# PDF/A: Produktereview batchtauglicher PDF/A-Validatoren

***Wie ist PDF/A definiert? Welche Schwierigkeiten gilt es zu berücksichtigen? Welche gängigen PDF/A-Validatoren gibt es? Wie ist die Genauigkeit dieser PDF/A-Validatoren? Diese Fragen beantwortet die vorliegende Studie.***

***Stand: April 2018***

**Inhalt**

1. Einleitung ..... 1
2. Definition PDF/A ..... 1
3. Allgemeine Probleme bei der Validierung ..... 2
4. Beurteilung der PDF/A-Validatoren ..... 2
5. PDF/A-Validation Benchmarking ..... 2
6. Analyseübersicht ..... 4

## 1. Einleitung

Die Formate PDF/A-1 und PDF/A-2 sind in den Archiven weit verbreitet und gelten als unbestrittene Archivformate. Immer häufiger stellt sich 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 vorliegende Studie dokumentiert und untersucht die wichtigsten aktuellen Programme zur Validierung von PDF/A-Dateien. Sie kann als Basis für eine erste Evaluation durch die verschiedenen Archive dienen. Diese Studie ist die komplett überarbeitete und aktualisierte Version der ersten PDF/A-Validatoren-Studie der KOST von 2010[1].

## 2. Definition PDF/A

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

PDF/A-1 basiert auf PDF-1.4 und existiert in zwei Varianten:

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

PDF/A-2 basiert auf PDF-1.7 und existiert in drei Varianten:

- PDF/A-2a: vollständige Übereinstimmung mit dem Standard
- PDF/A-2u: wie 2b jedoch mit einer Unicode Abbildung (Text durchsuch- und extrahierbar)
- PDF/A-2b: Mindestanforderungen von PDF/A erfüllt (Barrierefreiheit nicht erfüllt)

PDF/A-3 wird für die Langzeitarchivierung nicht empfohlen[2]!

[1] https://kost-ceco.ch/cms/download.php?3fce892c3ec7d401b521e736a0949ba1

[2] Siehe die KOST-Studie „PDF/A-2 und PDF/A-3: Was ist neu?“: https://kost-ceco.ch/cms/index.php?pdf-a-2_3_study_de

## 3. Allgemeine Probleme bei der Validierung

Die Dokumente, die zusammen den PDF/A Standard definieren, sind sehr umfangreich und sehr technisch. Die Beurteilung, ob ein Dokument dem Standard entspricht, kann 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, da er nur den Metadateneintrag von PDF/A prüft. Auf eine systematische Validierung mit speziellen Validatoren sollte deshalb nicht verzichtet werden.

PDF/A ist zwar als ISO-Standard genormt, jedoch wird im Standard nur aufgelistet, welche einzelnen Funktionen von PDF obligatorisch, empfohlen, eingeschränkt oder verboten sind. Diese Vorgaben werden vereinzelt in den Details unterschiedlich interpretiert.

## 4. Beurteilung der PDF/A-Validatoren

Wegen der obenerwähnten Problematik ist die Beurteilung der existierenden PDF/A-Validatoren nicht einfach. Es gibt grundsätzlich zwei Möglichkeiten, Validatoren zu testen:

- Prüfung gegen eine Testsuite mit erwarteten Ergebnissen
- Validatoren-Benchmarking

In der ersten Version der PDF/A-Validatoren-Studie aus dem Jahr 2010 konnte die KOST die zu diesem Zeitpunkt relativ unbekannte "Bavaria-Testsuite" der Firma PDFlib verwenden. Für die Neuauflage im Jahr 2017 wurde mangels einer neuen Testsuite ein Validatoren-Benchmarking durchgeführt.

## 5. PDF/A-Validation-Benchmarking

### a) Testset

Das Testset besteht aus diversen realen PDF-Dateien aus dem KOST-Umfeld und der nestor-AG Formaterkennung. Dieses Testset beinhaltet teils vertrauliche Dokumente und kann deshalb weder publiziert noch ausserhalb der KOST-Geschäftsstelle zur Verfügung gestellt werden.

