

White Paper

PDF/A - The Basics

From the [Understanding PDF](#) White Papers
PDF Tools AG

- **Why is PDF/A necessary?**
- **What is the PDF/A standard?**
- **What are PDF/A-1a, PDF/A-1b, PDF/A2?**
- **How should the PDF/A Standard be applied?**
- **Is PDF/A the answer to long-term archiving?**



Version: 2.0

Date: January 22, 2007

Copyright ©2007 PDF Tools AG. All Rights Reserved.

Other names and brands may be claimed as the property of others. Information regarding third party products is provided solely for educational purposes.

PDF Tool AG is not responsible for the performance or support of third party products and does not make any representations or warranties whatsoever regarding quality, reliability, functionality, or compatibility of these devices or products.

Contents

- Contents.....2**
- Introduction3**
 - Background.....3**
 - Why the PDF/A initiative?3**
- The PDF/A Standard5**
 - The Goal of PDF/A.....5**
 - PDF vs PDF/A.....5**
 - The PDF/A, A-1a, A-1b, A-2 "Babylon"5**
- Using the Standard.....7**
 - Obtaining a Copy7**
 - Who should read the PDF/A Standard.....7**
 - What tools are available?7**
 - PDF/A requires a complete solution.....7**
- Summary.....8**
 - PDF/A as the new archiving standard8**
 - How will the market react.....8**
 - Hot air or a long-term solution?8**

→ Introduction

PDF/A - A new Standard for Long-Term Archiving

Background

On September 28, 2005 the International Standards Organization (ISO) approved a new Standard governing electronic document archiving:

ISO-19005-1 - Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A-1).

This standard is a product of over 3 years of meetings, discussions, and review by organizations and companies worldwide.

The initiative to create a standard format for electronically archived documents based on Adobe's PDF format was launched in May 2002 in the United States by AIIM (Association for Information and Image Management), the NPES (National Printing Equipment Association) and the Administrative Office of the U.S. Courts. The kick-off meeting was held in October 2002 and included numerous electronic documentation users and PDF suppliers including Adobe Systems, Library of Congress, Surety Inc., Quality Associates Inc., Appligent, Merck, EMC, PDF Sages, and NARA (National Archives & Records Administration). Subsequent attendees included Xerox, Honeywell, EDS, and Glaxo Smith Kline to name a few.

The US initiative prepared a first draft and submitted their project to the ISO to be registered as an international Standard. The ISO assigned the project to a Technical Committee (TC 171 - Document Management Applications). TC 171 is comprised of 13 participating countries (who each have one vote) and 21 observer countries. After numerous reviews and amendments the Standard was approved by the ISO participating companies in September 2005.

PDF/A was approved as an international standard in September 2005.

Numerous organizations, users and suppliers were involved in the process.

Why the PDF/A initiative?

Approved archiving formats vary from country to country. Traditional archiving methods (paper and microfilm / microfiche) guarantee reproducibility but are outdated for modern technology. Large documents cannot be quickly sent around the globe and it is extremely difficult to search archived documents for specific content. In a first step towards electronic archiving, many organizations implemented TIFF archives. TIFF guarantees reproducibility in the long-term and has an established structure. TIFF is also easy to transmit in a worldwide business environment but is not easily searchable.

A movement then began towards PDF. PDF is a more attractive archiving format than TIFF for a variety of reasons:

- PDF stores structured objects (e.g. text, vector graphics, raster images), allowing for an efficient full-text search in an entire archive. TIFF is a raster format and must first be scanned with an OCR engine (optical character recognition) before it can be searched.

New technologies have paved the way for archiving documents electronically.

PDF has numerous advantages over TIFF format.

- PDF files are more compact and require only a fraction of the memory space of respective TIFF files, often with a better quality. The smaller file size is especially advantageous for electronic file transfer (FTP, e-mail attachment etc.).
- Metadata like title, author, creation date, modification date, subject, keywords, etc. can be embedded in a PDF file. PDF files can be automatically classified based on the metadata, without requiring human intervention.
- Page content in a PDF document is usually device-independent, i.e. it does not depend on a specific raster resolution, color system etc. The pages are first mapped to a raster when they are reproduced for viewing or printing (rendering process). PDF will therefore profit from technological advances in reproduction devices (printers, monitors etc.) even long into the future.

