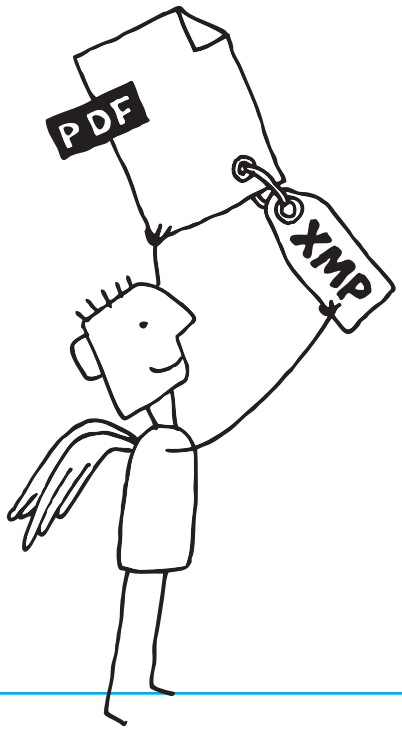
XMP-Metadaten in PDFlib-Produkten



PDFlib Whitepaper

XMP-Metadaten

Die Bedeutung von Metadaten

Der Ausdruck »Metadaten« bedeutet wörtlich »Daten über Daten«. Metadaten stellen eine Art Visitenkarte für ein digitales Dokument dar. Sie enthalten in der Regel einen Satz von Eigenschaften (Properties), wobei jede dieser Properties über eine bestimmte Bedeutung im Dokumentkontext verfügt. Beispiele für gängige Metadaten-Properties sind:

- Der Verfasser eines PDF-Dokuments.
- Das Erstellungsdatum des PDF-Dokuments oder das Datum, zu dem ein JPEG-Bild mit der Kamera aufgenommen wurde.
- Der Name des Fotografen eines Bildes.
- Die Seriennummer eines personalisierten Dokuments.
- Die GPS-Position, an der ein Bild aufgenommen wurde.
- Die Artikelnummer eines Artikels, der im Dokument beschrieben wird.
- Das Herstellungsdatum eines technischen Produkts, mit dem sich das Dokument befasst.
- Das Aktenzeichen eines Dokuments in einem Gerichtsverfahren.

Mit wachsender Anzahl von vollständig digitalisierten Arbeitsabläufen, z.B. in den Bereichen Publishing, Dokumentation, Übersetzung oder anderen Abläufen, spielen Metadaten eine immer wichtigere Rolle im Umgang mit digitalen Dokumenten über deren gesamte Lebensdauer.

Extensible Metadata Platform (XMP) von Adobe

Aus dem allgemeinen Bedarf nach einem gemeinsamen Metadatenformat heraus, das über verschiedenste Anwendungen und Dateiformate hinweg verwendbar ist, konzipierte Adobe das Format *Extensible Metadata Platform* (XMP). Es basiert auf XML und wurde nach dem vom W3C entwickelten RDF *(Resource Description Framework)* entworfen, das die Grundlage der Semantic Web Initiative bildet. XMP wurde 2012 als ISO 16684-1:2012 standardisiert.

XMP-Metadaten wandern mit der Datei mit und lassen sich in viele gängige Dateiformate inklusive PDF, TIFF und JPEG einbetten. Die in den Metadaten enthaltenen Properties werden zu Schemas zusammengefasst. Jedes Schema enthält eine beliebige Anzahl von Properties und wird durch einen eindeutigen URI für den zugehörigen Namensraum identifiziert.

Die XMP-Spezifikation umfasst mehr als ein Dutzend vordefinierte Schemas mit Hunderten von Properties für übliche Dokument- und Bildmerkmale. Das am häufigsten verwendete XMP-Schema heißt Dublin Core oder *dc* und enthält allgemeine Properties wie *Title, Creator, Subject* und *Description*. Außerdem können benutzerdefinierte Schemas erstellt werden, wenn firmen- oder branchenspezifische Metadaten benötigt werden. Dublin Core wurde als ISO 15836 standardisiert.

XMP für PDF-Dokumente wurde mit Acrobat 5 und PDF 1.4 im Jahre 2001 eingeführt. Der Vorgänger von XMP in PDF bestand aus einfachen Schlüssel/Wert-Paaren, so genannten Dokument-Infofeldern, die vor der Einführung von XMP als einziger Träger von Metadaten dienten. Dokument-Infofelder werden in Acrobat und PDF zwar nach wie vor unterstützt, XMP-Metadaten sind aber ein weit leistungsfähigeres Konzept und gewährleisten zudem, dass Metadaten auch bei Formatkonvertierungen, z.B. von eingescanntem TIFF nach PDF, erhalten bleiben. Beachten Sie, dass im kommenden Standard ISO 32000-2 für PDF 2.0 die herkömmlichen Dokument-Infofelder als hinfällig erklärt und durch XMP-Metadaten abgelöst werden.

