SLUB
Wir führen Wissen.

Erstellung PDF/A-konformer Dokumente: Einführende Informationen in den PDF/A-Standard

Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden
Dresden, 2018

1 Einleitung

Die Erstellung von PDF/A-konformen Dateien ist in einigen Fällen nicht einfach und sorgt für Unverständnis, wenn entweder keine PDF-Datei nach PDF/A-Standard erstellt werden kann oder dass, wenn es klappt, diese Datei nicht immer der Darstellung der ursprünglichen Datei entspricht. Dies hat bestimmte Ursachen, die hier erläutert werden, um berechtigte Fragen, die beim Erstellen PDF/A-konformer Dateien auftreten können, im Vorfeld zu klären.

In diesem Dokument werden zunächst Informationsquellen zur PDF/A zusammengestellt. Danach wird erklärt was PDF/A ist. Diese Erklärung zielt vor allem darauf ab, die Ursachen zu beschreiben, weshalb der PDF/A-Standard in einigen Fällen schwer erreicht werden kann. Dies wird ergänzt durch die Erläuterungen der unterschiedlichen Spezifikationen des PDF/A-Standards. Abschließend werden Validierungsmöglichkeiten vorgestellt und die Herausforderung einer nachträglichen Konversion beschrieben.

2 Literatur / Informationen

Es gibt viele Online-Informationen zum PDF/A-Standard. Die PDF Association[1], bzw. das PDF/A Competence Center[2] bieten einige Informationen dazu, insbesondere PDF/A kompakt (Drümmer, Oettler and Seggern, 2007) und PDF/A kompakt 2.0 (Oettler, 2013). Hier finden sich aber auch FAQs, weiterführende Links und anderes. Natürlich bietet auch Wikipedia Informationen zum PDF/A-Standard. Dort findet sich ein guter Überblick, der jedoch nicht ausreicht, um das Konzept des Standards zu erfassen. Wichtig (vor allem für das Erstellen von PDF/A-konformen Dateien) sind die aufgelisteten Eigenschaften der PDF/A-Varianten.
Validierung von PDF/A (2011) bietet einen Überblick über die Validierung von PDF/A-konformen Dateien. In diesem Zusammenhang ist auch Frieses (2014) Diskussion über JHOVE[3] als Validierungsinstrument zu nennen.
Zudem gibt es viele Anleitungen, wie PDF/A-konforme Dateien erstellt werden können. Da die Erstellung von PDF/A-konformen Dateien aus LaTeX sehr schwierig ist soll hier exemplarisch auf eine Seite verwiesen werden, die recht gut das Problemfeld umreißt[4].

Bevor erklärt wird, wie PDF/A-konforme Dateien erstellt werden können, wird im folgenden Abschnitt erläutert, was PDF/A bedeutet. Ohne Verständnis für diesen Standard kann die Erstellung von PDF/A-konformen Dateien sehr schwer sein.

[1] http://www.pdfa.org/
[2] http://www.pdfa.org/competence-center/pdfa-competence-center/?lang=de
[3] http://sourceforge.net/projects/jhove/
[4] http://texwelt.de/wissen/fragen/758/pdfa-konforme-dokumente-mit-latex

3 Was ist PDF/A?

PDF/A[5] ist ein spezieller PDF-Standard, der bestimmte Dokumenteigenschaften von einer PDF-Datei verlangt, um eine Wiedergabe des Dokuments auch in einigen Jahrzehnten garantieren zu können. Sind Inhalte in einem Dokument enthalten, die dem Standard nicht entsprechen, kann keine PDF/A-konforme Datei erstellt werden.

> „Ziel des PDF/A-Standards ist, dass PDF-Dokumente erstellt werden können, deren visuelles Erscheinungsbild über die Zeit erhalten bleibt“ (Drümmer, Oettler and Seggern, 2007, p.9).