Das Testset besteht aus 2980 verschiedenen PDF-Dateien, welche sich wie folgt auf die verschiedenen Ausprägungen von PDF und PDF/A aufteilen:

- 1a: 182 Dateien
- 1b: 2137 Dateien
- 2a: 51 Dateien
- 2b: 471 Dateien
- 2u: 98 Dateien
- pdf: 41 Dateien

### b) Durchführung

Das ursprüngliche PDF/A-Validation-Benchmarking wurde durch die KOST mit folgender Testumgebung durchgeführt:

- Windows 7 Enterprise mit SP1
- 32 Bit-Betriebssystem
- 8.00 GB RAM (3.41 GB verwendbar)
- JRE 1.8.0_131 x86

Die verwendeten PowerShell-Scripte sowie die einzelnen Ergebnisse sind auf GitHub ersichtlich: github.com/nestorFormatGroup/PdfaValidationBenchmarking

### c) Validatoren

Beim PDF/A-Validation-Benchmarking wurden folgende Validatoren getestet:

- *veraPDF* (Version 1.6.1, 1.8.4 und 1.10.4)
- *pdfaPilot CLI* von callas software GmbH (Version 7.0.267)[3]
- *3-Heights(TM) PDF Validator Shell* von PDF Tools AG (Version 4.9.20.0)
- *PDF/A-Manager* von PDFTron (Version 6.7152209)[3]
- *PDF/A Live! (CLI mit Serverlizenz)* von intarsys (Version 7.0.6.215)[3]

Nicht berücksichtigt wurden die folgenden Validatoren:

- *JHOVE* von OPF: kein PDF/A-Validator
- *KOST-Val* von der KOST: stützt sich auf Drittvalidatoren, welche bereits separat getestet werden
- *Adobe preflight*: entspricht dem pdfaPilot von callas, welcher in den Test aufgenommen ist

### d) Analyse

Für die Analyse hat die KOST mit jedem der vier Validatoren das gesamte Testset validiert und die Resultate analysiert. Die Analyse umfasste die folgenden Kriterien:

- Kosten: Preis des Produkts inklusive Wartungsvertrag für ein Jahr, gemäss Angaben der Hersteller.
- Geschwindigkeit: Dauer der kompletten Validierung in der beschriebenen Testumgebung. Eine Änderung in der Testumgebung kann die Geschwindigkeit potentiell stark beeinflussen.
- Robustheit: Anzahl der unkontrollierten Ausgaben im Lauf der kompletten Validierung. Damit sind Meldungen, welche nur via Konsole ausgegeben werden, gemeint, wie zum Beispiel „Out of Memory“ Fehler oder Fehler beim Öffnen der Datei.
- Einigkeit: Im Juni 2017 waren sich die Validatoren bei 82.58% einig und bei lediglich 3.52% gab es kein eindeutiges Resultat (2vs2). Analysiert wurde die Abweichung von der Mehrzahl der anderen Validatoren beim Testergebnis valid oder invalid. Im Juni 2017 waren dies gesamthaft 13.89%. Angegeben wird zusätzlich, wie sich die Abweichungen auf valide und invalide Dokumente verteilen.
- Genauigkeit: Manuelle Qualitätskontrolle über 30 Testdateien, welche alle die Validierung im Juni 2017 nicht bestanden hatten; festgehalten wurde der Prozentsatz der übereinstimmenden Fehlermeldungen. (Der Versuch der nestor-AG Formaterkennung, ein Mapping der Fehlermeldungen zu erstellen, muss als gescheitert betrachtet werden: Ein eindeutiges 1:1 Mapping über alle vier Validatoren hat lediglich zu einer Handvoll Fehlertypen geführt.)

Festgehalten wurden ferner die getestete Version, der Tester und der Testzeitpunkt. Spezielle Ergebnisse wurden in Bemerkungen erläutert.

Die Ergebnisse der Analyse sind in der Übersicht im Kapitel 6 zusammengefasst.

### e) Neue Versionen oder weitere Programme

Weil die Tests sehr aufwendig sind, muss die KOST darauf verzichten, jede neue Version eines Produkts separat zu testen. Eine Ausnahme wurde für veraPDF gemacht, welches ein neues Produkt mit gegenwärtig noch grossen Unterschieden zwischen den Versionen ist. Die zwei getesteten Versionen unterscheiden sich massiv in der Qualität.

Die KOST kann jedoch auf Anfrage gerne neue Produktversionen oder weitere Produkte testen, wenn ihr Aufwand entschädigt wird (neue Version CHF 2‘300, neues Produkt CHF 9‘200). Dies war bislang im Dezember 2017 der Fall mit dem Testen der neuen veraPDF Version 1.10.4 und im April 2018 mit dem Testen des Validators von intarsys.