XMP ist in allen Publishing-Produkten von Adobe implementiert und wird von zahlreichen unabhängigen Software-Anbietern und Anwendervereinigungen unterstützt. Adobe Bridge, das mit der Creative Suite ausgeliefert wird, verarbeitet XMP-Metadaten in verschiedenen Dateiformaten. In Acrobat *(Datei, Dokumenteigenschaften..., Zusätzliche Metadaten...),* in Photoshop, InDesign und anderen Adobe-Anwendungen erfolgt die Anzeige und Bearbeitung von XMP-Metadaten im Panel *Dokumenteigenschaften* bzw. *Dateiinformationen*.

XMP für verschiedene Industriezweige

Zur Umsetzung der jeweiligen Metadaten-Anforderungen kommt in verschiedenen Industriezweigen zunehmend XMP zum Einsatz. Einige Beispiele:

- Das AdsML-Konsortium erstellt Spezifikationen und Prozessabläufe für den Austausch von Anzeigen und -inhalten.
- Das International Press Telecommunications Council (IPTC) ist der Weltverband von Nachrichtenagenturen und Zeitungen. Er entwickelt Industriestandards für den Austausch von Nachrichten und publiziert das weit verbreitete XMP-Schema »IPTC Core«, das zur Übertragung von Metadaten für Bilder und andere Datentypen verwendet wird.
- Der DICOM-Standard zum Austausch digitaler Bilder in der Medizin unterstützt PDF und spezifiziert ein benutzerdefiniertes XMP-Schema zur Speicherung von Patientendaten, Befunden, Geräteparametern und anderen Metadaten.
- Die Publishing Requirements for Industry Standard Metadata (PRISM) definieren ein Metadatenvokabular zur Verarbeitung der Inhalte von Zeitschriften, Nachrichten, Katalogen, Büchern und Zeitungen.

XMP in ISO-Standards

Mehrere veröffentlichte oder geplante ISO-Standards spezifizieren PDF-Teilmengen für bestimmte Anwendungsgebiete wie grafische Industrie, Archivierung oder Ingenieurwesen. Mit Ausnahme der Prepress-Standards PDF/X-1 und PDF/X-3, die in den Jahren 2001 bzw. 2002 eingeführt wurden, berücksichtigen alle ISO-Standards für PDF die Verwendung von XMP-Metadaten (diese sind meist obligatorisch, außer bei ISO 32000). Sofern nicht anders beschrieben, basieren alle Standards auf XMP 2005:

- PDF/A-1 in ISO 19005-1 (veröffentlicht 2005): »Electronic document file format for long-term preservation – Use of PDF 1.4«. PDF/A-1 benötigt XMP zur Identifikation konformer Dateien und unterstützt benutzerdefinierte Metadaten mittels XMP-Extension-Schemas. Die XMP-Unterstützung in PDF/A-1 basiert auf der Spezifikation XMP 2004.
- PDF/A-2 in ISO 19005-2 (veröffentlicht 2011): »Electronic document file format for long-term preservation – Part 2: Use of ISO 32000-1 (PDF/A-2)«
- PDF/A-3 in ISO 19005-3 (veröffentlicht 2012): »Electronic document file format for long-term preservation – Part 3: Use of ISO 32000-1 with support for embedded files (PDF/A-3)«
- PDF/E-1 in ISO 24517-1 (veröffentlicht 2008): »Engineering document format using PDF – Use of PDF 1.6«. XMP-Unterstützung in PDF/E entspricht im Wesentlichen derjenigen in PDF/A-1, basiert aber auf der neueren Spezifikation XMP 2005.
- PDF/X-4 in ISO 15930-7 (veröffentlicht 2008, überarbeitet 2010): »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«. Ähnlich wie bei PDF/A-1 wird XMP benötigt, um die Konformität zum Standard PDF/X-4 auszudrücken.
- PDF/X-5 in ISO 15930-8 (veröffentlicht 2008, überarbeitet 2010): »Partial exchange of printing data using PDF 1.6 (PDF/X-5)«. PDF/X-5-Dokumente verweisen auf andere PDF/X-Dokumente, wobei das Verweisziel durch verschiedene XMP-Einträge beschrieben wird. XMP wird damit zu einem wesentlichen Bestandteil von PDF/X-5.
- ISO 32000-1 (veröffentlicht 2008): »Document management – Portable document format – PDF 1.7«. ISO 32000 ist die standardisierte Fassung von PDF 1.7. Der technische Inhalt entspricht PDF 1.7 (dem Dateiformat von Acrobat 8), das vollständige Unterstützung von XMP-Metadaten bietet.

