# PDF/A: Validatoren

***Wie definiert sich PDF/A? Welche Schwierigkeiten gibt es zu berücksichtigen? Welche gängigen PDF/A-Validatoren gibt es? Wie ist die Genauigkeit dieser PDF/A-Validatoren?***

## 1. Einleitung

Das PDF/A Format ist in den Archiven weit verbreitet und gilt als unbestrittenes Archivformat. Nur stellt sich immer häufiger die Frage, wann ein PDF/A auch wirklich ein PDF/A ist, mit welchen Mitteln dies überprüft werden kann und weshalb die Validierungsergebnisse keine strikten Resultate sein können.

Die Zusammenstellung verschiedener Programme zur Validierung von PDF/A-Dateien soll den Archiven in zweierlei Hinsicht als Handreichung dienen: Erstens zeigt sie auf, welche Validatoren nach heutigem Kenntnisstand für die Umwandlung von Dateien zu PDF/A existieren. Zweitens kann sie als Basis für eine erste Evaluation durch die verschiedenen Archive dienen.

## 2. Definition PDF/A

PDF/A ist ein Portable Document Format, das für die Langzeitarchivierung geschaffen wurde. Das Format wurde im Standard "ISO-19005-1 - Document management – Electronic document file format for long-term preservation" genormt. Dieser Standard entspricht der PDF-Version 1.4. Im Standard wird nur aufgelistet, welche Funktionen der einzelnen PDF-Versionen obligatorisch, empfohlen, eingeschränkt oder verboten sind.

PDF/A existiert in zwei Varianten:

- PDF/A 1a: vollständige Übereinstimmung mit dem Standard
- PDF/A 1b: Mindestanforderungen von PDF/A erfüllt (Barrierefreiheit nicht erfüllt)

## 3. Allgemeine Probleme bei der Validierung

PDF/A ist zwar als ISO-Standard genormt, jedoch wird im Standard nur aufgelistet, welche einzelnen Funktionen von PDF-Version 1.4 obligatorisch, empfohlen, eingeschränkt oder verboten sind. Diese werden vereinzelt in den Details unterschiedlich interpretiert. Zudem sind alle Dokumente *i*, die den PDF/A Standard zusammen definieren, sehr umfangreich und sehr technisch. Entsprechend kann die Beurteilung ohne die Hilfe von PDF/A-Validatoren nur von Experten mit fundiertem Wissen über Seitenbeschreibungssprachen wie PostScript und PDF vollzogen werden.

Die Anzeige des Adobe Readers ist nicht genügend aussagekräftig. Er prüft nur einige wenige Haupteigenschaften von PDF/A. Lässt man die nichtkonformen Testdokumente der Bavaria-Testsuite mit dem Adobe Reader 9.3.3 öffnen, erkennt er von den 226 nicht konformen Dokumente nur gerade deren 4 als nicht PDF/A (<2%). Diese Anzeige kann auch im Adobe Acrobat enthalten sein und entspricht nicht den Preflight Ergebnissen *ii*.

Auf eine systematische Validierung mit speziellen Validatoren sollte deshalb nicht verzichtet werden.

---

*i* Der ISO 19005-1 Standard mit der PDF Referenz 1.4 (ca. 1'000 Seiten) plus zusätzlich die darin referenzierten Dokumente wie Fontformate, XML-Spezifikation, Kompressionsformate, RFCs usw.

*ii* Preflight ist der eingebaute PDF/A-Validator in Adobe Acrobat und wurde von der Firma callas software GmbH entwickelt.

## 4. Beurteilung der PDF/A-Validatoren

Wegen der obenerwähnten Problematik ist die Beurteilung der existierenden PDF/A-Validatoren nicht einfach. Die KOST alleine wäre dazu nicht in der Lage. Sie stützt sich deshalb auf Vorarbeiten der Firma **PDF**lib *iii*, die nicht nur eine öffentliche Bewertung abgegeben, sondern auch die Grundlagen, in welchen das Know-How enthalten ist, online als "Bavaria-Testsuite" zur Verfügung gestellt hat *iv*.

Mit diesen Grundlagen und der vorliegenden Zustimmung von **PDF**lib ist es der KOST möglich, weitere Validatoren zu beurteilen und die existierende Bewertung auszubauen. Zudem wurden die einzelnen Testergebnisse zusammengefasst und grob beurteilt. Das Ergebnis ist in der Tabelle auf Seite 3 ersichtlich.

## 5. Anmerkungen zum Bavaria Report und Testsuite von PDFlib

**PDF**lib möchte als Hersteller vom Bavaria Report und Testsuite die Anwender auf folgende Punkte hinweisen:

- Die Testdaten zum Bavaria-Report können aufgrund der Komplexität von PDF nicht alle denkbaren Standardverletzungen überprüfen.
- Es wäre zwar wünschenswert und interessant, die Bavaria-Testsuite zu erweitern, **PDF**lib GmbH hat aber derzeit keine Pläne, die Tests zum Bavaria-Report dahingehend zu erweitern, dass zusätzliche PDF-Elemente getestet werden.
- Aufgrund der Verfügbarkeit der Testdaten seit Veröffentlichung des Bavaria-Reports (April 2009) können Hersteller von PDF/A-Validierungs-Software theoretisch die im Bavaria-Report getesteten PDF-Konstrukte gesondert behandeln, also sozusagen nur nach den getesteten Konstrukten suchen anstatt nach allen kritischen Elementen. Im Ergebnis würde ein Tool zwar gute Test-Ergebnisse erzielen, in der Praxis aber doch nicht unbedingt genau genug arbeiten.
- Aus Aufwandsgründen ist **PDF**lib GmbH nicht in der Lage, für Endanwender Support zur Bavaria-Testsuite und deren Anwendung auf PDF/A-Validierer zu leisten.