Was macht eine PDF/A aus? PDF/A kompakt gibt folgende exemplarische Hinweise, was enthalten sein muss und was nicht enthalten sein darf:

> „**Erforderlich**: Ein „Muss“ ist etwa der vollständige Zugriff zu allen zum Dokument gehörenden Elementen. Ein Beispiel: Schriften müssen eingebettet sein, ein Verweis auf die vorgesehene Schrift reicht nicht aus. Hat ein Leser in 10 Jahren die erforderliche Schrift nicht auf dem Rechner, könnten zum Beispiel Sonderzeichen oder Symbole nicht dargestellt werden.“ (Drümmer, Oettler and Seggern, 2007, p.9)
>
> „**Untersagt**: Zudem gibt es PDF-Merkmale, die zu vermeiden sind. Sie sind verboten, weil sie die gewünschte Beständigkeit unterlaufen, wie etwa interaktive Elemente oder PDF-Ebenen. Solche Eigenschaften verhindern die Eindeutigkeit, die eine gültige PDF/A-Datei erreichen muss. Bei einem PDF-Dokument mit Ebenen stellt sich für eine Druckausgabe in 50 Jahren womöglich die Frage, welche Ebene nun gelten soll, und welche nicht. Diese Entscheidung muss jetzt, zum Zeitpunkt der PDF-Erstellung, gefällt werden.“ (Drümmer, Oettler and Seggern, 2007, p.9)

Der Wikipedia-Artikel[6] zum PDF/A-Standard bietet hier detailliertere Informationen.

In diesem Dokument werden PDF-Dateien, die dem PDF/A-Standard entsprechen ab jetzt **PDF/A-Dateien** genannt. Eine PDF/A-Datei ist jedoch kein eigenes Format, sondern eine „normale“ PDF-Datei, die bestimmten Standards genügen muss. Die Bestätigung, dass es sich um eine PDF/A-Datei handelt, erhält man jedoch weder dadurch, dass eine Datei über eine bestimmte Dateiendung verfügt, noch, dass sie sich in einem bestimmten Programm öffnen lässt.
Ein Hinweis, dass zumindest ein Versuch unternommen wurde, eine PDF/A-Datei zu erstellen, gibt der Blaue Balken in PDF-Dateien an, der folgenden Text enthält:

- Die geöffnete Datei entspricht dem PDF/A-Standard. Sie wurde schreibgeschützt geöffnet, um Änderungen zu verhindern (ADOBE ACROBAT PRO)
- Diese Datei verlangt Konformität mit dem PDF/A-Standard und wurde schreibgeschützt geöffnet, um Änderungen zu verhindern (ADOBE READER)

Um sicher zu gehen, dass die Datei wirklich dem PDF/A-Standard entspricht, muss in ADOBE ACROBAT PRO (oder einem anderem Programm) die Konformität geprüft werden.

Letztendlich ist eine PDF-Datei nur PDF-konform, wen sichergestellt wird, **dass sie bestimmten Validierungsroutinen standhält**. Für „Otto-Normalverbraucher“ ist es jedoch nicht einfach zu überprüfen, ob es sich bei einer PDF-Datei nun um eine PDF/A-Datei handelt, denn

[5] http://de.wikipedia.org/wiki/PDF/A
[6] https://de.wikipedia.org/wiki/PDF/A

- Es müssen Validierungsprogramme verfügbar sein
- Es müssen Programme verfügbar sein, die PDF/A-Dateien erzeugen können.

Hierzu finden sich in dem Abschnitt 5 Validierung weitere Informationen.

Zudem ist PDF/A ist auch nicht gleich PDF/A. Es gibt mehrere Versionen, die unterschiedliche Inhalte erlauben:

- PDF/A-1b
- PDF/A-1a
- PDF/A-2a
- PDF/A-2b
- PDF/A-2u

Natürlich gibt es Gründe für die unterschiedlichen Versionen, die für die korrekte Erstellung der Datei wichtig sind. Deshalb sollen hier die Unterschiede erläutert werden, mit dem Ziel, einen praktischen Leitfaden für die PDF/A-Erstellung zu formulieren.

