The PDF – we can't do without it.

Document formats, ISO standards, long-term archiving

Not all document formats have managed to become a standard and nowhere near enough deliver what they promise. So what can or should we expect from a format? What are the most important quality characteristics of a format and how can we find out whether a format is 'good'? Where do the dangers lurk and how can we avoid them? The choice of format is very important, especially in the field of long-term archiving. Documents that have been archived for several years in an unsuitable format or in poor quality can have serious consequences.

PDF and the ISO standards

When it comes to archiving, one format stands out from the rest: PDF, the native format of Adobe Acrobat. Originally designed for the exchange of documents irrespective of platform or software, PDF has firmly established itself over the past 21 years and is now implemented by almost every software house that generates electronic documents. But the increasing use of PDF has also raised fears of a dependency on Adobe. To overcome this, Adobe and some users and industrial enterprises have committed themselves to making PDF the industry standard and developing it through the ISO boards. The first result of this initiative was the publication of the ISO 32000-1 standard in 2008, based on Adobe's PDF 1.7 version. ISO is currently working on the PDF 2.0 version.

PDF's range of functions, which has grown considerably over the years, is not suitable or necessary for every area of application. That is why ISO has developed a number of substandards based on the PDF standard and specifically tailored to the most important applications.

www.pdf-tools.com

Dr. Hans Bärfuss is founder and CEO of PDF Tools AG, an internationally successful software development and distribution company. As a delegate of the Swiss Association for Standardization (SNV) at ISO, he helps to standardize file formats and digital signatures. Dr. Bärfuss is one of the founders and initiators of the PDF/A Competence Center, an association that aims to promote and raise awareness of the ISO standard for PDF and PDF/A, and is chairman of the Swiss Chapter. He gives numerous lectures at conferences and seminars, and publishes articles on the subject of digital documents.

Standard	Purpose
PDF/X	Exchanging print templates: the print result must be predictable
PDF/A	Long-term electronic archiving: see text
PDF/E	Engineering: interactive 3D models for design drawings and production documents
PDF/VT	Variable data printing and transactional printing: large volumes with caching and streaming
PDF/UA	Universal access: barrier-free, with operating aids such as screen readers

Fig. 1: the main PDF substandards

The PDF/X, PDF/A, PDF/E, PDF/VT and PDF/UA standards are not separate file formats. Instead, they are based on the overriding PDF 1.7 standard, defining the requirements and prohibitions for the intended use in each case. These rules limit PDF 1.7's range of functions accordingly and form a defined subset.

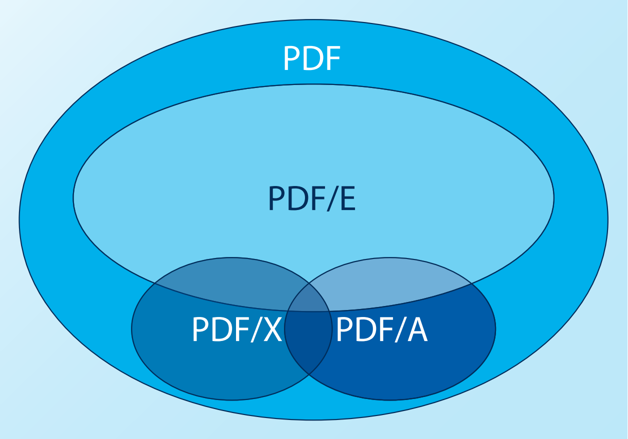

Fig. 2: PDF substandards define a subset of the range of functions

PDF for long-term electronic archiving

Although PDF was developed for the exchange of documents irrespective of the operating system or the creator/viewer's software, the format on its own is not suitable for archiving. That is why PDF/A was developed.

PDF/A is the name given to a range of standards that describe the characteristics of PDF files for the long-term electronic archiving of documents. The purpose of the standards is to ensure that the archived documents remain accessible in the long term and that they are self-documenting, unambiguous, searchable, and described with metadata. For these reasons, encryption is prohibited, fonts and color profiles must be embedded, alternative and dynamic content must be removed, and references to external content must be removed.

A PDF file can also correspond to several substandards at once. This is particularly important when archiving PDF/X and PDF/VT-compliant files that must also meet the PDF/A standard. The range of functions of a file that requires compliance with several standards will therefore be limited to the number of functions offered by the individual standards.

Format vs. format

PDF is a very capable format, but by no means an allrounder. A large number of formats have been optimized for specific uses (see Fig. 3).

When it comes to describing the pros and cons of file formats, many authors automatically look to PDF and publish informed – and sometimes less informed – comparisons between these formats and PDF. However, to make the comparison as fair as possible, we should not lose sight of the intended use. Comparisons make sense only if the areas of application overlap, which brings us to the most frequently discussed areas.