[3] Bei diesem Produkt handelt es sich um einen Konverter und Validator, was zum Teil die höheren Kosten erklärt.

## 6. Analyseübersicht

| PDF/A Validatoren 2017 | Callas: pdfaPilot | PDF Tools: 3Heights PDF Validator | PDFTron: PDF/A Manager | veraPDF (v1.6.1) | veraPDF (v1.8.4) | veraPDF (v1.10.4) | Intarsys PDF/A Live! |
|---|---|---|---|---|---|---|---|
| **Kosten**<br>Gratis: Kostenlos<br>Gering: 1 - 499 CHF<br>Mässig: 500 - 999 CHF<br>Teuer: > 999 CHF | Teuer [3]<br>EUR 5'399.- | Gering<br>CHF422.- | Mässig [3]<br>USD 699.- | Gratis<br>CHF 0.00 | Gratis<br>CHF 0.00 | Gratis<br>CHF 0.00 | Mässig [3]<br>EUR 474.10 bis EUR 1'200.- |
| **Geschwindigkeit**<br>Sehr gut: ≤ 30 Minuten<br>Gut: 31 - 120 Minuten<br>Ausreichend: 121 - 240 Minuten<br>Mangelhaft: > 240 Minuten | Gut<br>1:58:50 | Sehr gut<br>0:18:00 | Sehr gut<br>0:13:38 | Mangelhaft<br>8:30:20 | Mangelhaft<br>9:22:45 | Ausreichend<br>3:28:33 | Sehr gut<br>0:15:59 |
| **Robustheit**<br>Sehr gut: ≤ 5 unkontrollierte Ausgaben<br>Gut: 6 - 10 unkontrollierte Ausgaben<br>Ausreichend: 11-30 unkontrollierte Ausgaben<br>Mangelhaft: > 30 unkontrollierte Ausgaben | Sehr gut<br>0 | Sehr gut<br>1 | Sehr gut<br>3 | Mangelhaft<br>49 | Mangelhaft<br>55 | Mangelhaft<br>48 | Gut<br>6 |
| **Einigkeit**<br>Sehr gut: < 1.0% Abweichung<br>Gut: 1.0-4.9% Abweichung<br>Ausreichend: 5.0-9.9% Abweichung<br>Mangelhaft: ≥ 10.0% Abweichung | Gut | Sehr gut | Ausreichend | Gut | Mangelhaft | Ausreichend | Gut |
| **Total Abweichung** | 2.38% | 0.87% | 5.74% | 4.90% | 19.53% | 5.91% | 1.51% |
| **Rest Valid** | 0.00% | 0.81% | 5.00% | 2.28% | 10.23% | 3.22% | 1.17% |
| **Rest Invalid** | 2.38% | 0.07% | 0.74% | 2.62% | 09.30% | 2.68% | 0.34% |
| **Genauigkeit**<br>Sehr gut: ≥ 90% treffende Fehlermeldung<br>Gut: 80-89% treffende Fehlermeldung<br>Ausreichend: 70-79% treffende Fehlermeldung<br>Mangelhaft: < 70% treffende Fehlermeldung | Gut<br>87.56% | Sehr gut<br>99.33% | Gut<br>86.78% | Sehr gut<br>93.11% | Mangelhaft<br>28.78% | Sehr gut<br>90.44% | Gut<br>80.06% |
| **Getestete Version** | CLI v7.0.267 | Shell v4.9.20.0 | v6.7152209 | v1.6.1 | v1.8.4 | v1.10.4 | v7.0.6.215 |
| **Tester** | KOST | KOST | KOST | KOST | KOST | KOST | KOST |
| **Testzeitpunkt** | Juni 2017 | Juni 2017 | Juni 2017 | Juni 2017 | August 2017 | Dezember 2017 | April 2018 |
| **Bemerkungen** | Einigkeit und Genauigkeit wären deutlich besser mit der Option „N-Eintrag im OutputIntent prüfen". | | Die Kosten stammen aus dem Jahr 2010. | | Einigkeit und Genauigkeit haben sich in der Version 1.8.4 massiv verschlechtert. Von den 30 invaliden PDF-Dateien wurden 18 als valide ausgegeben. | Geschwindigkeit hat sich in der Version 1.10.4 massiv verbessert.<br>Einigkeit und Genauigkeit ist vergleichbar mit der Version 1.6.1. | Für die Verwendung der CLI ist eine Serverlizenz erforderlich. Diese kostet EUR 1'200.-. |