---- Hinweis: Wenn ein Dokument auch in Jahrzehnten noch korrekt dargestellt werden soll, ist es am besten, das Dokument so zu erstellen, dass problemlos eine PDF/A-konforme Datei erstellt werden kann. -----

4 PDF/A-Versionen

4.1 PDF/A-1 – PDF/A-2

Ein Hauptmerkmal ist die zugrundeliegende PDF-Version:

- PDF/A-1 basiert auf PDF 1.4
- PDF/A-2 basiert auf PDF 1.7

Somit sind in PDF/A-1-Dateien nur Inhalte erlaubt, die in PDF 1.4-Dateien enthalten sein dürfen, wohingegen PDF/A-2-Dateien auch Inhalte erlauben, die in PDF 1.7-Dateien enthalten sein dürfen. Wenn Objekte mit Transparenz in der Datei enthalten sind, sollte zum Beispiel PDF/A-2 gewählt werden. Eine Übersicht über die PDF-Spezifikationen findet sich in PDF/A kompakt (Drümmer, Oettler and Seggern, 2007, p.8).
Auch wenn im Gegensatz zu PDF/A-1 einige Elemente zusätzlich erlaubt sind, werden von PDF/A-2 andere Eigenschaften bemängelt. Somit kann keine eindeutige Empfehlung formuliert werden, welche PDF/A-Version genutzt werden soll. Es hängt sehr von den Inhalten der Quelldatei und den Programmen ab, mit denen die Quelldatei erstellt wurde, bzw. mit welchen Programmen die PDF erstellt werden soll. Wenn es möglich ist, sollte folgende Hierarchie beachtet werden, wobei PDF/A-2a die erstrebenswerteste ist:

1. PDF/A-2a
2. PDF/A-2b
3. PDF/A-1a
4. PDF/A-1b

Doch was bedeuten diese Standards? Im folgenden Abschnitt werden diese Spezifikationen erläutert.

4.2 PDF/A-1b, PDF/A-1a

Es gibt zwei Konformitätsebenen von PDF/A-1[7]:

- a (Level A (Accessible) conformance: sowohl eindeutige visuelle Reproduzierbarkeit als auch Abbildbarkeit von Text nach Unicode und inhaltliche Strukturierung des Dokuments, so dass es im Sinne der Barrierefreiheit von einem Screenreader vorgelesen werden kann.)
- b (Level B (Basic) conformance: eindeutige visuelle Reproduzierbarkeit)

Eine Übersicht über die Unterschiede der beiden Konformitätsebenen findet sich in PDF/A kompakt (Drümmer, Oettler and Seggern, 2007, p.13). Zudem kann auf Wikipedia[8] verwiesen werden.
Eine wichtige Eigenschaft von PDF/A-1b ist, dass nur die rein visuelle Reproduzierbarkeit garantiert wird (Drümmer, Oettler and Seggern, 2007, p.13). Somit ist es möglich, dass Textstellen:

- nicht suchbar sind
- nicht durch copy and paste in andere Dokumente übernommen werden können.

Vor allem bei nachträglichen Konversionen muss unter Umständen damit gerechnet werden, dass diese Mängel nicht sofort auffallen, da sie ja für das Auge des Lesers ja korrekt dargestellt werden. Ein ähnliches Verhalten tritt auf, wenn über GHOSTSCRIPT oder POSTSCRIPT PDF/A-Dateien erstellt werden.

4.3 PDF/A-2b, PDF/A-2a

Es gibt drei Konformitätsebenen von PDF/A-2[9]:

- PDF/A-2a: realisiert vollständig alle Anforderungen der ISO 19005-2, insbesondere alle strukturellen und semantischen Eigenschaften.
- PDF/A-2b: Mindestanforderung an eine PDF/A-2 Datei, garantiert das richtige Erscheinungsbild des Dokuments für eine Langzeitarchivierung.
- PDF/A-2u: wie 2b, plus: der gesamte Text ist in Unicode abgebildet, so dass der gesamte Text indexiert und dargestellt werden kann

5 Validierung

Um zu bestimmen, ob eine PDF/A-Datei korrekt ist, beziehungsweise um zu bestimmen, an welchen Inhalten die Erstellung einer PDF/A scheitert, kann die Datei validiert werden. Dazu gibt es unterschiedliche Programme. In diesem Dokument werden nur Validierungsprogramme beschrieben, die frei nutzbar sind, oder die in relativ weit verbreiteten kostenpflichtiger Software enthalten ist.