File format	Created for
TIFF	Exchange format for raster images (scanning, archiving)
PostScript/ PCL	Page description language for printing
PDF	Exchange of documents with fixed layout and interactive elements
AFP	Transactional printing with variable data
Office formats	Proprietary format for work documents within the company
OOXML	XML-based Open Office format of Microsoft
ODF	Open Office format in competition with Microsoft
XML	Storage of object-oriented data with reference to a schema
XMP	Extensible metadata format based on XML
XPS	XML-based page description language of Microsoft
EPUB	Format for publishing documents online
PRC	Storage of 3D data and attributes
Multimedia	Numerous formats for audio and video streaming

Fig. 3: file formats and intended use

Incoming mail

When it comes to images in scanned documents, TIFF is neither better nor worse than PDF. The advantages of PDF become apparent only if there are additional requirements that go beyond the mere representation of the page. In addition, a TIFF archive should not be migrated blindly to a PDF archive if no added value compensates for the effort. The criteria for a migration from TIFF to PDF are shown in Fig. 4.

☑ Embedded OCR text

☑ Scanned and digitally generated

☑ Embedded digital signature

☑ Modern compression procedure

☑ Standardized comments (XFDF)

☑ Embedded metadata (XMP)

Fig. 4: criteria for the migration from TIFF to PDF/A

Work documents

Office formats are used most frequently for work documents. OOXML is the native format of Word, Excel and PowerPoint. However, it is not possible to make a fair comparison between OOXML and PDF. OOXML has been designed to create and edit documents. It contains structure information and makes readable text available (copy/paste). To describe PDF as an author format or even as suitable for editing would be rather bold. On the other hand, as a format PDF scores highly with its fixed layout and its suitability for archiving, which OOXML is definitely not suitable for, even though some users would like to think so.

The XPS and PDF/A subsets, however, are comparable and have many things in common, such as static content. The motive behind developing XPS was to eliminate incompatibilities between the document format and the operating system. For instance, the XPS graphics model is the same as the new WPF graphics subsystem in Vista. More precisely, XPS elements are a subset of XAML, the description language for documents and user interfaces contained in WPF.

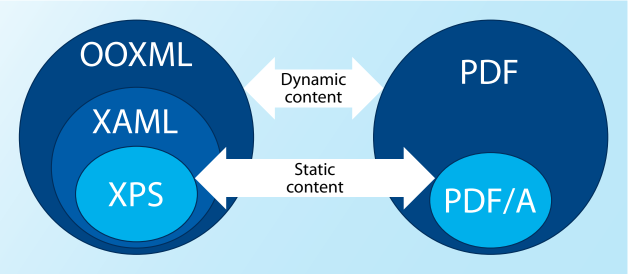

Fig. 5: comparison of Office formats with PDF

This example in particular shows the importance of the dispersion of a format. PDF/A is now the de facto standard for archiving; in contrast, XPS is almost unknown. Similar conclusions can be drawn about the Microsoft-independent Office format ODF.

Outgoing mail

Print data streams in PostScript, PCL and AFP format are sometimes archived directly, but more usually after conversion to TIFF (COLD). Over the past years, there has been talk of replacing 'tiffing', as it is sometimes known, with PDF/A conversion. Criteria such as file size (necessary to embed fonts), conversion effort and final quality play an important role. There is no archiving standard for PostScript and PCL, but work is currently underway on an AFP/A standard as an alternative to PDF/A.

These considerations are particularly relevant for companies that want to implement their own archiving solution for outgoing mail. If a company-wide archiving solution is preferred, then PDF/A is usually the obvious choice.

3D data for engineering

An ISO standard called PRC (Product Representation Compact) is available for 3D data that can be manipulated interactively and is enriched with descriptions (e.g. parts lists). PRC can be embedded in PDF and is an important feature of the PDF/E standard. PDF/E-2 is designed in a way that enables compliant files to be archived directly.

Metadata

There are numerous proprietary formats for metadata and often it is stored directly in the archive system. However, it is strongly recommended to use a standard format. ISO offers XMP (Extensible Metadata Platform), which is based on XML; it can be embedded in almost any image format (e. g. JPEG and TIFF) and is an important component of PDF/A. In addition, XML data can be embedded directly in PDF/A-3, such as for electronic invoicing (ZUGFeRD).

Multimedia

Audio and video files can also be archived, of course, either individually or as embedded data streams in PDF and other formats. However, no generally recognized standards for these formats exist at present, although this might change in the future with EPUB, a format used for electronic publication, playing a driving role.

Archiving – is PDF/A the solution?

In a world of electronic documents, PDF/A is certain to fulfill most file format requirements. But this by no means applies to all archived material. For interactive communication tools in particular, such as websites, programs and multimedia content, new standards are required. The ISO experts are therefore not likely to run out of work any time soon.

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