## 6. Neue Versionen oder weitere Programme

Weil die Tests sehr aufwendig sind, verzichtet die KOST darauf, jede neue Version separat zu testen. Auch ist es uns nicht möglich, alle existierenden Validatoren abzubilden. Auf Anfrage ist die KOST aber jederzeit gerne bereit, Aktualisierungen oder Ergänzungen vorzunehmen.

---

*iii* Die Grundlagen der Firma **PDF**lib sind vertrauenswürdig, da sie einerseits Mitglied des 'PDF/A Competence Center' ist und anderseits diese Bewertung für ihre eigenen Produkte (Erstellung von PDF/A-Dokumente) benötigt.

*iv* www.pdflib.com/knowledge-base/pdfa/validation-report

## 7. Produktübersicht

| **PDF/A Validatoren** (Zusammenstellung diverser Desktopprogramme) | | **Adobe: Adobe Acrobat 9.1** www.adobe.com | **Callas: pdfaPilot** www.callassoftware.com | **Intarsys: PDF/A Live** www.intarsys.de | **PDF Tools: 3Heights PDF Validator Shell** www.pdf-tools.com | **Seal Systems: PDF Longlife Suite / PDF Checker** www.sealsystems.de | **Solid Documents: Solid Framework** www.soliddocuments.com | **PDFTron: PDF/A Manager** www.pdftron.com |
|---|---|---|---|---|---|---|---|---|
| **Geschwindigkeit & Robustheit:**<br>Sehr gut = <1 Minute & ohne Absturz<br>Gut = 1 - 5 Minuten & ohne Absturz<br>Ausreichend = >5 Minuten & ohne Absturz<br>Mangelhaft = Absturz | | Ausreichend | Ausreichend | **Gut** | **Sehr gut** | **Gut** | Ausreichend | **Sehr gut** |
| **Genauigkeit:** *v*<br>Sehr gut = Mittelwert >=95%<br>Gut = Mittelwert 90% - 94%<br>Ausreichend = Mittelwert 75 - 89%<br>Mangelhaft = Mittelwert <75% | | **Gut** | **Gut** | **Sehr gut** | **Gut** | Ausreichend | Mangelhaft | **Sehr gut** |
| **Isartor testsuite** (non-conforming) | | | | | | | | |
| 6.1 File structure | 31x | 97% | 100% | 90% | 100% | 42% | 100% | 90% |
| 6.2 Graphics | 47x | 100% | 100% | 100% | 100% | 83% | 100% | 100% |
| 6.3 Fonts | 28x | 100% | 100% | 96% | 100% | 50% | 100% | 100% |
| 6.4 Transparency | 6x | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| 6.5 Annotations | 25x | 96% | 100% | 100% | 100% | 100% | 100% | 100% |
| 6.6 Actions | 37x | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| 6.7 Metadata | 27x | 100% | 100% | 100% | 100% | 41% | 37% | 96% |
| 6.9 Interactive Forms | 3x | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| **Other non-conforming** | | | | | | | | |
| ISO 19005 violations | 9x | 89% | 100% | 89% | 89% | 33% | 67% | 100% |
| XMP 2004 violations | 5x | 20% | 40% | 60% | 20% | 40% | 40% | 80% |
| PDF 1.4 violations | 8x | 38% | 38% | 100% | 63% | 25% | 38% | 75% |
| **Conforming** | | | | | | | | |
| Real world | 34x | 88% | 100% | 97% | 85% | 82% | 41% | 100% |
| PDFlib samples | 8x | 100% | 100% | 100% | 88% | 75% | 13% | 88% |
| Advanced XMP | 16x | 100% | 100% | 100% | 100% | 100% | 0% | 100% |
| **Mittelwert** (conforming / non-conforming) | | **91%** | **94%** | **97%** | **90%** | **75%** | **49%** | **95%** |
| **Getestete Version:** | | 9.1.0 | 1.2.077 | 5.0.4 | 1.8.32.1 | 2.1.1.2 | 5.1.168 | 1.00 (CLI) |
| **Tester / Testjahr:** | | **PDF**lib / 2009 | **PDF**lib / 2009 | KOST / 2010 | **PDF**lib / 2009 | **PDF**lib / 2009 | **PDF**lib / 2009 | KOST / 2010 |
| **Bemerkungen:** | | Die Version 9.0 ist bei der Geschwindigkeit und Robustheit mangelhaft. | | Test mit einer neueren Version nochmals komplett durchgeführt. | | | | Report ist sehr übersichtlich (Kompakt mit guter Fehlermeldung). |



*v* Bei der Genauigkeit ist nicht nur das Ergebnis (bestanden / durchgefallen) wesentlich, zusätzlich muss die Fehlermeldung mindestens einen reellen Fehler beschreiben.