Validierung von PDF/A

Dieser Artikel befasst sich mit dem Thema Validierungssoftware und stellt die Pläne zur Entwicklung der Testsuite des PDF/A Competence Centers vor.

Der Autor und PDF-Experte Thomas Merz ist Geschäftsführer der PDFlib GmbH. Das Unternehmen mit Sitz in München ist seit 1997 auf PDF-Entwicklung fokussiert. Die PDFlib-Produktfamilie bietet Komponenten zur dynamischen Erstellung von PDF, auch zur Herstellung von PDF/A-1a (Tagged PDF) und PDF/A-1b. PDFlib GmbH ist Gründungsmitglied des PDF/A Competence Centers.

Zudem ist Thomas Merz in der Technical Working Group (TWG) tätig, die sich mit der Erarbeitung gemeinsamer technischer Positionen befasst. Ergebnis sind hier unter anderem die „Technical Notes" (diese werden auf pdfa.org veröffentlicht), etwa zu den Bereichen „PDF/A-1 und Namespaces", „Farbe in PDF/A-1", „Metadaten (XMP/Docinfo) in PDF/A-1" sowie „Digitale Signatur und PDF/A-1". Die TWG kooperiert mit dem ISO-Komitee, etwa in der Erarbeitung des Standards PDF/A-2 und in der Erstellung der PDF/A-Testsuite.

Grundlagen

Als ISO 19005-1 (Oktober 2005) wurde der Standard PDF/A-1 im Herbst 2005 veröffentlicht. Als Ergänzung und Klarstellung wurde im zweiten Quartal 2007 ein Korrigendum publiziert. Der ISO-Standard selbst umfasst nur 36 Seiten, bezieht sich aber auf weitere, viel umfangreichere Spezifikationen (etwa die Referenz zu PDF 1.4, Spezifikationen zu XMP, Fontformate, ICC und weitere mehr).

Die Konformität umfasst nicht nur die Angaben im ISO-Standard selbst, sondern auch in den Sekundärspezifikationen. Es gibt keine Referenzimplementierung (Dokumente/Software).

Der Standard PDF/A-1a zeichnet sich durch weiter gehende Anforderungen aus als PDF/A-1b. PDF/A-1a verlangt unter anderem Tagged PDF und Unicode.

Die PDF/A-Konformität bedeutet einen beträchtlichen Aufwand für die Hersteller von PDF-Software.

Aspekte der PDF/A-Konformität

Die PDF/A Konformität umfasst mehrere Bereiche, die ein Dokument im Laufe seiner Existenz durchläuft beziehungsweise durchlaufen kann:

- **Erstellung:** Dieser Schritt bezieht sich auf die Generierung PDF/A-konformer Dokumente aus diversen Ursprungsformaten.
- **Korrektur:** Eine Modifikation von PDF ist unter Umständen zur Erreichung der PDF/A-Konformität notwendig.
- **Verarbeitung:** Eine Modifikation von PDF/A muss unter Erhaltung der Konformität durchgeführt werden.
- **Anzeige:** Hier ist die Darstellung gemäß den Vorgaben von PDF/A gemeint. Eine PDF/A-Datei lediglich „irgendwie" anzuzeigen, wie es viele Viewer erlauben, reicht in diesem Zusammenhang nicht aus.
- **Validierung:** Die Überprüfung der PDF/A-Konformität von Dokumenten gemäß Standard.