The inventor of the de facto PDF Standard, Adobe Systems, has published seven new versions of their "PDF Reference Manual" during the past 12 years. Each new version has enriched the format with countless new features and has updated some of the older features. It was therefore necessary to define a stable derivative of the PDF format, based on Adobe's proprietary PDF specification, that could be internationally accepted as a Standard for long-term electronic archiving. The result: PDF/A.

➔ The PDF/A Standard

The Goal of PDF/A

ISO 19005-1 defines "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." (from ISO 19005-1). The Standard does not define an archiving strategy or the goals of an archiving system. It identifies a "profile" for electronic documents that ensures the documents can be reproduced in years to come.

A key element to this reproducibility is the requirement for PDF/A documents to be 100 % self-contained. All of the information necessary for displaying the document in the same manner every time is embedded in the file. This includes all visible content like text, raster images, vector graphics, fonts, color information and much more. A PDF/A document however is not permitted to be reliant on any information from direct or indirect external sources, for example links to external image files or font that are not embedded.

PDF/A files are self-contained.

All of the information required to display a document the same way every time is embedded in the file.

PDF/A is currently based on the PDF Reference Version 1.4.

PDF vs PDF/A

PDF in its native form cannot guarantee long-term reproducibility and not even the "WYSIWYG" principle (what you see is what you get). Certain restrictions and amendments had to be incorporated into the Standard. To be accepted, PDF/A needed to be based on an existing version of the PDF Reference and not on anticipated functionality in a future version. The ISO TC 171 chose the Adobe PDF Reference 1.4, which Adobe implemented in Acrobat 5, as the basis for the Standard. The ISO Standard states that PDF/A "shall adhere to all requirements of PDF Reference as modified by this part of ISO 19005". The Standard itself identifies only differences with respect to the PDF Reference. In order to fully understand PDF/A, you have to also understand the PDF Reference 1.4.

Certain functionality allowed in PDF 1.4 has been specifically excluded from PDF/A, for example transparency and sound and movie actions. There are also elements described in the PDF Reference 1.4 that are not mandatory. PDF/A on the other hand requires these elements to be implemented, for example embedded fonts. In short, PDF/A is based on the PDF Reference 1.4, with specific features being either mandatory, recommended, restricted, or prohibited.

The PDF/A, A-1a, A-1b, A-2 "Babylon"

PDF/A has been established as a row of standards with several parts. Currently only PDF/A-1 (Part 1) has been approved. PDF/A-1 is further subdivided into two levels of compliance: PDF/A-1a and PDF/A-1b.

PDF/A-1a (referred to as Level A Conformance) denotes full compliance with the currently approved PDF/A Standard (ISO 19005-1): Part 1.

There are different levels of PDF/A.

PDF/A-1a denotes full compliance with the currently approved PDF/A Standard.

There is also a "minimal compliance" level for PDF/A: PDF/A-1b (referred to as Level B Conformance). PDF/A-1b requirements are meant to ensure that the rendered visual appearance of the file is reproducible over the long-term.

PDF/A-1a and PDF/A-1b differ primarily with respect to text extraction.

- PDF/A-1a ensures the preservation of a document's logical structure and content text stream in natural reading order. The text extraction is especially important when the document must be displayed on a mobile device (for example a PDA) or other devices in accordance with Section 508 of the US Rehabilitation Act. In such cases the text must be reorganized on the limited screen size (re-flow). This feature is also known as "Tagged PDFs".
- PDF/A-1b ensures that the text (and additional content) can be correctly displayed (e.g. on a computer monitor), but does not guarantee that extracted text will be legible or comprehensible. It therefore does not guarantee compliance with Section 508.

The difference between PDF/A-1a and -1b has no impact for scanned documents, provided the files have not been enhanced by means of OCR for searching.

A new part to the standard, ISO 19005-2, Part-2 (PDF/A-2), is currently being worked on by the Technical Committee. PDF/A-2 will address some of the new feature added with versions 1.5, 1.6 and 1.7 of the PDF Reference. PDF/A-2 should be backwards compatible, i.e. all valid PDF/A-1 documents should also be compliant with PDF/A-2. However PDF/A-2 compliant files will not necessarily be PDF/A-1 compliant.

PDF/A- 1b does not guarantee full text extraction functionality (as required by Section 508 of the US Rehabilitation Act).

