

Whitepaper: XMP Metadata Support in PDFlib Products

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 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. Adobe makes the XMP specification freely available, and offers an open-source XMP toolkit for software developers.

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.

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, which served as the sole carrier of metadata prior to the introduction of XMP. While document info entries are still supported in Acrobat and PDF, XMP metadata is a much more powerful concept and allows metadata to survive format conversions, e.g. from scanned TIFF to PDF.

XMP is implemented in all Adobe publishing products and supported by dozens of independent software vendors and user groups. Adobe Bridge, part of the Creative Suite, deals with XMP metadata in various file formats. 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. While the File Info panel groups metadata properties according to the predefined XMP schemas, custom panels can be defined to tailor metadata display and editable fields according to the requirements of various application domains.

XMP for verticals. XMP is increasingly used by industry groups 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 an industry group established by news organizations. It develops industry standards for the interchange of news data. It published the »IPTC Core« schema for XMP which is widely used for transferring metadata for images and other news items.
- ▶ 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 several existing and planned ISO standards which specify PDF subsets for certain application domains, such as the graphic arts industry, archiving, or engineering. Except for the prepress standards PDF/X-1 and X-3 which have been introduced in 2001 and 2002, all ISO standards for PDF include the use of XMP metadata (even mandatory in most cases except ISO 32000):

- ▶ PDF/A-1 in ISO 19005-1 (published in 2005): »Electronic document file format for long-term preservation – Use of PDF 1.4«. PDF/A-1 requires XMP for identifying conforming files and supports custom metadata through XMP extension schemas. A formal description of all extension schemas must be embedded in PDF/A to maximize the future use of custom metadata. PDF/A-1 allows the use of document info entries, but requires synchronization between common PDF document info entries and certain predefined XMP properties to allow pure XMP-based workflows. The standard defines this »crosswalk« between document info entries and XMP properties. XMP support in PDF/A-1 is based on the XMP 2004 specification.
- ▶ PDF/E in ISO 24517-1 (expected for publication in 2008): »Engineering document format using PDF – Use of PDF 1.6«. XMP support in PDF/E is almost identical to PDF/A-1, except that it is based on the newer XMP 2005 specification.
- ▶ PDF/X-4 in ISO 15930-7 (published in 2008): »Complete exchange of printing data (PDF/X-4) and partial exchange of printing data with external profile reference (PDF/X-4p) using PDF 1.6«. Similar to PDF/A-1, XMP is required to express standards conformance in PDF/X-4. Document info entries may be used in PDF/X-4, but must be synchronized with corresponding XMP entries. XMP extension schemas for custom metadata are allowed. However, unlike PDF/A-1 these can be used without embedding a formal description. XMP support in PDF/X-4 is based on the XMP 2005 specification.
- ▶ PDF/X-2 in ISO 15930-5 (published in 2003) and PDF/X-5 in ISO 15930-8 (expected for publication in 2008): »Partial exchange of printing data using PDF 1.6 (PDF/X-5)«. PDF/X-2 and X-5 documents reference other PDF/X documents, where the target of such a reference is identified by using various XMP entries. This makes XMP a crucial component of PDF/X-2 and X-5.
- ▶ ISO 32000 (expected for publication in 2008): »Document management – Portable document format – PDF 1.7«. ISO 32000 is the standardized version of PDF 1.7. The technical content is identical to PDF 1.7 (the file format of Acrobat 8) which fully supports XMP metadata.

The Dublin Core, one of the most common predefined XMP metadata schemas has been standardized as ISO 15836 (published in 2003): »Information and documentation — The Dublin Core metadata element set«.

XMP support in the PDFlib product suite. Simple XMP support has been introduced in the PDFlib product family in 2004. With PDF/A-1 support in PDFlib 7 (released in 2006) the XMP features were expanded to match the requirements of PDF/A-1. In particular, automatic synchronization of document info entries to XMP properties (as specified in the PDF/A-1 crosswalk) was implemented, as well as automatic creation of several internal XMP properties required for PDF/A-1. As a result, PDFlib users can generate XMP for PDF/A-1 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 7.0.3 adds support for XMP extension schemas according to PDF/A-1. Users can embed »extension schema container schemas« for custom metadata in PDF/A-1. Since PDFlib fully validates user-supplied XMP extension schemas for internal consistency and standards conformance, the output is guaranteed to conform to the PDF/A-1 standard.

This feature makes PDFlib 7.0.3 the first product worldwide to support XMP extension schemas for PDF/A-1. As a result of PDFlib GmbH's participation in the PDF/A Competence Center, all PDF/A activities are closely coordinated with other vendors of PDF/A software to ensure the highest possible degree of standards conformance and adherence to industry practises.

Since XMP validation is active even when no PDF/A output is created, all XMP users benefit from the improved XMP support in PDFlib 7.0.3.

More details on XMP in PDF/A, plus an online validator for XMP extension schemas can be found on www.pdflib.com.

Injecting XMP in PDF with PDFlib PLOP 3.1. In addition to various other features including encryption, decryption, optimization, and digital signature, PDFlib PLOP can insert XMP metadata in existing PDF documents. This function comes handy in situations where existing PDF documents do not contain all required metadata properties. It is especially useful in PDF/A workflows since XMP support in PLOP 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 any kind of information from PDF documents. It is available as a stand-alone product, and also integrated in all other 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 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, PDF documents exhibit 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 the XMP block will likely retrieve the wrong instance; only software which carefully follows the PDF object structure will successfully retrieve XMP metadata in all cases. This is the reason why Adobe's free XMP Toolkit does not fully support XMP retrieval from PDF, while it does support XMP in other file formats such as TIFF and JPEG.

Searching for XMP metadata with PDFlib TET PDF IFilter. TET PDF IFilter is the latest product released by PDFlib GmbH. It implements Microsoft's IFilter interface and can be used with various Microsoft and third-party desktop and enterprise

search products, such as Windows Desktop Search (WDS), Office SharePoint Server (MOSS), Indexing Server, or SQL Server. XMP support in TET PDF IFilter makes it very 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 items.

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.

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 Office SharePoint Server (MOSS) on server systems, Windows Desktop Search (WDS) on workstations, or other retrieval products. After attaching TET PDF IFilter to these products users can search for documents based on both XMP metadata properties and the actual page contents. 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
phone +49 • 89 • 452 33 84-0
info@pdflib.com
www.pdflib.com

