

Whitepaper: XMP-Metadaten in PDFlib-Produkten

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 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 oder Übersetzung, 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 von Adobe nach dem vom W3C entwickelten RDF (Resource Description Framework) entworfen, das die Grundlage der Semantic Web Initiative bildet. Adobe stellt die XMP-Spezifikation nicht nur kostenlos zur Verfügung, sondern bietet zudem ein Open-Source XMP-Toolkit für Software-Entwickler an.

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 Eigenschaften (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.

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 Dokumentinfoeinträgen, die vor der Einführung von XMP dazu dienten, Metadaten zu transportieren. Dokumentinfoeinträge 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.

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...*), Photoshop, InDesign und anderen Adobe-Anwendungen erfolgt die Anzeige und Bearbeitung von XMP-Metadaten im Panel *Dokumenteigenschaften* bzw. *Dateiinformatioenen*. Dieses Panel zeigt Metadaten-Properties entsprechend der vordefinierten XMP-Schemas in übersichtlicher Anordnung an. Außerdem können so genannte Custom Panels definiert werden, die die Anzeige der Metadaten sowie editierbare Felder für die Bearbeitung an verschiedene Anwendungsgebiete anpassen.

XMP für verschiedene Industriezweige. Zur Implementierung der notwendigen Metadaten kommt in der Industrie zunehmend XMP zum Einsatz. Einige Beispiele:

- ▶ Das AdsML-Konsortium erstellt Spezifikationen und Prozessabläufe für den Austausch von Werbedaten und -inhalten.
- ▶ Das International Press Telecommunications Council (IPTC) ist ein Konsortium der Zeitungswirtschaft. Es entwickelt Industriestandards für den Austausch von Zeitungsdaten und publiziert das weit verbreitete Schema »IPTC Core« für XMP, das zur Übertragung von Metadaten für Bilder und andere Zeitungselemente 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 Zeitungen, Zeitschriften, Katalogen und Büchern.

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

- ▶ 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. Eine formale Beschreibung der Extension-Schemas muss im PDF/A-Dokument eingebettet sein, damit benutzerdefinierte Metadaten zu einem späteren Zeitpunkt auch garantiert korrekt interpretiert werden können. PDF/A-1 erlaubt die Verwendung von Dokumentinfoeinträgen, einige gängige PDF-Dokumentinfoeinträge müssen aber mit entsprechenden vordefinierten XMP-Properties synchronisiert werden, damit rein auf XMP basierende Arbeitsabläufe gewährleistet sind. Der Standard definiert den Abgleich (Crosswalk) zwischen Dokumentinfoeinträgen und XMP-Properties. Die XMP-Unterstützung in PDF/A-1 basiert auf der Spezifikation XMP 2004.
- ▶ PDF/E in ISO 24517-1 (veröffentlicht 2008): »Engineering document format using PDF – Use of PDF 1.6«. Die 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): »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 muss die Konformität zu PDF/X-4 mit XMP ausgedrückt werden. Dokumentinfoeinträge sind in PDF/X-4 erlaubt, müssen aber mit den entsprechenden XMP-Einträgen syn-

chronisiert werden. XMP-Extension-Schemas für benutzerdefinierte Metadaten sind zulässig. Im Gegensatz zu PDF/A-1 können diese aber ohne Einbettung einer formalen Beschreibung verwendet werden. Die XMP-Unterstützung in PDF/X-4 basiert auf der Spezifikation XMP 2005.

- ▶ PDF/X-2 in ISO 15930-5 (veröffentlicht 2003) und PDF/X-5 in ISO 15930-8 (veröffentlicht 2008): »Partial exchange of printing data using PDF 1.6 (PDF/X-5)«. PDF/X-2- und X-5-Dokumente verweisen auf andere PDF/X-Dokumente, wobei das Verweisziel durch verschiedene XMP-Einträge identifiziert wird. XMP ist damit ein wesentlicher Bestandteil von PDF/X-2 und X-5.
- ▶ ISO 32000 (veröffentlicht 2008): »Document management – Portable document format – PDF 1.7«. ISO 32000 ist die standardisierte Fassung von PDF 1.7. Der technikbezogene Inhalt entspricht PDF 1.7 (dem Dateiformat von Acrobat 8), das vollständige Unterstützung von XMP-Metadaten bietet.

Dublin Core, eines der am häufigsten verwendeten Schemas für vordefinierte XMP-Metadaten, wurde als ISO 15836 standardisiert (veröffentlicht 2003): »Information and documentation — The Dublin Core metadata element set«.