- PDF/VT in ISO 16612 (veröffentlicht 2010): »Variable data exchange – Part 2: Using PDF/X-4 and PDF /X-5 (PDF/VT-1 and PDF/VT-2)«
- PDF/UA-1 in ISO 14289-1 (veröffentlicht 2012): »Document management applications – Electronic document file format enhancement for accessibility - Part 1: Use of ISO 32000-1 (PDF/UA-1)«
- ISO 32000-2 (Standardisierungsprozess mit Stand 2016 noch nicht abgeschlossen): »Document management – Portable Document Format – PDF 2.0«. In PDF 2.0 werden die herkömmlichen Dokument-Infofelder durch XMP-Metadaten abgelöst (außer *CreationDate* und *ModDate)*.

XMP-Unterstützung in Produkten der PDFlib GmbH

XMP-Unterstützung in der PDFlib-Produktfamilie

Einfache XMP-Unterstützung wurde bereits 2004 in die PDFlib-Produktfamilie integriert. Mit der Unterstützung von PDF/A ab 2006 wurde die XMP-Funktionalität erweitert, um den Anforderungen von PDF/A zu genügen. Implementiert wurde insbesondere der automatische Abgleich von Dokument-Infofeldern mit entsprechenden XMP-Properties (festgelegt im PDF/A-Crosswalk) sowie die automatische Erstellung verschiedener interner XMP-Properties, die für PDF/A erforderlich sind. PDFlib-Benutzer können seitdem XMP für PDF/A erstellen, ohne sich um die Einzelheiten des XMP-Formats kümmern zu müssen. Fortgeschrittene Benutzer können alle vordefinierten XMP-Metadatenschemas an PDFlib übergeben, um sie in die generierten PDF-Dokumente einzubetten. Da PDFlib auf allen relevanten Betriebssystemen verfügbar ist und keine Produkte von Fremdanbietern voraussetzt, ist XMP-Unterstützung auf allen Plattformen gewährleistet.

Darüber hinaus bietet PDFlib Unterstützung für XMP-Extension-Schemas gemäß PDF/A. Benutzer können die von PDF/A vorgeschriebene Beschreibung von Extension-Schemas für benutzerdefinierte Metadaten einbetten. Da PDFlib externe XMP-Extension-Schemas vollständig auf interne Konsistenz und Konformität zum Standard überprüft, ist die Ausgabe garantiert PDF/A-konform.

PDFlib war damit das weltweit erste Produkt, das XMP-Extension-Schemas für PDF/A unterstützt. Weitere Informationen zu XMP in PDF/A finden Sie unter www.pdflib.com. Außer XMP auf Dokumentebene unterstützt die PDFlib-Produktfamilie auch XMP auf Objektebene. Seitenbasiertes XMP kann zum Beispiel benutzerdefinierte Angaben zur Verarbeitung dieser Seite, bildbezogenes XMP Angaben zu Rechteinhabern lizenzierter Bilder enthalten u.ä.

Suche nach XMP-Metadaten mit PDFlib TET PDF IFilter

TET PDF IFilter implementiert die IFilter-Schnittstelle von Microsoft und kann mit verschiedenen von Microsoft und anderen Herstellern angebotenen Produkten zur computer- oder unternehmensweiten Suche eingesetzt werden, zum Beispiel mit Windows Search, Microsoft SharePoint, FAST Search oder SQL Server. Die XMP-Unterstützung in TET PDF IFilter ermöglicht einen bequemen Umgang mit XMP-Metadaten in Umgebungen, in denen Microsoft-Lösungen zur Textsuche zum Einsatz kommen.

Die leistungsfähige Metadaten-Implementierung von TET PDF IFilter unterstützt das Property-System von Windows für Metadaten. Neben Seiteninhalten werden auch XMP-Metadaten sowie Standard- und benutzerdefinierte Dokument-Infofelder indexiert. Die Indexierung der Metadaten lässt sich auf verschiedenen Ebenen konfigurieren:

- Dokument-Infofelder und gängige XMP-Properties werden auf Standard-Windows-Properties wie *Title, Subject, Author* abgebildet.
- TET PDF IFilter ergänzt nützliche PDF-spezifische Pseudo-Properties wie Seitengröße, PDF/A-Konformitätslevel oder Fontlisten.
- Nach allen relevanten vordefinierten XMP-Properties kann gesucht werden, z.B. nach *dc:rights, xmpRights:UsageTerms* oder *xmp:CreatorTool*.
- Die Suche umfasst auch benutzerdefinierte XMP-Properties, z.B. firmenspezifische Klassifizierungen.
- Zusätzlich zu Dokument-Metadaten werden auch XMP-Metadaten für Bilder indexiert, z.B. der Name des Fotografen eines Bildes oder Copyright-Angaben.

TET PDF IFilter bietet optional die Möglichkeit, Metadaten in den indexierten Rohtext zu integrieren. Damit können auch Volltextsuchmaschinen ohne Metadaten-Unterstützung (z.B. SQL Server) nach Metadaten suchen.