PDF/A-Validierung: Produkte

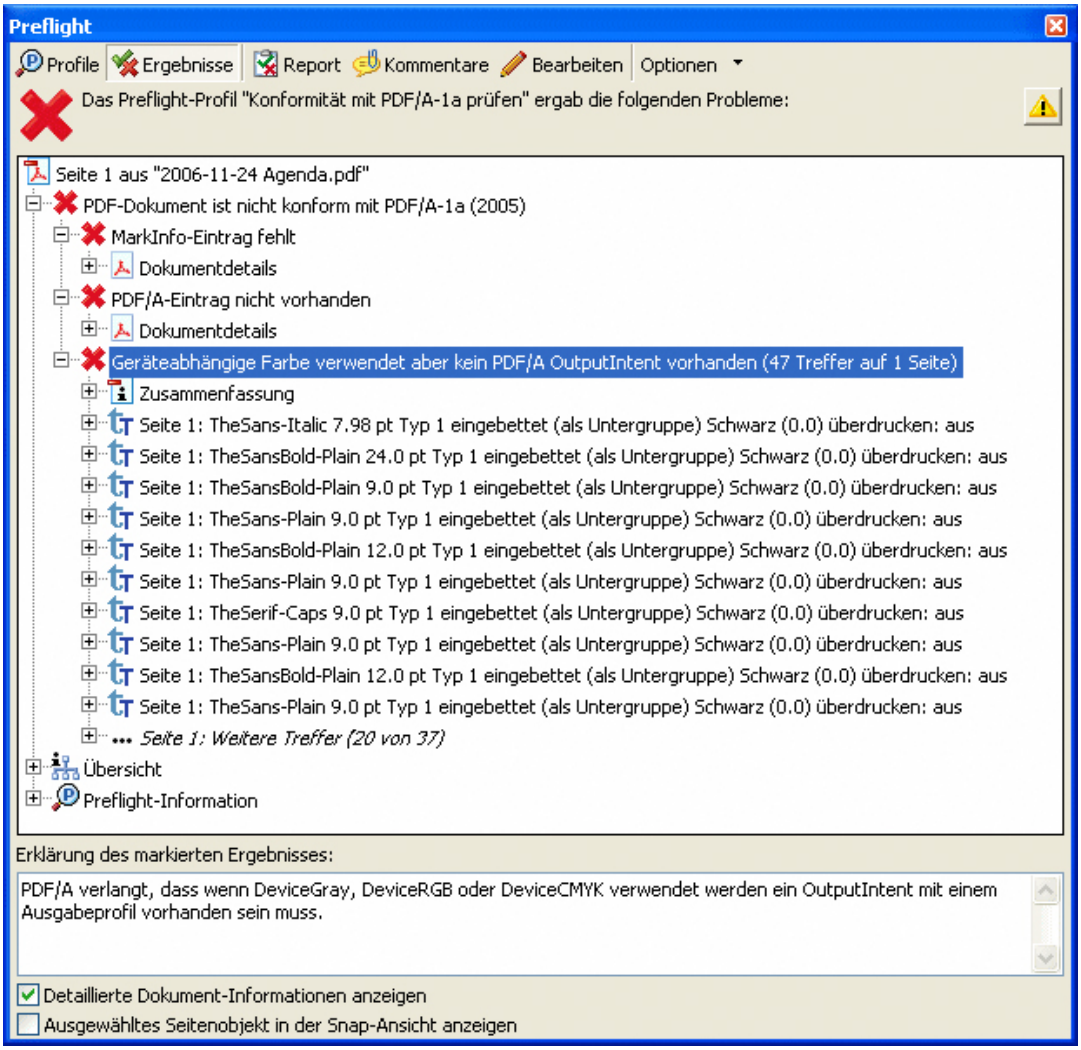
Für die Validierung von PDF/A-Dokumenten sind eine Reihe von Anwendungen auf dem Markt. Zu den Lösungen zählen auch die folgenden Validierer von PDF/A Competence Center Mitgliedsfirmen:

- Acrobat 8 Preflight (entwickelt von callas software)
- PDF Tools AG: 3-Heights PDF Validator
- LuraTech: LuraDocument PDF Validator
- Seal Systems: PDF Checker
- Intarsys: PDF/A Live!
- callas: pdfaPilot
- callas: pdfInspektor
- Apago: PDF Appraiser (Vertrieb durch Actino)

Es gibt zudem weitere Produkte anderer Anbieter.