PDF/A-2 is currently being worked on and is taking into account the PDF Reference Versions 1.5, 1.6 and 1.7

➔ Using the Standard

Obtaining a Copy

The ISO 19005-1 Standard can be purchased from the [ISO website](#). Copies can be ordered in paper or in PDF format and, like all other ISO Standards, they are protected by copyright. For this reason, publishing freely available version on the internet is illegal. The Standard is currently only available in English.

Who should read the PDF/A Standard

PDF/A format is meant to support and enhance a good archiving strategy. The Standard itself is quite technical and can only be fully understood by experts with a fundamental knowledge in page description languages like PostScript and PDF. The main standard is fairly small, however the volume of related documents is enormous. The PDF Reference alone contains almost 1000 pages, excluding the additional reference documents like font formats, XML specification, compression formats, RFCs etc. In addition, PDF/A alone does not guarantee long-term archiving. A good approach is to enlist an expert who will help you understand the requirements of PDF/A, determine how to implement PDF/A into your archiving strategy, and explain what steps you need to take to ensure your overall archiving goals can be met.

The PDF/A Standard (ISO 19005-1) can be obtained at: www.iso.org.

PDF/A is complex: alone the PDF Reference 1.4 has almost 1000 pages.

What tools are available?

Tools for creating, processing and validating PDF/A documents have been on the market since mid 2006. Adobe itself has integrated respective functions in Version 8 of the Adobe Acrobat, released in the fall of 2006. Even Microsoft has made available a separate downloadable plug-in for their new Office 2007 package, allowing for the creation of PDF/A compliant files directly from Office products. Due to the number of products that have already appeared for creating PDF/A, it has become extremely important to properly verify if the PDF/A documents are fully compliant with the ISO standard.

PDF/A requires a complete solution

PDF/A is only part of a complete archiving solution. PDF/A alone does not guarantee long-term archiving and it does not guarantee that information will be displayed as desired. PDF/A also does not claim that a PDF/A-based archive is always the best solution. However, if you decide to use PDF/A, then PDF/A defines a set of requirements that make long-term archiving possible.

Other aspects that must be taken into account when implementing a PDF/A-compliant archive include, for example, corporate standards and procedures, reliable data sources, reliable fonts, quality management and special individual requirements. The migration of current paper- or TIFF-based archives to PDF/A compliant archives is not an insignificant task and must be well planned.

PDF/A is only one element of a complete archiving strategy.

PDF/A alone cannot guarantee long-term archiving, but it makes long-term archiving possible.

→ Summary

PDF/A as the new archiving standard

PDF/A is expected to establish itself as the new electronic archiving standard. PDF is prevalent in public and private sectors worldwide and is already an accepted archiving format in countless markets. The PDF/A Standard will help ensure that users get the guarantee of long-term reproducibility.

The establishment of the PDF/A Standard will probably (and it should) also have an impact on the future development of PDF itself. Adobe will continue to improve its PDF offerings and add new technology. Examples of these are 3D and XFA for dynamic PDF forms. The Standard will therefore come under further pressure, because a principle concept behind standards, and especially an archiving standard, is that they remain constant and don't regularly change.

PDF/A is expected to quickly establish itself as a preferred electronic archiving format.

PDF/A tools are readily available. Verify the quality of the tool and the expertise of the supplier.

How will the market react?

Don't expect the market to be flooded with PDF/A products in the short term. Understanding the technology behind PDF/A requires considerable knowledge. In addition, users have higher quality expectations for standards-compliant software. The first tools have been on the market since mid 2006. In demand are tools for creating and validating PDF/A compliant documents, as well as simple conversions of existing PDF files into compliant PDF/A files.

The appearance of the first professional PDF/A tools has initiated processes for implementing PDF/A compliant archives. One cannot however expect too much functionality too quickly. Plan on only the more restrictive PDF/A-1b format being readily available, with the full functionality of PDF/A-1a coming later. Expect also to find numerous products that claim to support PDF/A, but don't really. Expertise for evaluations and serious suppliers will especially be in demand during the market introduction phase.

Hot air or a long-term strategy?

PDF/A is not being viewed as just "hot air". The drive towards PDF-based archives has been going on for years, and PDF is already an established archiving format. The PDF/A Standard will help ensure the long-term preservation of the electronic files. With Microsoft now supporting the direct generation of PDF/A from their new Office products, the signal is loud and clear. PDF/A, internationally accepted, is here to stay.