Einfügen von XMP in PDF mit PDFlib PLOP und PLOP DS

Neben zahlreichen Funktionen wie Ver- und Entschlüsselung, Optimierung und digitaler Signatur bieten PDFlib PLOP und PLOP DS die Möglichkeit, XMP-Metadaten in vorhandene PDF-Dokumente einzufügen. Dies ist nützlich bei bereits vorhandenen PDF-Dokumenten, die noch nicht alle erforderlichen Metadaten-Properties enthalten. Profitieren können hier insbesondere PDF/A-Workflows, da

die XMP-Unterstützung von PLOP und PLOP DS die Konformität zu PDF/A erhält. So kann benutzerdefiniertes XMP mit Extension-Schemas in existierende PDF/A-Dokumente eingebracht werden, die aus Workflows stammen, die keine Extension-Schemas unterstützen.

Extrahieren von XMP aus PDF mit PDFlib pCOS

PDFlib GmbH bietet mit der produktübergreifenden pCOS-Schnittstelle die Möglichkeit, verschiedenste Informationen aus PDF-Dokumenten abzufragen. pCOS ist als eigenständiges Produkt verfügbar und darüber hinaus in alle anderen Produkte integriert. pCOS bietet eine einfache Programmiermethode, um XMP-Metadaten aus PDF-Dokumenten zu extrahieren. XMP-Metadaten werden in Unicode normalisiert, so dass Benutzer sich nicht um Encoding-Fragen zu kümmern brauchen.

Die XMP-Extraktion erfolgt unabhängig von Kompression, Verschlüsselung und PDF-Objektstruktur. Adobe definiert einen XMP-Paketmechanismus, mit dem sich XMP-Datenpakete auf einfache Weise in verschiedenen Dateiformaten einfügen und abfragen lassen. Das PDF-Dateiformat weist jedoch einige Besonderheiten auf, die die Sache verkomplizieren. So können PDF-Dokumente über mehrere Update-Abschnitte verfügen. Dies hat zur Folge, dass mehrere Instanzen eines XMP-Streams in der Datei vorhanden sind, auch wenn nur eine einzige davon relevant ist. Eine einfache textuelle Suche nach dem XMP-Block liefert dann höchstwahrscheinlich die falsche Instanz. Ein Programm muss die PDF-Objektstruktur durchlaufen, um die XMP-Metadaten zuverlässig zu finden.

Workflow-Szenarios, die von XMP-basierter Dokumentsuche profitieren

Die Verarbeitung von XMP-Metadaten beim Durchsuchen digitaler Dokumente lässt sich auf verschiedenen Ebenen konfigurieren. Im Folgenden finden Sie zwei typische Beispiele.

Publishing: Creative Professionals nutzen Publishing-Software von Adobe und anderen Herstellern zur interaktiven Erstellung von Dokumenten und Metadaten. Sie versehen die Dokumente mit Schlüsselwörtern, Verfassernamen, Copyright-Informationen und anderen XMP-Properties. Mit Adobe Bridge können sie die Dokumente gemäß den ihnen zugeordneten Metadaten-Properties durchsuchen oder gruppieren, wobei sie vorwiegend gängige XMP-Schemas wie Dublin Core und IPTC einsetzen.

Technische Dokumentation: Eine große Zahl von Dokumenten wird manuell oder automatisch generiert und nach Abteilung oder Firma gruppiert abgelegt. Auf diese Dokumentsammlungen kann mit gängigen Windows-Retrieval-Produkten zugegriffen werden, zum Beispiel mit Microsoft SharePoint auf Servern, Windows Search auf Desktops oder mit anderen Retrieval-Produkten. Sobald TET PDF IFilter an diese Produkte angeschlossen ist, können Benutzer Dokumente nicht nur nach dem eigentlichen Seiteninhalt, sondern auch nach den XMP-Metadaten oder Bildeigenschaften durchsuchen. Während vordefinierte XMP-Schemas gängige Anforderungen abdecken, lassen sich mit benutzerdefinierten XMP-Schemas spezielle firmenspezifische Bedürfnisse befriedigen.

PDFlib GmbH
Franziska-Bilek-Weg 9
D-80339 München
Tel. +49 • 89 • 452 33 84-0
support@pdflib.com
www.pdflib.com/knowledge-base/xmp-metadata

PDFlib GmbH ist auf die Entwicklung von PDF-Technologie spezialisiert. PDFlib-Produkte sind seit 1997 weltweit im Einsatz. Das Unternehmen berücksichtigt wichtige technologische Trends, etwa ISO-Standards für PDF. PDFlib GmbH vertreibt alle Produkte weltweit, wobei Nordamerika, Europa und Japan die wichtigsten Märkte darstellen.