SLUB
Wir führen Wissen.

Handreichung Publikationen und Pflichtexemplare

SLUB Dresden

Version 1.1.1, 2022-03-28

Inhaltsverzeichnis

Vorwort . . . 1
Allgemeines . . . 1
Abgrenzung . . . 1
Nutzungsszenarien und signifikante Eigenschaften . . . 1
- Lesen und Anschauen . . . 2
- Einfache Nachnutzung der Information . . . 2
- Analyse . . . 2
- Bibliographische Einordnung . . . 2
- Barrierefreier Zugang . . . 2

Anforderungen . . . 3
- Allgemein . . . 3
- Metadaten . . . 3
- Datenstruktur . . . 4
- Grafische Darstellung . . . 4
- Farben und Farbräume . . . 5
- Schriftarten (Fonts) . . . 5
- Glyphen . . . 6
- Bilder . . . 6
- Eigenständige grafische Objekte (XObject) . . . 6
- Dokumentenstruktur . . . 6
- Inhalt / Text . . . 7
- Zulässige PDF/A Formate . . . 7

Quellenverweise . . . 8

Release Note 1.1.1 vom 2022-03-28

- Ergänzung PDF/UA
- Update Struktur

Vorwort

Dieses Dokument richtet sich an Produzenten, die digitale Objekte in das SLUBArchiv.digital einliefern und diese langfristig benutzbar erhalten wollen.

Allgemeines

Elektronische Publikationen und Pflichtexemplare sind oft als Portable Document Format (PDF) gespeichert. Dieses Format bietet als offener Standard die Möglichkeit, elektronische Dokumente unabhängig von der verwendeten Hard- und Software auszutauschen. Eine PDF Datei umfasst genau ein Dokument und die vollständige Beschreibung des Seiteninhaltes mit festem Layout, Text, Schriften, Abbildungen und weiteren Informationen, die für die Darstellung eines Dokuments erforderlich sind.

Um auch in Zukunft eine originalgetreue Darstellung des Dokuments und seiner Inhalten garantieren zu können, müssen bestimmte Vorgaben eingehalten werden. Diese Vorgaben werden durch die PDF/A Standards und die darin enthaltenen Normen für die Langzeitarchivierung beschrieben.

Vorbehaltlich gesonderter Absprachen mit dem SLUBArchiv sind von dieser Handreichung abweichende Änderungen **nicht** gestattet. Informationspakete, die diese Anforderung nicht erfüllen, sind nicht langzeitarchivfähig und werden vom SLUBArchiv zurückgewiesen.

Abgrenzung

Dieses Dokument beschreibt die Anforderungen des SLUBArchiv an Monographien und vergleichbare Schriften, die als elektronische Publikationen und Pflichtexemplare vorliegen und in das SLUBArchiv aufgenommen werden sollen.

Nicht von diesem Dokument erfasst sind Materialien,

- die als OpenAccess Publikationen basierend auf HTML/JATS vorliegen
- die als Druckvorstufe basierend auf PDF/X generiert wurden
- die Abzüge von Webseiten enthalten (WebArchiv)
- die auf PDF Varianten PDF/X, PDF/E, PDF/H, PDF/VT basieren

Nutzungsszenarien und signifikante

Eigenschaften

Auf Grundlage bei der SLUB eingelieferter Publikationen, Dissertationen, E-Journals und elektronischen Pflichtexemplaren wurden deren Nutzungsszenarien ermittelt. Aus diesen Nutzungsszenarien wurden die im Langzeitarchiv dauerhaft zu erhaltenden (d.h. signifikanten) Eigenschaften abgeleitet. Sie bilden die Grundlage für den verbindlichen Regelsatz, der unten näher erläutert wird.

Im Folgenden sind die vom SLUBArchiv adressierten Nutzungsszenarien gemeinsam mit den zu erhaltenden Eigenschaften aufgeführt, diese sind aus https://git.slub-dresden.de/digital-preservation/significantproperties/-/blob/master/sigprops_publikation_pflichtexemplare.xml abgeleitet.

