

Whitepaper: Creating PDF/A with PDFlib

The PDF/A formats specified in the ISO 19005 standard strive to provide a consistent and robust subset of PDF which can safely be archived over a long period of time, or used for reliable data exchange in enterprise and government environments. This whitepaper discusses PDFlib features for creating PDF/A output suitable for long-term document archival.

PDF/A-1a and PDF/A-1b. PDF/A, formally the international standard ISO 19005, is targeted at reliable long-time preservation of digital documents. The standard is based on PDF 1.4 and imposes some restrictions regarding the use of color, fonts, annotations, and other elements. There are two flavors of PDF/A-1, both of which can be created and processed with PDFlib:

- ▶ ISO 19005-1 Level B conformance (PDF/A-1b) ensures that the visual appearance of a document is preservable over the long term. Simply put, PDF/A-1b ensures that the document will look the same when it is processed some time in the future.
- ▶ ISO 19005-1 Level A conformance (PDF/A-1a) is based on level B, but adds crucial properties from »Tagged PDF«: it requires structure information and reliable text semantics in order to preserve the document's logical structure and natural reading order. Simply put, PDF/A-1a not only ensures that the document will look the same when it is used in the future, but also that its contents (semantics) can be reliably interpreted and will be accessible to physically impaired users.

When PDF/A (without any conformance level) is mentioned below, both conformance levels are meant. The PDF/A implementation in PDFlib, PDFlib+PDI and the PDFlib Personalization Server (PPS) is based on the ISO 19005-1 standard plus the 2006 corrigendum.

PDF/A requirements and restrictions. PDF/A requires certain PDF features, and prohibits certain other properties. For example, in order to guarantee exact text reproduction all fonts used in a document must be embedded; in order to guarantee exact color reproduction all colors must be specified in a device-independent way. Metadata must be embedded using the XMP format; encryption must not be used.

In addition to these straight-forward requirements, however, PDF/A requires various other PDF features which are more subtle (e.g. certain entries in font data structures), and prohibits some critical structures (e.g. certain combinations of TrueType fonts and encodings). There are many aspects which must be implemented and checked by software developers before they arrive at fully standards-compliant PDF/A products. PDF/A is much more than simply »PDF with fonts embedded«!

PDF/A support in the PDFlib product suite. PDFlib provides application developers with a toolkit which allows the following PDF/A-related operations:

- ▶ Create PDF/A from scratch, e.g. based on text from a database
- ▶ Convert raster images (e.g. scans) to PDF/A
- ▶ Process existing PDF/A documents, e.g. merge or split PDF documents
- ▶ Create PDF/A-1a with structure information (Tagged PDF)

All of these operations can be implemented with simple PDFlib function calls. Sample code for a variety of programming languages and development environments is provided with the PDFlib distribution. Due to the significant overlap between PDF/A and the PDF/X standard (ISO 15930) for the graphics arts industry, the PDF/A support in PDFlib took advantage of the fact that we've been supporting various flavors of PDF/X for several years.

Creating PDF/A-compliant output. Creating PDF/A-compliant output with PDFlib is achieved by the following means:

- ▶ PDFlib will automatically take care of several formal settings for PDF/A, such as PDF version number and required XMP conformance entries.
- ▶ The PDFlib client program must explicitly use certain function calls and options (e.g. for font embedding).
- ▶ The PDFlib client program must refrain from using certain function calls and option settings (e.g. encryption).

If the PDFlib client program obeys to these rules, valid PDF/A output is guaranteed. If PDFlib detects a violation of the PDF/A creation rules it will throw an exception which must be handled by the application. No PDF output will be created in case of an error; there is no danger of creating non-compliant output if an error occurs. Details of required and prohibited operations are discussed in the PDFlib documentation.

Output intents and ICC profiles made easy. In order to maintain consistent color reproduction, PDF/A requires the use of device-independent color, usually achieved via ICC profiles or CIE Lab color specifications. While these concepts are widely used in the graphics arts industry, enterprise PDF developers are not necessarily familiar with color management concepts. In this situation PDFlib makes it easy to create device-independent output regardless of the source of input data:

- ▶ PDF/A output can be created with or without an output intent ICC profile.
- ▶ In the common case of black text PDFlib will automatically choose an appropriate color space (Lab or DeviceGray) depending on whether or not an ICC output intent profile has been specified.
- ▶ External ICC profiles and profiles embedded in images allow fine-grain color control.
- ▶ ICC profiles for common application scenarios are provided with the PDFlib distribution so that valid PDF/A output can be achieved quickly.

