, topic, content, keywords, date created, date modified, publisher, etc.) are embedded directly in the PDF file using a standardized format (XMP). This means that the metadata can be automatically and systematically amended without manual steps.

- The PDF format is generally conceived such that it is not tied to any one particular device (raster resolution, color system, etc.). Only once the document is displayed is the page content represented in the viewer or on the printer; this process is called "rendering". This means that PDF documents adapt to the technological development of the output devices (printer, monitor, scanner, etc.) and remain up-to-date for years after their creation.

Over the past fifteen years, Adobe Systems, the author of the PDF format, has published a total of eight versions of its PDF Reference Manual. In each revision, the PDF format was expanded to include new functionalities and existing ones were overhauled. For this reason, it was unavoidable that an enduring, stable and internationally valid standard for long-term preservation would be based on the Adobe PDF format. The result of this development is the PDF/A Standard.

The PDF/A Standard

Goals of PDF/A

ISO Standard 19005 defines a file format based on PDF called PDF/A. The format offers a mechanism that represents electronic documents such that the visual appearance remains preserved for an extended period, independent of tools and systems for producing, saving and reproducing it. This Standard specifies neither the methods nor the intention or the purpose of preservation. The Standard is thus intended to guarantee that electronic documents can be viewed in their original appearance, even in the future. For this reason, the document may not refer, either indirectly or directly, to an external source, for example an external image or a font that is not embedded in the document itself.
PDF/A is structured as a series that includes multiple Standards. At present however, only PDF/A-1 (ISO 19005 Part 1) has entered into effect.

PDF/A files are self-describing.

All information necessary to display the document is embedded.

PDF/A-1 is based on PDF Reference 1.4. PDF/A-2 will be based on the ISO Standard for PDF 1.7 (ISO-32000).

The distinction between PDF and PDF/A

The PDF format does not guarantee long-term reproducibility or complete independence from software and the output device. In order to guarantee both principles, it was necessary to both limit and expand the existing PDF Standard. It was clear from the outset that PDF/A-1 would have to be based on an existing version of PDF in order for it to be accepted by the broadest possible swathe of the public. The ISO committee TC 171 chose the Adobe PDF Reference 1.4 as the basis for the PDF/A-1 Standard.

This Reference was implemented by Adobe in their Acrobat 5 product. Since it is a standard, PDF/A-1 must fulfil all requirements of this Reference, and must also respect certain technical limitations of Acrobat 5. The original PDF Reference and ISO 19005-1 together comprise the current PDF/A-1 Standard. ISO Standard 19005-1 only identifies the differences with respect to the PDF Reference. Accordingly, PDF Reference 1.4 is the central basis on which to comprehend the PDF/A-1 Standard.

Certain functionalities of PDF 1.4, such as transparency or the integration of audio and video, are not permitted by the PDF/A-1 Standard. Certain options outlined in PDF 1.4 are mandatory in PDF/A-1: for example, all fonts used must be embedded in the document. Essentially, the PDF/A-1 Standard does nothing other than specifically identify individual characteristics of PDF Reference 1.4 and to indicate whether each is absolutely necessary, recommended, limited, or not permitted.

PDF/A, A-1a, A-1b, A-2 „Babylon"

The PDF/A-1 Standard is divided into two levels of conformance: PDF/A-1a and PDF/A-1b.

PDF/A-1a (Level A Conformance) defines conformance with all requirements of the PDF/A-1 Standard.

There are two levels of conformance in PDF/A-1.

PDF/A-1a meets all requirements. PDF/A-1b meets the minimum requirements.

The minimum requirements for conformance with PDF/A-1 are contained in PDF/A-1b (Level B Conformance). The PDF/A-1b requirements are generally sufficient for unequivocal reproduction over an extended period.

PDF/A-1a differs from PDF/A-1b mainly with respect to accessibility requirements (Paragraph 508 of the US Rehabilitation Act).

- PDF/A-1a guarantees that the document text is extractable and that the logical structure of the document as well as the natural reading process of integrated text material remain intact. Text extraction is mainly of interest if documents are to be displayed on mobile devices (e.g. PDA) or visualized in the sense of Paragraph 508 of the US Rehabilitation Act. This includes the requirement that the representation of the text fit on the reduced screen by being restructured (re-flow). This functionality is also known as "tagged PDF".