Lesen und Anschauen

- Erhalt des optischen Eindrucks
- Erhalt der Seitenanzahl
- Erhalt der Navigationselemente (Lesezeichen, Links)
- Erhalt der logischen Struktur
- Erhalt kodierter Informationen

Einfache Nachnutzung der Information

- Verwendbarkeit von Inhalten durch Kopieren und Einfügen in andere Dokumente

Analyse

- Erhalt der Möglichkeit zur Volltextsuche innerhalb des Dokuments

Bibliographische Einordnung

- Erhalt bibliographischer Metadaten zum Aufbau eines rudimentären Katalogs:
  - Autor
  - Erscheinungsjahr
  - Titel
  - Medienart
- Erhalt persistenter Identifikatoren (PPN, URN, DOI)

Barrierefreier Zugang

- Erhalt Inhaltsstruktur (Titelhierachie, Tabellen, Listen)
- Erhalt der Lesezeichen zum Navigieren

- Erhalt der Lesereihenfolge (vor allem bei mehrspaltigen Text)
- Erhalt der Sprachdefinition (vor allem bei mehrsprachigen Text)
- Alternativer Text für Bilder und Grafiken
- Erhalt von Artifakte
- Erhalt der standartisierte Zeichenkodierung (Unicode - UTF-8)
- Erhalt der Metadaten (Informationen über das Dokument)

Anforderungen

Allgemein

Verpflichtend

Die Regelsätze des SLUBArchivs bauen auf dem Dateiformat PDF/A und PDF/UA einschließlich deren Spezifikationen auf. Entsprechend fordert das SLUBArchiv deren Einhaltung. Zusätzlich beinhaltet das SLUBArchiv-spezifische Einschränkungen sowie Klarstellungen hinsichtlich der Funktionalität von PDF-Dokumenten.

Für die Regelsätze des SLUBArchivs existiert eine technische Implementierung in Form eines jeweiligen Profils für das Validierungswerkzeug veraPDF [https://verapdf.org/software/]. Es kommt für die maschinelle Prüfung der Archivfähigkeit von PDF/A Dokumenten und barrierefreien PDF/UA Dokumenten zum Einsatz.

Für den PDF/UA-1 Standard reicht die technische Prüfung mittels des Validierungswerkzeug nicht aus. Zum Beispiel kann technisch überprüft werden, ob es einen Alternativtext zu einer grafischen Darstellung gibt, aber die semantische Korrektheit, also die inhaltlich richtige und vollständige Beschreibung kann nur manuell geprüft werden. Deshalb sind folgende Element zu prüfen:

- Zuordnung Inhaltsdokumente den richtigen Tags
- Überschriften, Listen richtig getaggt
- Gibt es leere Tags
- Kontrolle der Vorlese-Reihenfolge im Tag Baum
- Prüfung der Umfließansicht, Kontrastmodus, Navigation über Lesezeichen

Metadaten

Verpflichtend

Alle notwendigen Basisinformationen für die Identifizierung eines PDF Dokuments sind zu erfassen. Die bibliographischen Metadaten umfassen dabei Titel, Verfasser und Schlüsselwörter.

Zusätzlich sind die PDF/A und PDF/UA Version und die Anwendung, mit der die PDF Datei erstellt wurde, abzulegen.

Für alle Metadaten gilt:

- Metadaten müssen unkomprimiert und unverschlüsselt sein.
- Metadaten sind im XMP (Extensible Metadaten Platform) Standard eingebettet abzulegen.
- Für erweiterte Anforderungen sind definierte und konforme Extension-Schemas als Container-Schema mit Namen und Beschreibung aller Eigenschaften und Datentypen zu verwenden.
- Metadaten, die im Dokument als Dokumentinformation abgelegt sind, werden nicht ausgewertet und berücksichtigt.
- Metadaten zu verwendeten Bilder sollen direkt am Bildobjekt und in den dafür von PDF verwendeten Objekttypen abgelegt werden.
- Metadatenstream im Katalog Verzeichnis muss vorhanden sein und muss einen Eintrag dc:title beinhalten (s. PDF/UA-1[e])

Datenstruktur

Verpflichtend

Generell sind nur Objekte oder Datentypen zu verwenden, die im PDF Standard oder in der entsprechenden PDF/A Spezifikation beschrieben sind. Die durch den PDF Standard beschriebene Dateistruktur und Syntax ist strikt einzuhalten. Teilweise zeigen sich einzelne Anwendungen gegenüber diesbezüglichen Abweichungen fehlertolerant. Allein die Einhaltung des Standards ist jedoch maßgeblich für die Archivfähigkeit.

Das PDF darf nicht passwortgeschützt sein. Einschränkungen durch Mechanismen des Digital Rights Management dürfen nicht existieren.

Gute Praxis

Es sollen möglichst keine unreferenzierten Objekte in der PDF Datei vorhanden sein, da sie die Darstellung des Inhaltes beeinflussen können.

Bei jeder Änderung des primären Dokuments sollte eine neue PDF Datei produziert werden. So wird verhindert, dass inkonsistente und unübersichtliche, inkrementelle Updates in der Datenstruktur des PDF Dokumentes abgelegt werden. Beispiele hierfür sind Referenzierung von Löschungen bzw. Änderungen von Seiten.