Beispiel: Acrobat 8 Preflight

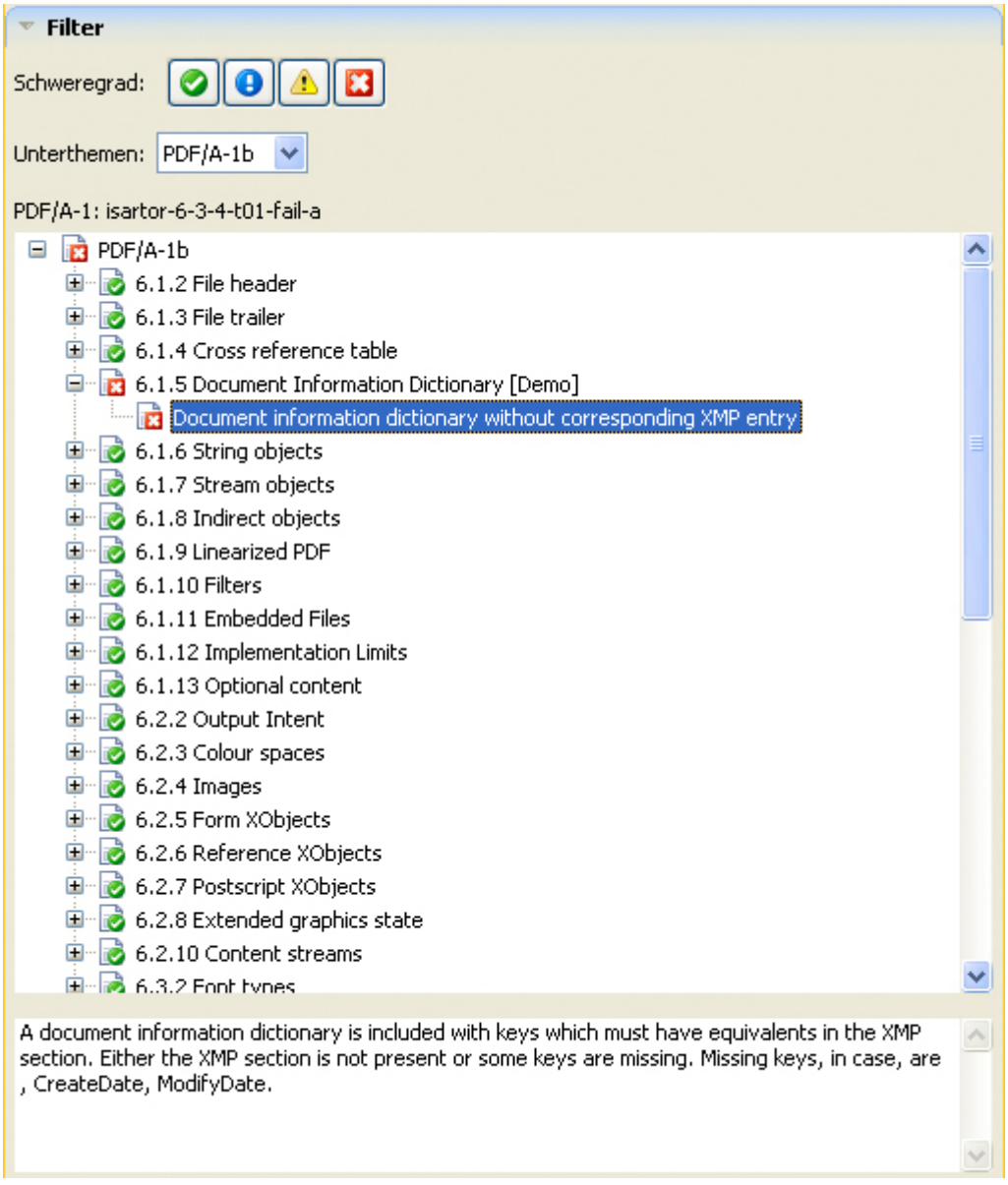
Als erstes Beispiel für die PDF/A-Validierung sei hier die Acrobat-Funktion Preflight genannt, die eine Entwicklung von Adobe/callas software ist. PDF/A-Fähigkeiten sind in Preflight von Acrobat 7 nur als Draft (Entwurf) enthalten, der verabschiedete Standard ist in Acrobat 8 implementiert.



Preflight ist in der Lage, PDF-Dateien auf die Einhaltung von PDF/A zu validieren. Die Abbildung zeigt das Ergebnis einer Prüfung, bei der das PDF nicht PDF/A-konform ist.

Beispiel: Cabaret Stage 3

Das Programm Cabaret Stage 3 (Vertrieb über Intarsys) ist eine Software, mit der sich PDF-Dokumente anzeigen, ausfüllen, bearbeiten, speichern, drucken und validieren lassen.



Validierung in Cabaret Stage 3: Die geprüfte Datei ist nicht PDF/A-konform.

PDF/A-Validierung: Aspekte

Für die Überprüfung der PDF/A-Konformität von Dokumenten sind einige Aspekte zu beachten. Zu den Hauptaspekten der Validierung zählen sowohl die Breite als auch die Tiefe der Prüfung.

Breite der Prüfung:

Alle Vorschriften des Standards müssen abgedeckt sein. Das bedeutet, dass sämtliche Verbote und Gebote (=Regeln) im Standard überprüft werden müssen. Dabei betreffen einige Regeln alle Dokumente, wie z.B. XMP oder Farbräume. Manche Regeln betreffen nur spezielle Datenstrukturen, die nicht unbedingt in jedem PDF-Dokument vorhanden sein müssen. Hierzu zählen etwa Font-Untergruppen (Subsets) und Annotationen.

Tiefe der Prüfung:

Bei der Tiefe der Prüfung stellt sich die Frage, wie genau die Datenstrukturen untersucht und wie detailliert einzelne Regeln überprüft werden. Folgende Sekundärspezifikationen müssen berücksichtigt werden:

- Fonts: TrueType/OpenType/PostScript Type 1
- ICC-Farbprofile
- XMP -> RDF -> XML, Namespaces

Auch die Kombination von Eigenschaften kann von Bedeutung sein:

- Beispielsweise Fonts und Formularfelder: sind auch die Fonts eingebettet, die in Formularfeldern benutzt werden?

Der Bereich PDF/A-1a und Tagged PDF:

- Hier gibt es durchaus einen Ermessensspielraum für “sinnvolle” Strukturinformationen.