[7] http://de.wikipedia.org/wiki/PDF/A#PDF.2FA-1
[8] http://de.wikipedia.org/wiki/PDF/A#PDF.2FA-1
[9] http://de.wikipedia.org/wiki/PDF/A#PDF.2FA-2

5.1 veraPDF

> „veraPDF is an open source conformance checker for PDF/A files. It is designed to help archives and libraries check that their PDF/A collections conform to the appropriate ISO 19005 archiving standard specification.“[10]

Dieses Programm kann kostenfrei heruntergeladen und installiert werden[11]. Eine Installationsanleitung ist in der Datei *veraPDFPDFAConformanceCheckerGUI.pdf*[12] unter GitHub[13] enthalten.

Um das Programm mit der grafischen Benutzeroberfläche zu starten, muss auf Windows-Systemen die Datei *verapdf-gui.bat* aufgerufen werden. Diese Datei befindet sich in der Regel in dem Ordner *verapdf* im persönlichen Benutzerverzeichnis, wenn kein anderer Speicherort während der Installation angegeben wurde.

Die Ergebnisse der Validierung können als HTML oder XML-Datei aufgerufen und gespeichert werden. In diesen Reports werden die jeweiligen Fehler beschrieben und die Stellen genannt, an denen diese auftreten.
Eine Korrektur der Fehler mit veraPDF ist nicht möglich.

5.2 3-Heights™ PDF Validator Online Tool

Dieses Online Validierungs-Programm[14] prüft, ob eine PDF-Datei PDF/A-konform ist. Es können alle PDF/A-Spezifikationen geprüft werden. Es wird auch ein Protokoll ausgegeben, das beschreibt, welche Elemente nicht PDF/A-konform sind.
Die Fehler werden jedoch nicht direkt in der Datei lokalisiert, so dass die Stelle in dem Dokument, in der der Fehler auftritt, gesucht werden muss.

5.3 Preflight

Preflight ist ein Validierungstool in Adobe Acrobat Professional. Dieses Tool hat den Vorteil, dass auch die Stellen des Dokuments angezeigt werden, an denen sich der Fehler befindet. Somit ist es recht einfach Korrekturen (Ändern der Schriftart, ...) vorzunehmen. Zudem werden in PDF KOMPAKT (Drümmer, Oettler and Seggern, 2007) die Fehlermeldungen aus Preflight ausführlich kommentiert.

5.4 Jhove

Jhove ist ein Programm mit dem unterem anderem auch PDF-Dateien validiert werden können. Für die Validierung von PDF/A-Dateien ist es nicht wirklich geeignet, trotzdem soll es der Vollständigkeit halber hier aufgeführt werden. Friese (2014) hat die Verwendbarkeit von Jhove recht verständlich erläutert.

5.5 Weitere

Es gibt weitere Validierungssoftware (PDF/A Competence Center, 2011), doch diese sind kostenpflichtig und können hier nicht vorgestellt werden.

---

[10] http://openpreservation.org/about/projects/verapdf/
[11] http://verapdf.org/software/
[12] https://github.com/veraPDF/veraPDF-library/blob/integration/veraPDFPDFAConformanceChecker-GUI.pdf
[13] https://github.com/veraPDF/veraPDF-library#install-from-zip-package
[14] http://www.pdf-tools.com/pdf/validate-pdfa-online.aspx

5.6 Hinweise

Es ist jedoch möglich, dass bestimmte Inhalte nicht als PDF/A abgespeichert werden können. Hier muss entschieden werden, was wichtiger ist:

- Bewahrung des originären Inhalts und dessen Darstellung -> Gefahr, dass das Dokument nach einigen Jahren nicht mehr dargestellt werden kann
- Bewahrung des Inhalts, der über lange Zeit dargestellt und angezeigt werden kann -> Verlust des Inhalts, der dieses verhindert.

