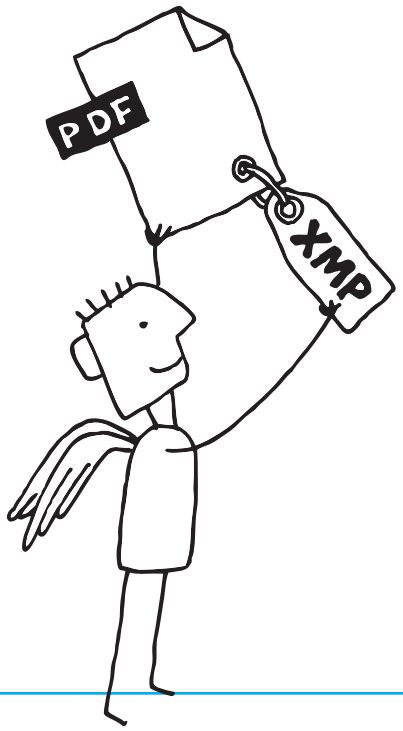
XMP Metadata Support in PDFlib Products



PDFlib Whitepaper

XMP Metadata

The importance of metadata

The term metadata literally means »data about data«. Metadata has been described as the business card of a particular digital document. Metadata often comprises a set of properties, where each property has specific meaning in the context of the document. Some examples for common metadata properties:

- The author of a PDF document.
- The date a PDF document was created or a JPEG image was taken with a camera.
- The name of the photographer who took an image.
- The serial number of a personalized document.
- The GPS position where an image was taken.
- The stockkeeping unit (SKU) of the item described in a document.
- The year of manufacture of the engineering product described in a document.
- The reference number of a document in a legal case.

As an increasing number of publishing, documentation, translation, and other workflows are implemented in a completely digital manner, metadata plays a crucial role for handling digital documents during their lifetime.

Adobe's Extensible Metadata Platform (XMP)

As Adobe recognized the need for a common metadata format which can be used across applications and file formats, they designed the Extensible Metadata Platform (XMP). This is an XML-based format modelled after W3C's RDF (Resource Description Framework) which forms the foundation of the semantic Web initiative. In 2012 XMP has been standardized as ISO 16684-1:2012.

XMP metadata travels with the file, and can be embedded in many common file formats including PDF, TIFF, and JPEG. Metadata properties are grouped in schemas. Each schema is identified by a unique namespace URI and holds an arbitrary number of properties.

The XMP specification includes more than a dozen predefined schemas with hundreds of properties for common document and image characteristics. The most widely used predefined XMP schema is called the Dublin Core, or *dc*. It includes general properties such as *Title, Creator, Subject,* and *Description*. In addition to predefined schemas, custom schemas can be defined to cover company- or industry-specific metadata requirements. The Dublin Core has been standardized as ISO 15836.

XMP for PDF documents has been introduced with Acrobat 5 and PDF 1.4 in 2001. The predecessor of XMP in PDF was formed by simple key/value pairs, so-called document info entries. While document info entries are still supported in PDF 1.7, XMP metadata is a much more powerful concept and allows metadata to survive format conversions, e.g. from scanned TIFF to PDF. Note that document info entries are deprecated in favor of XMP metadata in the PDF 2.0 standard ISO 32000-2.

XMP is implemented in all Adobe publishing products and supported by dozens of independent software vendors. XMP metadata can be displayed and edited in the *File Info/Document Properties* panel in Acrobat *(File, Properties..., Additional metadata...)*, Photoshop, InDesign, and other Adobe applications.

XMP for verticals

XMP is increasingly used by various industries to cover their metadata requirements. Some examples:

- The AdsML consortium creates specifications and processes for the exchange of advertising information and content.
- The International Press Telecommunications Council (IPTC) is a consortium of the world's major news agencies. It develops industry standards for the interchange of news data. It published the »IPTC Core« XMP schema which is widely used for transferring metadata for images and other data types.
- The DICOM standard for exchanging medical images supports the use of PDF and specifies a custom XMP schema for storing patient data, study description, equipment details, and other metadata.
- The Publishing Requirements for Industry Standard Metadata (PRISM) defines a metadata vocabulary for processing magazine, news, catalog, book, and journal content.

XMP mandated by ISO standards

There are various ISO standards which specify PDF subsets for certain application domains, such as the graphic arts industry, archiving, or engineering. Except for the outdated prepress standards PDF/X-1 and PDF/X-3 which have been introduced in 2001 and 2002, all ISO standards for PDF include the use of XMP metadata. XMP is even mandatory except in the base PDF 1.7 standard ISO 32000-1.

XMP support in PDFlib GmbH products

XMP support in the PDFlib product family

Simple XMP support has been introduced in the PDFlib product family in 2004. With the introduction of PDF/A support in 2006 the XMP features were expanded to match the requirements of PDF/A. In particular, automatic synchronization of document info entries to XMP properties (as specified in the PDF/A-1/2/3 crosswalk) was implemented, as well as automatic creation of several internal XMP properties required for PDF/A. As a result, PDFlib users can generate XMP for PDF/A without having to struggle with the internals of the XMP format. Advanced users can directly feed all of the predefined XMP metadata schemas to PDFlib for inclusion in the generated PDF documents. Since PDFlib is available on all relevant operating systems and does not require any third-party products, it brings XMP support to all platforms.