XMP-Unterstützung in der PDFlib-Produktfamilie. Die PDFlib-Produktfamilie unterstützt XMP in einfacher Form seit 2004. Mit der Unterstützung von PDF/A-1 in PDFlib 7 (freigegeben 2006) wurde die XMP-Funktionalität erweitert, um den Anforderungen von PDF/A-1 zu genügen. Implementiert wurde insbesondere der automatische Abgleich von Dokumentinfoeinträgen mit entsprechenden XMP-Properties (festgelegt im PDF/A-1-Crosswalk) sowie die automatische Erstellung verschiedener interner XMP-Properties, die für PDF/A-1 erforderlich sind. PDFlib-Benutzer können seitdem XMP für PDF/A-1 erstellen, ohne sich um die Einzelheiten des XMP-Formats kümmern zu müssen. Fortgeschrittene Benutzer können alle vordefinierten XMP-Metadaten-Schemas 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.

PDFlib 7.0.3 bietet erstmals Unterstützung für XMP-Extension-Schemas gemäß PDF/A-1. Benutzer können die von PDF/A-1 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 konform zum Standard PDF/A-1. PDFlib 7.0.3 ist damit das weltweit erste Produkt, das XMP-Extension-Schemas für PDF/A-1 unterstützt. Durch die Mitwirkung von PDFlib GmbH im PDF/A Competence Center erfolgen alle PDF/A-Aktivitäten in enger Absprache mit anderen Anbietern von PDF/A-Software, so dass ein höchstmögliches Maß an Konformität zum Standard gewährleistet ist und die gängige Praxis berücksichtigt wird.

Da die XMP-Validierung auch dann zum Einsatz kommt, wenn keine PDF/A-Ausgabe generiert wird, profitieren alle Benutzer von den Verbesserungen der XMP-Unterstützung in PDFlib 7.0.3.

Weitere Informationen zu XMP in PDF/A sowie einen Online-Validierer für XMP-Extension-Schemas finden Sie unter www.pdflib.com.

Einfügen von XMP in PDFs mit PDFlib PLOP 3.1. Neben zahlreichen Funktionen wie Ver- und Entschlüsselung, Optimierung und digitaler Signatur bietet PDFlib PLOP 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 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 PDFs mit PDFlib pCOS. PDFlib GmbH nutzt generell die pCOS-Schnittstelle, um 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 zurückgegeben, so dass Benutzer sich nicht um Encoding-Fragen zu kümmern brauchen.

Die XMP-Extraktion erfolgt unabhängig von Komprimierung, 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 sehr sorgfältig durchlaufen, um die XMP-Metadaten zuverlässig zu finden. Dies ist auch der Grund, warum das kostenlose XMP-Toolkit von Adobe die XMP-Abfrage aus PDF-Dokumenten nur begrenzt unterstützt, während XMP aus anderen Dateiformaten wie TIFF und JPEG ohne Einschränkungen extrahiert werden kann.

Suche nach XMP-Metadaten mit PDFlib TET PDF IFilter. TET PDF IFilter ist das jüngste Produkt der PDFlib GmbH. Es 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 Desktop Search (WDS), Office SharePoint Server (MOSS), Indexing Server 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 leistungsstarke Metadaten-Implementierung von TET PDF IFilter unterstützt das Property-System von Windows für Metadaten. Neben den Seiteninhalten indiziert TET PDF IFilter XMP-Metadaten sowie Standard- und benutzerdefinierte Dokumentinfoeinträge. Die Indizierung der Metadaten lässt sich auf verschiedenen Ebenen konfigurieren:

- ▶ Dokumentinfoeinträge und gängige XMP-Properties werden auf Standard-Windows-Properties wie *Title*, *Subject* oder *Author* abgebildet.
- ▶ TET PDF IFilter ergänzt nützliche PDF-spezifische Pseudo-Properties, z.B. Seitengröße, PDF/A-Konformitätsstufe, 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 indiziert, z.B. der Name des Fotografen eines Bildes oder Copyright-Angaben.

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

Workflow-Szenarios, die von XMP-basierter Dokumentsuche profitieren. Die Verarbeitung von XMP-Metadaten lässt sich in verschiedene Szenarien integrieren, in denen digitale Dokumente durchsucht werden müssen. Zwei typische Beispiele werden im folgenden beschrieben.

Publishing: Creative Professionals nutzen Publishing-Software von Adobe und anderen Herstellern zur interaktiven Erstellung von Dokumenten und Metadaten. Sie versehen die Dokumente mit Schlüsselwörtern, Verfasseramen und anderen



üblichen 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 sehr 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 Office SharePoint Server (MOSS) auf Serversystemen oder Windows Desktop Search (WDS) auf Workstations. 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 Eigenschaften von Bildern 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
info@pdflib.com
www.pdflib.com*