6 Konversion

In einigen Fällen ist es nicht möglich, aus bestehenden Dateien eine PDF/A zu erstellen, weil zum Beispiel die Quelldatei nicht mehr existiert, oder weil ein Programm keine PDF/A-Dateien erstellen kann. In diesen Fällen kann eine PDF/A-Datei nur durch Konversion aus einer „normalen“ PDF erstellt werden. Dabei muss damit gerechnet werden, dass einige Funktionen oder Eigenschaften nicht mehr verfügbar sein werden, wie zum Beispiel:

- Volltextdurchsuchbarkeit
  Bei PDF/A-1b-Dateien werden Schriften, die nicht eingebettet werden können als Bild eingefügt, um das Dokument zumindest visuell zu erhalten. Dadurch lassen sich die betreffenden Textabschnitte nicht mehr durchsuchen und werden somit bei einer Suche auch nicht gefunden.
- Verfälschung von Bildern
  Wenn Bilder in einem Dokument nicht dem PDF/A-Standard entsprechen (Komprimierung, Transparenz, ...), werden diese nicht mehr korrekt dargestellt.

Es muss also vor der Konversion überlegt werden, was wichtiger ist:

- Bewahrung des originären Inhalts und dessen Darstellung -> Gefahr, dass das Dokument nach einigen Jahren nicht mehr dargestellt werden kann
- Bewahrung des Inhalts, der über lange Zeit dargestellt und angezeigt werden kann.

7 Schlussfolgerungen

Auch wenn es schon häufig im Text genannt wurde, muss hier nochmal mit Nachdruck darauf verwiesen werden, dass es bei der Erstellung von PDF/A-Dateien vor allem darauf ankommt, die Inhalte über lange Zeit darstellen zu können!
Dies ist nicht mit allen Inhalten in PDF-Dateien möglich, so dass beim Erstellen von Dokumenten darauf geachtet werden muss, „einfache“ Inhalte zu verwenden. Ansonsten kann diese Datei nicht in ein Langzeitarchiv aufgenommen werden, das dafür garantiert, dass die Dateien auch nach Jahrzehnten dargestellt werden können.
Dies betrifft unter anderem Inhalte wie

- Schriften
- Grafiken
- Metadaten

Bei Schriften sollte darauf geachtet werden, dass:

- keine proprietären Schriften verwendet werden
- keine speziellen Sonderzeichen verwendet werden
- so „normale“ Schriften wie möglich verwendet werden
- darauf geachtet wird, dass die jeweilige Software die Schrift auch tatsächlich komplett und nicht nur als Untergruppe einbettet

Bei Grafiken sollte darauf geachtet werden, dass:

- bei PDF/A-1 keine Transparenz erlaubt ist
- bei PDF/A-1 keine JPEG-2000-Kompression erlaubt ist
- Schriften in Grafiken häufig nicht korrekt eingebettet werden

Es ist generell zu empfehlen, Grafiken als einfaches Bild in eine PDF/A-Datei zu integrieren. Diese sind für den Menschen lesbar und bieten wenige Fehlerquellen. Dadurch können jedoch textliche Inhalte nicht mehr durch Volltextsuche gefunden werden, bzw. durch copy and paste extrahiert werden.

Bei Metadaten kann nur empfohlen werden, in der jeweiligen Erstellungssoftware zu prüfen, dass nur Standardmetadaten eingefügt werden. Diese sollten PDF/A-konform sein. Auf Grund der Vielfalt der Programme und Erstellungsroutinen sind die Metadaten recht individuell, was pauschale Hinweise zur Ursachenbehebung verhindert.

8 Literatur

Drümmer, O., Oettler, A. and Seggern, D. von, 2007. *PDFA kompakt: digitale Langzeitarchivierung mit PDF*. {online} Callas Software GmbH. Available at: ‹http://www.pdfa.org/wp-content/uploads/2011/08/PDFA_kompakt_pdfa1b.pdf›.

Friese, Y., 2014. *Langzeitverfügbarkeit sichern: PDF-Validierung durch JHOVE? | PDF Association*. {online} Available at: ‹http://www.pdfa.org/2014/12/langzeitverfugbarkeit-sichern-pdf-validierung-durch-jhove/?lang=de› {Accessed 25 Mar. 2015}.

Oettler, A., 2013. *PDF/A kompakt 2.0 : PDF für die Langzeitarchivierung*. Available at: ‹http://www.pdfa.org/wp-content/uploads/2013/03/PDFA-kompakt-2_0_screen.pdf›.

PDF/A Competence Center, 2011. *Validierung von PDF/A | PDF Association*. {online} Available at: ‹http://www.pdfa.org/2011/09/validierung-von-pdfa/?lang=de› {Accessed 27 Mar. 2015}.