- PDF/A-1b ensures that text and other content on pages is reproduced uniformly; it is not a guarantee, however, that the embedded text is comprehensible and legible. The creator of a PDF/A-1b conformant file is at liberty to embed the text in a readable form, even if the more stringent requirements pursuant to the aforementioned Section 508 are not met.

For scanned documents, conformance with PDF/A-1b is completely sufficient, even if they have been processed using OCR to enable a full text search.

The Technical Committee is currently working on a new component of the Standard: PDF 19005 2 (PDF/A-2). PDF/A-2 is being developed in order to take account of the expanded scope of functionality outlined in PDF Reference 1.7. In the meantime, PDF 1.7 itself has been standardized, i.e. PDF/A-2 will no longer be based on the Adobe PDF Reference, but rather on the new ISO Standard 32000-1 (PDF 1.7).

Contrary to usual practice with PDF References, PDF/A-2 does not replace the existing Standard PDF/A-1, but rather will exist alongside it for perpetuity. PDF/A-2 conformant viewers must also be able to simultaneously display PDF/A-1 conformant documents correctly.

PDF/A-1a was drafted to meet the accessibility requirements set out in Paragraph 508 of the US Rehabilitation Act.

PDF/A-1b is sufficient for uniform visual display of documents.

PDF/A-2 is not intended to replace the existing Standard, but rather to incorporate new PDF functions.

Use of the PDF/A Standard

Where do I get a copy?

The PDF/A-1 Standard ISO 19005-1 is distributed directly from the ISO Website (www.iso.org). Both paper copies and electronic versions (as PDF) are available. As is the case for all other ISO Standards, the document is copyright protected. It is therefore illegal to offer free copies via the internet. The PDF/A-1 Standard is only available in English.

PDF/A-1 Standard ISO 19005-1 is available from the ISO website: www.iso.org

PDF/A is a purely technical standard and expert knowledge is required to implement it.

To whom is the Standard addressed?

The objective of the PDF/A Standard is to optimize archiving methods. The Standard is purely technical in nature. For this reason, it is essentially only fully comprehensible to specialists with extensive knowledge about page description languages such as PostScript and PDF. The main document itself is small, however the scope of the basis document is very large. PDF Reference 1.4 alone consists of almost one thousand pages – and this does not include all information associated with the Reference, such as font and compression formats, XML specifications, ICC color profiles, digital signatures, RFCs, etc. In addition, the Standard alone cannot guarantee long-term preservation. A strategy for developing company-wide archiving is generally the result of a comprehensive project. Collaboration with experts who understand the requirements of the PDF/A Standard and can apply them is recommended. Only in this manner can a consistent strategy be produced that ensures long-term document preservation goals.

What tools are available?

Various tools to create, process and verify PDF/A documents have been on the market since 2006. Version 8 of Adobe Acrobat includes appropriate tools. Microsoft offers an Add-in for Office 2007 that can be downloaded separately. It allows users to produce PDF/A conformant documents directly, using the Office palette. As innumerable products to create PDF/A documents already exist, the results of the different products with respect to unobjectionable PDF/A conformity must be verified.

PDF/A as a component of a comprehensive long-term archiving concept

In itself, the PDF/A Standard is merely a component of a comprehensive solution. In isolation, the Standard does not guarantee long–term preservation or reproduction parameters. Moreover, it is not the ideal solution for every project. PDF/A defines the specific requirements for electronic documents so that they can be archived over the long-term. To build an archive that is conformant to the PDF/A Standard, other aspects must be taken into consideration. These include, among other things, in-house company standards and processes, quality management, reliable data sources and dedicated requirements tailored to the specific application purpose. In particular, the transfer of existing paper or TIFF archives to a PDF/A conformant archive requires careful planning.

PDF/A is a component of a comprehensive archiving strategy.

The Standard alone does not guarantee long-term preservation, however it is an essential requirement to achieve that objective.