Weitere Aspekte

Weitere Punkte spielen in der tatsächlichen Anwendung eine Rolle. Ist in der aktuellen Umgebung eher eine interaktive oder eine Batch-orientierte Prüfung sinnvoll? Beim Reporting wird der Detaillierungsgrad der Fehlermeldungen je nach Einsatzzweck unterschiedlich stark sein. Häufig wird auch eine automatische Korrektur im Anschluss an die Prüfung gewünscht.

PDF/A-Validierung: Besonderheiten

Ein PDF/A-Validierer muss das Korrigendum zum Standard berücksichtigen. Dies ist etwa in Acrobat 8 der Fall, nicht aber in Acrobat 7.

Eine unzureichende Breite oder Tiefe der Prüfung wird ein unzuverlässiges Ergebnis liefern.

Es sind unterschiedliche Fehlersituationen beim Validieren möglich. So kann es etwa vorkommen, dass der Validierer einen Fehler meldet, obwohl das Dokument PDF/A-konform ist. Auch das Gegenteil ist möglich, wenn ein Validierer trotz PDF/A-Verletzung keinen Fehler anzeigt.

Andere Probleme in der Validierung können auftreten, wenn ein gültiger Input schon vor der Prüfung abgelehnt wird. Problematisch kann auch ein Report sein, der nicht detailliert genug ist.

Testsuite für PDF/A

PDF/A Competence Center und TWG erarbeiten eine Testsuite mit PDF/A-Dokumenten. Das Ziel sind standardkonforme Dokumente, die viele Aspekte von PDF/A ausloten. Eine Strategie liegt darin, bewusst erzeugte Verletzungen des Standards einzubauen. Die „synthetische Testsuite" wird schrittweise erstellt. Dabei wird die Breite und Tiefe der Abdeckung dokumentiert. Ein wichtiger Punkt ist schließlich der Abgleich mit den am Markt verfügbaren Validierern.

Ausgangspunkte:

Im Blickfeld sind (vermeintlich oder tatsächlich) konforme Erstellungsprogramme, sowie diverse Methoden zur manuellen Erstellung.

Wie überzeugt sich die TWG von der Konformität der Testsuite? Es erfolgt eine Prüfung mit verschiedenen Validierungstools, außerdem wird eine manuelle Inspektion mit speziellen Analysetools durchgeführt.

Für die Herstellung der Testsuite steht der gesammelte Sachverstand der TWG bereit. Ein iteratives Vorgehen sorgt für Genauigkeit.

Validierung der Validierer

Das Anwendungsziel ist die Überprüfung von Validierern durch Anwenden der Testsuite. Eine zunehmende Genauigkeit wird durch den Ausbau der Testsuite erreicht. Die Kriterien zur Bewertung von Validierungstools ergeben sich durch die folgenden Fragen:

- Welche Fehler werden erkannt?
- Welche gültigen Elemente werden irrtümlich abgewiesen?
- Wie aussagekräftig sind die Fehlerreports?

Weiteres Vorgehen

Nach Vervollständigung der Testsuite wird das Testverfahren formalisiert. Die Durchführung der Tests soll durch eine unabhängige Prüfstelle erfolgen. Schließlich kann über die Testsuite die Zertifizierung von Produkten vorgenommen werden.

Praktische Empfehlungen zur Validierung

Bei hohem Volumen muss mit Performance-Einbußen durch die Validierung gerechnet werden. Daher ist es sinnvoller, Produkte und Workflows für die PDF/A-Erstellung überprüfen, statt jedes erzeugte Dokument zu validieren. Extern zugelieferte Dokumente müssen trotzdem in der Regel validiert werden, da der Herstellungsweg in der Regel nicht bekannt ist. Zur Erinnerung: nichtkonforme Dokumente können nicht immer in PDF/A konvertiert werden! Um etwa einen Font nachträglich einzubetten, muss dieser natürlich erst einmal verfügbar sein.

Thomas Merz, PDFlib GmbH, aoe

Anhang: PDFlib GmbH

- Seit 1997 auf PDF-Entwicklung fokussiert
  - Ca. 11 000 Lizenzen in mehr als 60 Ländern
  - Hunderte von OEM-Kunden
- PDFlib-Produktfamilie: Komponenten zur dynamischen Erstellung von PDF
  - PDF/A-1a (Tagged PDF) und PDF/A-1b
- PLOP: Verschlüsselung, Optimierung, Signatur
  - Erhaltung von PDF/A-1 (falls vom Standard erlaubt)
- Text Extraction Toolkit (TET): PDF nach Text/XML/RTF
  - Normalisierung aller Texte zu Unicode
  - Erkennung von Wortgrenzen und Textfluss