

Whitepaper: Erstellen von PDF/A mit PDFlib-Produkten

Die im Standard ISO 19005 definierten PDF/A-Formate definieren eine konsistente und stabile Teilmenge von PDF, die der sicheren Langzeitarchivierung und dem zuverlässigen Datenaustausch in Unternehmen und Behörden dient. Das vorliegende Whitepaper erläutert PDFlib-Funktionen für die Erstellung von PDF/A-Dokumenten zur Langzeitarchivierung.

PDF/A-1a und PDF/A-1b. PDF/A-1, das im internationalen Standard ISO 19005-1 definiert ist, dient der zuverlässigen Langzeitarchivierung von digitalen Dokumenten. Der Standard basiert auf PDF 1.4 und definiert gewisse Einschränkungen bezüglich der Verwendung von Farben, Schriften, Anmerkungen und anderen Elementen. PDF/A-1 umfasst zwei Varianten. Beide können mit PDFlib-Produkten erstellt und verarbeitet werden:

- ▶ ISO 19005-1 Level B (PDF/A-1b) stellt sicher, dass das visuelle Erscheinungsbild eines Dokuments über lange Zeit erhalten bleibt. Das bedeutet, dass das Aussehen eines Dokuments unverändert ist, wenn es zu einem späteren Zeitpunkt angezeigt oder gedruckt wird.
- ▶ ISO 19005-1 Level A (PDF/A-1a) erweitert Level B um wesentliche Eigenschaften von »Tagged PDF«: Gefordert werden Strukturinformationen und Angaben zur inhaltlichen Gliederung, um die logische Dokumentstruktur sowie die natürliche Lesereihenfolge zu erhalten. PDF/A-1a sichert also nicht nur das Aussehen eines Dokuments, sondern sorgt auch dafür, dass der Dokumentinhalt (Semantik) zuverlässig interpretierbar bleibt und körperlich beeinträchtigten Benutzern zugänglich ist (Barrierefreiheit).

Wenn im Folgenden von PDF/A ohne Kompatibilitätsstufe die Rede ist, sind beide Kompatibilitätsstufen gemeint. Die PDF/A-Implementierung in allen PDFlib-Produkten basiert auf dem Standard ISO 19005-1:2005 einschließlich Technical Corrigendum 1 (2007).

Anforderungen und Einschränkungen in PDF/A. Bei PDF/A sind einige PDF-Funktionen zwingend erforderlich, andere dagegen sind untersagt. Um zum Beispiel Text exakt reproduzieren zu können, müssen alle in einem Dokument verwendeten Fonts eingebettet sein; für die exakte Farbproduktion dagegen sind alle Farben geräteunabhängig anzugeben. Metadaten müssen im XMP-Format eingebettet sein und Verschlüsselung ist nicht erlaubt.

Zusätzlich zu diesen leicht verständlichen Eigenschaften fordert PDF/A weitere PDF-Funktionen, die weniger offensichtlich sind (z.B. bestimmte Einträge in den Fontdatenstrukturen), und untersagt einige kritische Strukturen (z.B. bestimmte Kombinationen von TrueType-Schriften und Encodings). Software-Entwickler müssen sich mit zahlreichen Aspekten auseinandersetzen, die gründlich implementiert und getestet werden müssen, um vollständige Konformität zu PDF/A zu erreichen. Somit ist PDF/A weit mehr als nur »PDF mit eingebetteten Fonts«!

PDF/A-Unterstützung in der PDFlib-Produktfamilie. PDFlib bietet Anwendungsentwicklern ein Tool, mit dem sie folgende Operationen im Hinblick auf PDF/A durchführen können:

- ▶ PDF/A erstellen, z.B. aus Text, der aus einer Datenbank stammt
- ▶ Rasterbilder (z.B. Scans) nach PDF/A konvertieren
- ▶ Vorhandene PDF/A-Dokumente weiterverarbeiten, z.B. um PDF-Dokumente zu kombinieren oder aufzuteilen
- ▶ PDF/A-1a mit Strukturinformationen erstellen (Tagged PDF)

- ▶ XMP-Metadaten zu den erzeugten Dokumenten hinzufügen, einschließlich des diffizilen *Themas Extension-Schemas* (siehe unten).