Conclusion

PDF/A – the archiving standard

PDF/A is the standard for archiving electronic documents. The PDF format is widespread globally. It is used in both the public and private sectors for a wide range of purposes. The PDF/A Standard is the perfect instrument to ensure long-term preservation and reproducibility of documents over extended periods.

The PDF/A Standard also influences the future development of the PDF format itself. Independent of it, Adobe will continue to develop new functionalities. For example, 3-dimensional models or XFA for dynamic PDF forms. Conversely, these developments will influence the PDF/A Standard.

What is the market reaction?

It is not expected that PDF/A products will inundate the market. The knowledge required to understand the PDF/A Standard technology is considerable and specific. In addition, the user expects more sophisticated quality from software than is appropriate for a standard. The first applications have been on the market since 2006. Demand is primarily for PDF/A conformant production of documents , which check PDF/A conformity (validation) and enable simple conversion of existing PDF documents into PDF/A documents.

Comprehensive projects to build PDF/A conformant archives have arisen along with the first professional PDF/A tools. Currently however, one must not have excessively high functionality expectations. As is so often the case when introducing a new standard, many products will be released to the market that advertise PDF/A conformity yet do not actually fulfill the requirements of the Standard. The use of expert opinions for evaluation purposes is strongly recommended.

PDF/A as a long-term strategy

The PDF/A Standard will not be short-lived. Demand has existed for years for a standardized framework for archiving with PDF. The format is already used for precisely this purpose, even if many users must define specific guidelines in order to do so. The fact that Microsoft is responding to customer demand by making it possible to create PDF/A documents directly from the most recent Office palette is a clear signal: the internationally valid PDF/A Standard for long-term archiving is here to stay.

Information

PDF/A Competence Center

The PDF/A Competence Center was founded in 2006. The objective of this international organization is to promote the exchange of information and experience with regard to long-term archiving in conformance with ISO 19005 – PDF/A. The Board is composed of managers drawn from the following companies: callas software GmbH, Compart System-haus GmbH, intarsys consulting GmbH, LuraTech Europe GmbH, PDF Tools AG, PDFlib GmbH and Seal Systems. In less than two years, over 85 companies and organizations as well as numerous specialists from more than 20 countries have joined the organization. www.pdfa.org

About PDF Tools AG

PDF Tools AG counts more than 4,000 companies and organizations in 60 countries among its customers, making it one of the world's leading producers of software solutions and programming components for PDF and PDF/A products.

Dr. Hans Bärfuss, founder and CEO of PDF Tools AG, began using PDF technology in customer projects more than 15 years ago. Since then, the PDF and PDF/A format have evolved into a powerful, widely used format and ISO standard that can be used for almost any application. During this time, PDF Tools AG has developed into one of the most important companies on the market for PDF technology, and has played a significant part in developing the PDF/A ISO standard for electronic long-term archiving.

As the Swiss representative on the ISO committee for PDF/A and PDF, the company's knowledge flows directly into product development. The result is high quality, efficient products based on the 3-Heights™ philosophy of the development team, which consists of experienced engineers.

The portfolio of PDF Tools AG ranges from components to services through to solutions. The products support the entire document flow, from raw materials to scanning processes through to signing and storage in a legally compliant long-term archive. An advantage of the components and solutions is the broad range of interfaces, which ensure smooth and easy integration into existing environments.

Due to the growing demands of the market, the products are enhanced and refined continuously. Support is provided by the developers themselves, allowing them to identify trends and customer requirements quickly and use this knowledge when planning enhancements and components.

All development activities are performed in-house at PDF Tools AG in Switzerland. The company does not outsource any programming, so that the entire development process can take place centrally in a single location. This helps to ensure the high standards expected by the company, particularly with regard to the 3-Heights™ technology.
The effectiveness of this approach is confirmed by the success of the products on the market. Our customers include well-known global companies from every industry. That is the greatest compliment of all – and the perfect motivation to continue shaping the world of PDF and PDF/A.

PDF Tools AG | Kasernenstrasse 1 | 8184 Bachenbülach | Switzerland
Tel.: +41 43 411 44 51 | Fax: +41 43 411 44 55
pdfsales@pdf-tools.com | www.pdf-tools.com

PDF-TOOLS.COM
Premium PDF Technology