Raster images, e.g. TIFF and JPEG, play a vital role in document creation. Scanned paper documents and photographs from digital cameras are common examples of raster image data in document workflows. While in modern workflows raster image data may already be device-independent (usually by means of an embedded ICC color profile) and thus are PDF/A-compatible, legacy image data will in many cases be device-dependent, such as black-and-white or RGB scans without any ICC profile. PDFlib supports both situations:

- ▶ ICC profiles embedded in raster image files will be honored.
- ▶ External ICC profiles can be applied to an image
- ▶ As a fallback solution for legacy data of unknown origin PDFlib contains a builtin sRGB profile.
- ▶ By specifying an output intent ICC profile, device-dependent image data can be used without applying an ICC profile to the images.

The PDFlib documentation discusses PDF/A color strategies for common application scenarios.

XMP made easy. PDF/A mandates the use of XMP metadata for storing information about a document inside the PDF itself. XMP provides a powerful and flexible framework for storing standard and custom metadata. If you already use XMP metadata in your workflow you can create full XMP streams which PDFlib will integrate into the PDF/A output. However, developers who are not familiar with XMP are not required to delve into this topic. PDFlib will create the XMP output required for PDF/A, and will automatically map plain old document info fields to the corresponding XMP constructs. As a result, developers who wish to do so can leverage the power of XMP, while PDFlib's automatic XMP generation will be adequate in situations with simple metadata requirements.

Processing existing PDF/A documents. Additional rules apply when importing pages from existing PDF/A-conforming documents. For example, in Adobe Acrobat it is very easy to combine two PDF/A documents such that the resulting output document is no longer PDF/A-compliant (without any warning). When dealing with existing PDF/A documents, PDFlib+PDI carefully examines the PDF/A properties of all input and output documents to make sure that the output will again be PDF/A-compliant. For additional control the output intent of an imported document can be copied to the output PDF, effectively cloning the PDF/A color properties of an existing document.

Creating PDF/A-1a with Tagged PDF. PDF/A-1a can be regarded as PDF/A-1b plus Tagged PDF: it requires structure information for the document and imposes certain conditions on fonts to make sure that the text can be properly interpreted. As a result, PDF/A-1a documents are fully accessible for users with physical disabilities. In addition to the visual appearance they also preserve the meaning of its contents.

PDFlib's support for PDF/A-1a is based on the features for producing Tagged PDF: each content item can be placed at a particular location in the document's structure tree; content items which are not relevant for the document structure (e.g. headers and footers, pagination) can be tagged as artifacts which means that they will be ignored when repurposing the document (e.g. when the document is read aloud by software, or converted to some other format). Note that you will need detailed knowledge about the document's logical structure in order to create Tagged PDF. PDFlib will take care of the PDF-related details, but it cannot infer the document structure from its contents.

Tagged PDF support has been introduced in PDFlib 6. PDFlib 7 adds the capability to include annotations in the document's structure tree which improves the accessibility of links and other interactive elements.

Based on the existing Tagged PDF support, PDFlib 7 can produce output which conforms to PDF/A-1a. This makes PDFlib the first tool to support this advanced PDF/A level.

Validating PDF/A. When implementing a standards-based workflow it is good practice to deploy tools for validating the results against the standard. With respect to the PDF/A standard there are software tools available which check whether or not a given PDF document complies with the ISO standard.

The Preflight tool included in Acrobat 7 offers a test profile for PDF/A-1b. However, since the release of Acrobat 7.0 predates the final release of the ISO standard for PDF/A, this version implements only a draft version of PDF/A. Starting with Acrobat 7.07 the

Preflight plugin supports the published PDF/A standard. However, there are a few issues with the plugin since it deviates from the PDF/A standard in a few places. PDFlib GmbH has carefully examined several cases of alleged non-conforming PDFlib output together with the developers of the Acrobat Preflight plugin, and established that the output actually conforms to the standard. Preflight's inappropriate messages regarding non-compliant documents will be fixed in a future version.

Similarly, we are actively working with other vendors of PDF/A validation tools to make sure that creators and validators have a common understanding and interpretation of the PDF/A standard.

PDF/A competence center. PDFlib GmbH is actively involved in the PDF/A competence center to increase industry awareness of PDF/A, and achieve multi-vendor compatibility. See www.pdfa.org for more information.

PDFlib GmbH
Tal 40
80331 München, Germany
phone +49/89/29 16 46 86
info@pdflib.com
www.pdflib.com