All diese Operationen lassen sich mit einfachen PDFlib-Funktionsaufrufen implementieren. Beispielcode für verschiedene Programmiersprachen und Entwicklungsumgebungen wird mit der PDFlib-Distribution mitgeliefert. Weitere Programmierbeispiele für PDF/A finden Sie im *PDFlib Cookbook*. Da PDF/A zahlreiche Überlappungen zum Standard PDF/X (ISO 15930) für die grafische Industrie aufweist, profitierte die PDF/A-Entwicklung in PDFlib davon, dass wir bereits seit einigen Jahren mehrere PDF/X-Varianten unterstützen. Um die von PDF/A geforderte Fonteinbettung zu erleichtern, bietet das *Japanese Resource Kit for the PDFlib Family* kostenfrei einbettbare Fonts sowie sprachspezifische ICC-Profile, CMaps und Dokumentation für japanische Benutzer.

Erstellung PDF/A-konformer Ausgabe. Die Erstellung PDF/A-konformer Ausgabe funktioniert in PDFlib wie folgt:

- ▶ PDFlib setzt automatisch verschiedene formale für PDF/A erforderliche Einstellungen, z.B. die PDF-Versionsnummer oder die nötigen XMP-Einträge zur PDF/A-Identifizierung.
- ▶ Das PDFlib-Clientprogramm muss explizit bestimmte Funktionsaufrufe und -optionen verwenden (z.B. für die Fonteinbettung).
- ▶ Das PDFlib-Clientprogramm darf verschiedene andere Funktionsaufrufe und -optionen nicht verwenden (z.B. Verschlüsselung).

Genügt das PDFlib-Clientprogramm diesen Regeln, ist die Ausgabe garantiert PDF/A-konform. Entdeckt PDFlib eine Verletzung dieser Regeln, so wird eine Exception ausgelöst, die von der Anwendung abgefangen werden muss. Im Fehlerfall wird kein PDF erzeugt, so dass keine Gefahr besteht, dass nichtkonforme Dokumente erstellt werden. Die notwendigen bzw. verbotenen Operationen werden im Detail in der PDFlib-Dokumentation beschrieben.

Geräteunabhängige Farbangaben. Um eine konsistente Farbreproduktion zu gewährleisten, erfordert PDF/A geräteunabhängige Farbangaben, die in der Regel mittels ICC-Profilen oder durch den Farbraum CIE Lab erreicht werden. Eine optionale Druckausgabebedingung (Output Intent) beschreibt die Farbeigenschaften des Dokuments. Während die Anwendung dieser Konzepte in der grafischen Industrie weit verbreitet ist, sind PDF-Entwickler im Unternehmensbereich dagegen nicht unbedingt mit Farbmanagement vertraut. PDFlib erleichtert die Erstellung geräteunabhängiger Ausgabe, unabhängig von der Art der Eingabedaten:

- ▶ PDF/A-Ausgabe kann mit oder ohne ICC-Profil für die Druckausgabebedingung erstellt werden.
- ▶ Im häufigen Fall mit schwarzem Text wählt PDFlib automatisch den passenden Farbraum (Lab oder DeviceGray), abhängig davon, ob ein ICC-Profil für die Druckausgabebedingung spezifiziert wurde oder nicht.
- ▶ Externe ICC-Profile und in Bilder eingebettete Profile ermöglichen detaillierte Farbsteuerung.
- ▶ Die PDFlib-Distribution enthält ICC-Profile für häufige Anwendungsfälle. Damit lässt sich mit geringem Aufwand korrekte PDF/A-Ausgabe erstellen.

Rasterbilder, z.B. TIFF oder JPEG, spielen eine wichtige Rolle bei der Dokumenterstellung. Typisch im Dokumentenworkflow sind etwa eingescannte Dokumente oder Fotos aus Digitalkameras. Während Rasterbilder in modernen Workflows (meist mittels eingebetteter ICC-Farbprofile) bereits geräteunabhängig und damit PDF/A-konform sind, ist dies bei älteren Bilddaten oft nicht der Fall, etwa bei Bildern, die im Schwarzweiß- oder RGB-Modus ohne zugehöriges ICC-Profil eingescannt wurden. PDFlib unterstützt beide Anwendungssituationen:

- ▶ ICC-Profile, die in Rasterbildern eingebettet sind, werden berücksichtigt.
- ▶ Externe ICC-Profile können auf ein Bild angewandt werden.

- ▶ Als Behelfslösung für ältere Daten unbekannter Herkunft ist ein sRGB-Profil in PDFlib integriert, das kompatibel zu vielen Hard- und Software-Produkten ist.
- ▶ Durch Angabe eines ICC-Profiles für die Druckausgabebedingung können geräteabhängige Bilddaten verwendet werden, ohne dass ein ICC-Profil auf einzelne Bilder angewandt werden muss.