On top of this, PDFlib adds support for XMP extension schemas according to PDF/A-1/2/3. Users can embed extension schema descriptions for custom metadata as required by PDF/A. Since PDFlib validates user-supplied XMP extension schemas and the corresponding descriptions for internal consistency and standards conformance the output is guaranteed to conform to the PDF/A standard.

This feature made PDFlib the first product worldwide to support XMP extension schemas for PDF/A. More details on XMP in PDF/A can be found on www.pdflib.com.

In addition to document-related XMP the PDFlib product family also supports object-level XMP. For example, page-based XMP may carry custom processing information for that page, image-level XMP holds intellectual property information for licensed images, etc.

Searching for XMP metadata with PDFlib TET PDF IFilter

TET PDF IFilter implements Microsoft's IFilter interface and can be used with various Microsoft desktop and enterprise search products, such as Windows Search, Microsoft SharePoint or SQL Server. XMP support in TET PDF IFilter makes it easy to leverage XMP metadata in environments where Microsoft search solutions are deployed.

The advanced metadata implementation in TET PDF IFilter supports the Windows property system for metadata. In addition to page contents it indexes XMP metadata as well as standard or custom document info entries. Metadata indexing can be configured on several levels:

- Document info entries and common XMP properties are mapped to standard Windows properties, e.g. *Title, Subject, Author*.
- TET PDF IFilter adds useful PDF-specific pseudo-properties, e.g. page size, PDF/A conformance level, font lists.
- All relevant predefined XMP properties can be searched, e.g. *dc:rights, xmpRights:UsageTerms, xmp:CreatorTool*.
- Custom (user-defined) XMP properties can be searched, e.g. company-specific classification.
- In addition to document metadata, XMP attached to images can also be indexed, e.g. the name of the photographer of an image or copyright information.

TET PDF IFilter optionally integrates metadata in the indexed raw text. As a result, even full-text search engines without metadata support, e.g. SQL Server, can search for metadata.

Injecting XMP in PDF with PDFlib PLOP and PLOP DS

In addition to various other features including encryption, decryption, optimization, and digital signature, PDFlib PLOP and PLOP DS can insert XMP metadata in PDF documents. This function comes handy in situations where PDF documents do not contain all required metadata properties. It is especially useful in PDF/A workflows since XMP support in PLOP and PLOP DS is PDF/A-aware. For example, custom XMP with extension schemas can be injected in PDF/A documents from workflows which do not support extension schemas.

Extracting XMP from PDF with PDFlib pCOS

The pCOS interface is PDFlib GmbH's method for retrieving all kinds of information from PDF documents. It is integrated in several products. pCOS offers a simple programming method for extracting XMP metadata from PDF documents. XMP metadata is normalized to Unicode so that users don't have to worry about encoding issues.

XMP retrieval with pCOS works regardless of compression, encryption, and PDF object structure. While the XMP package mechanism defined by Adobe allows easy inclusion and retrieval of XMP data packages in various file formats, the PDF format exhibits several subtleties which complicate the issue. For example, PDF documents may contain several update sections which cause multiple instances of an XMP stream to be present in the file, where only one of these instances is relevant. A simple text search for XMP data will likely retrieve the wrong instance; only software which follows the PDF object structure retrieves the correct XMP metadata reliably.

Workflow scenarios which benefit from XMP-based document search

XMP metadata handling can be integrated in diverse scenarios which require searching digital documents. Two typical examples are described below.

Publishing: Creative professionals use Adobe and other publishing software to create documents and metadata interactively. They assign keywords, author name, copyright information and other common XMP properties to documents. They can use Adobe Bridge to search or group documents according to the assigned metadata properties, and are focused on common XMP schemas such as Dublin Core and IPTC.

Technical documentation: A large number of documents is created manually or automatically, and collected in departmental or company-wide collections. These document collections are accessed via common Windows retrieval tools, such as Microsoft SharePoint on server systems, Windows Search on workstations. After attaching TET PDF IFilter to these products users can search for documents based on XMP metadata properties, the actual page contents, or even image properties. While predefined XMP schemas cover the basic requirements, customized XMP schemas can be used in the queries to cover company-specific requirements.

PDFlib GmbH
Franziska-Bilek-Weg 9
80339 München, Germany
support@pdflib.com
www.pdflib.com/knowledge-base/pdfa

PDFlib GmbH is completely focused on PDF technology. Customers worldwide use PDFlib products since 1997. The company closely follows development and market trends, such as ISO standards for PDF. PDFlib GmbH products are distributed all over the world with major markets in North America, Europe, and Japan.

Founded in 2006 as PDF/A Competence Center, in 2011 the PDF association broadened its scope to cover all aspects of PDF technology. Today, it provides an industry meeting-place, and a platform for members to exercise thought-leadership in the community.

- Developers use the PDF Association to share knowledge and experience with PDF technology.
- Decision-makers use the PDF Association to learn about the role and capabilities of PDF and PDF's subset standards in ECM and other electronic document applications.
- End-users benefit from improved reliability, quality and functionality and interoperability in their experience of electronic documents.