In der PDFlib-Dokumentation werden PDF/A-Farbstrategien für häufige Anwendungsfälle durchgespielt.

XMP leicht gemacht. PDF/A verlangt, dass Metadaten, also Informationen über ein Dokument, im XMP-Format im PDF gespeichert werden. XMP definiert einen leistungsfähigen und flexiblen Rahmen für Standard- oder selbstdefinierte Metadaten (weitere Informationen hierzu finden Sie in unserem Whitepaper zu XMP). Falls Sie bereits XMP-Metadaten in Ihrem Workflow verwenden, können Sie vollständige XMP-Streams erzeugen, die PDFlib dann in die PDF/A-Ausgabe integriert. Entwickler, die mit XMP nicht vertraut sind, brauchen sich jedoch nicht in dieses Thema einzuarbeiten. PDFlib erstellt die von PDF/A geforderte XMP-Ausgabe und setzt die klassischen Dokumentinfo-Einträge automatisch in die entsprechenden vom Standard verlangten XMP-Konstrukte um. Damit können Entwickler die Möglichkeiten von XMP gezielt nutzen oder die automatische XMP-Generierung von PDFlib in Anspruch nehmen, falls nur einfachere Anforderungen an die Metadaten zu erfüllen sind.

XMP-Extension-Schemas. XMP ist grundsätzlich erweiterbar, so dass firmen- oder branchenspezifische Anforderungen durch XMP-Extension-Schemas erfüllt werden können. PDF/A unterstützt dieses Konzept, verlangt aber, dass eine maschinenlesbare Beschreibung des Schemas nach vorgegebenen Regeln in das Dokument eingebettet wird, um die zukünftige Auswertung zu erleichtern. XMP-Extension-Schemas für PDF/A werden von PDFlib-Produkten unterstützt (PDFlib 7.0.3 war das weltweit erste Produkt mit Unterstützung für Extension-Schemas). PDFlib validiert externe XMP-Metadaten inklusive Extension-Schemas, um sicherzustellen, dass die generierte Ausgabe vollständig konform zu PDF/A ist.

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

Verarbeitung vorhandener PDF/A-Dokumente. Weitere Regeln sind zu beachten, wenn Seiten aus vorhandenen PDF/A-konformen Dokumenten importiert werden. In Adobe Acrobat zum Beispiel können Sie zwei PDF/A-Dokumente sehr leicht so kombinieren, dass das Ergebnis nicht mehr PDF/A-konform ist (wobei keine Warnung erfolgt). Bei der Verarbeitung vorhandener PDF/A-Dokumente untersucht PDFlib+PDI sorgfältig die PDF/A-Eigenschaften aller Eingabe- und Ausgabe-Dokumente, um sicherzustellen, dass auch die Ausgabe PDF/A-konform wird. Außerdem kann die Druckausgabebedingung eines importierten Dokuments in das Ausgabe-PDF kopiert werden, um die PDF/A-Farbeinstellungen eines vorhandenen Dokuments exakt zu kopieren.

Erstellung von PDF/A-1a mit Tagged PDF. PDF/A-1a ist im Prinzip PDF/A-1b plus Tagged PDF: Strukturinformationen für das Dokument sind erforderlich; Fonts unterliegen bestimmten Bedingungen, damit sichergestellt ist, dass der Text fehlerfrei interpretiert werden kann. Dokumente im Format PDF/A-1a sind damit für körperlich beeinträchtigte Benutzer uneingeschränkt nutzbar (barrierefrei). Neben dem optischen Erscheinungsbild garantieren sie auch die Bedeutung des Dokumentinhalts.

Die Unterstützung für PDF/A-1a in PDFlib basiert auf den Funktionen zur Erzeugung von Tagged PDF: Jeder Dokumentbestandteil kann an einer bestimmten Stelle des Dokumentstrukturbaums platziert werden; Dokumentbestandteile, die für die Dokumentstruktur nicht relevant sind (z.B. Kopf- und Fußzeilen oder Seitennummern), können als Artefakte gekennzeichnet werden, um anzuzeigen, dass

man sie bei der Weiterverwendung des Dokuments ignorieren kann (z.B. wenn das Dokument von Software vorgelesen oder in ein anderes Format konvertiert wird). Bilder können mit Alternativtext versehen werden, der zum Beispiel sehbehinderten Benutzern von Acrobat vorgelesen wird.

Beachten Sie, dass Sie detaillierte Kenntnisse über die logische Dokumentstruktur benötigen, um Tagged PDF zu generieren. PDFlib kümmert sich um die Details der PDF-Ausgabe, kann aber nicht die Dokumentstruktur aus dem Dokumentinhalt erschließen.

Tagged PDF wird seit PDFlib 6 unterstützt. In PDFlib 7 können auch Anmerkungen in den Dokumentstrukturbaum integriert werden. Dies verbessert die barrierefreie Zugänglichkeit von Links und anderen interaktiven Elementen.

Auf Basis der bereits vorhandenen Unterstützung für Tagged PDF kann PDFlib 7 Ausgabe gemäß PDF/A-1a erstellen. Damit ist PDFlib das erste Werkzeug, das diesen fortgeschrittenen PDF/A-Level unterstützt.

Validierung von PDF/A. Bei der Einrichtung von Workflows, die auf einem Standard basieren, sollten Tools zum Einsatz kommen, die überprüfen, ob die Ergebnisse auch wirklich diesem Standard entsprechen. Für PDF/A gibt es Validierer, die überprüfen, ob ein PDF-Dokument dem ISO-Standard entspricht.

Die PDF/A-Ausgabe von PDFlib entspricht vollständig den Validierungsregeln des Preflight-Werkzeugs von Acrobat 9, das einen der striktesten PDF/A-Validierer auf dem Markt darstellt. Beachten Sie, dass die PDF/A-Validierung von Acrobat 8 den ISO-Standard noch nicht vollständig umsetzt.

Außerdem arbeitet PDFlib GmbH aktiv mit den Anbietern von Software für die PDF/A-Validierung zusammen, um sicherzustellen, dass Ersteller und Validierer den PDF/A-Standard auf die gleiche Art interpretieren.

Derzeit bietet PDFlib GmbH keine Produkte zur Validierung von PDF/A-Dokumenten (d.h. zur Überprüfung der Konformität zum Standard) oder zur Konvertierung von beliebigen PDF-Dokumenten nach PDF/A.

Verarbeitung von PDF/A-konformen Dokumenten mit PDFlib PLOP. Die Produkte PDFlib PLOP und PLOP DS bieten verschiedene PDF-Verarbeitungsfunktionen, die PDF/A berücksichtigen. So werden PDF/A-konforme Eingabedateien auch in PDF/A-konforme Ausgabe umgesetzt. Kann die Verarbeitung nur unter Verletzung des Standards erfolgen, wird eine Fehlermeldung ausgegeben. Einige Beispiele:

- ▶ Die Verschlüsselung von PDF/A-Dokumenten führt zu einer Fehlermeldung, da Verschlüsselung in PDF/A nicht erlaubt ist. Um ein PDF/A-Dokument dennoch mit PLOP zu verschlüsseln, müssen Sie den PDF/A-Status explizit angeben.
- ▶ Wenn Sie mit PLOP DS ein PDF/A-Dokument mit einer digitalen Signatur versehen, wird das Signaturfeld konform zum PDF/A-Standard generiert.
- ▶ Mit PLOP können Sie XMP-Metadaten in vorhandene PDF/A-Dokumente aufnehmen. Da PLOP 3.1 XMP-Extension-Schemas für PDF/A unterstützt, können diese in einem zweiten Bearbeitungsschritt in existierende Dokumente eingebracht werden. Damit lassen sich XMP-Extension-Schemas auch dann nutzen, wenn sie von der ursprünglichen Erstellungssoftware der PDF/A-Dokumente nicht unterstützt werden.

PDF/A Competence Center. PDFlib GmbH ist Gründungsmitglied des PDF/A Competence Center, das die Nutzung von PDF/A fördern und die Kompatibilität zwischen Anbietern gewährleisten will. PDFlib GmbH ist aktives Mitglied der Technischen Arbeitsgruppe (TWG), die TechNotes zu PDF/A-Themen publiziert. Die TWG veröffentlichte zudem die Isartor-Testsuite für PDF/A-1. Dabei handelt es sich um eine Sammlung von Testdateien, mit denen man die Standardkonformität und Abdeckung eines PDF/A-Validierers überprüfen kann. Weitere Informationen hierzu finden Sie unter www.pdfa.org.

PDFlib GmbH
Franziska-Bilek-Weg 9
D-80339 München
Tel. +49 • 89 • 452 33 84-0
info@pdflib.com