Grafische Darstellung

Verpflichtend

Die grafische Darstellung bezieht sich auf das Erscheinungsbild der einzelnen Seiten des Dokuments, d.h. auf deren Aussehen und die darauf verankerten Objekte. Es werden Stream-Objekte genutzt. Stream-Objekte beschreiben durch eine Sequenz von Befehlen die Darstellung und werden von einem PDF Reader nacheinander abgearbeitet.

Gute Praxis

Für die Barrierefreiheit müssen Inhalte der Grafik oder der grafischen Darstellung (Art und Zweck) so bereitgestellt werden, dass diese durch Vorlesefunktionen zugänglich sind (s. {pdfua1}). D.h. diese müssen mit einem Figure-Tag gekennzeichnet werden und eine alternative Darstellung oder einen Ersatztext haben.

Bildunterschriften müssen den Caption-Tag tragen.

Semantische Zusammenhänge zwischen grafischen Darstellungen müssen ausgezeichnet sein, um diese Informationen ebenfalls barrierefrei zu gestalten.

Farben und Farbräume

Verpflichtend

Jedes verwendete Farbprofil ist in das PDF Dokument als Profil nach dem Standard "International Color Consortium (ICC)" einzubetten. Folgende ICC Spezifikationen sind zugelassen: ICC.1:1998-09 [h], ICC.1:2001-12[i], ICC.1:2003-09[j] oder ISO 15076-1:2010[l]. Der Ausgabefarbraum ist mittels "OutputIntent" anzugeben[1].

Gute Praxis

Der Einsatz von Prozess- und Schmuckfarben sollte vermieden werden.

Schriftarten (Fonts)

Verpflichtend

Um sicherzustellen, dass der Textinhalt und die semantischen Eigenschaften jedes Zeichens bei der Wiedergabe der Originaldatei übereinstimmen, ist es notwendig, folgendes einzuhalten:

- Fonts sind in das PDF entsprechend PDF 1.7[a], 9.9 einzubetten.
- Ein eingebetteter Font muss alle Glyphen für alle Zeichen enthalten, die im Text des Dokuments verwendet werden. Laut ISO für PDF/A sind Untergruppen (Subsets) von Fonts erlaubt. Es muss aber sichergestellt werden, dass die eingebettete Schriftart für alle verwendetem Zeichen im PDF Text eine Glyphendefinition beinhaltet. Ausnahmen bilden CID Fonts[2]. Hier müssen alle im Font erlaubten Glyphen innerhalb des PDF Dokuments abgelegt werden (s. PDF/A-2[c], NOTE 2 und NOTE 3).

Gute Praxis

Es sollten nur Fonts benutzt werden, die für eine unbegrenzte, universelle Wiedergabe frei in das PDF eingebettet werden können und dürfen.

Glyphen

Verpflichtend

Glyphen-Metriken sind einzuhalten und so im PDF abzubilden, wie die Metriken (z.B. Breite) im Zeichensatz angeben sind.

Alle referenzierten Glyphen müssen auf ihren Unicode Wert abgebildet werden (s. PDF/UA-1[e]).

Es dürfen keine Referenzen auf das .notdef Glyphen vorkommen.

Bilder

Verpflichtend

Für im PDF Dokument verwendete Bilder dürfen keine Alternativbilder im gleichen Dokument[3] referenziert werden. Für Bilder dürfen keine OPI Informationen[4] verwendet werden. Für die Kompression ist JPEG2000[m] als Baseline JPX zu verwenden (M 9.2).

Eigenständige grafische Objekte (XObject)

Verpflichtend

- Formulare dürfen keine OPI Informationen und keine Beschreibung durch PostScript verwenden.
- Referenzobjekte und PostScript-Objekte dürfen nicht verwendet werden.

Dokumentenstruktur

Verpflichtend

Die Dokumentenstruktur ist besonders wichtig, um PDF Dokumente Barrierfrei zugänglich zu erhalten und bildet die logische Reihenfolge von PDF Objekten ab.

- Der Inhalt wird im Strukturbaum mit semantisch geeigneten Tags in einer logischen Lesereihenfolge abgebildet.
- Grundsätzlich sind die Standard-Tags zu verwenden, die in (s.{pdfa}) festgelegt sind.
- Inhalte, die keine Aussage haben, sondern als Schmuck gelten (Artifakte), dürfen nicht in dem Strukturbaum auftauchen.
- Kopf- und Fußzeilen (Pagination) werde grundsätzlich als Artifakte gekennzeichnet und kommen nicht im Strukturbaumvor.

Inhalt / Text

Verpflichtend

- Der Inhalt eines PDF Dokuments und dessen logische Lesereihenfolge müssen erhalten bleiben.
- Zeichen müssen als Unicode abgebildet sein, damit sie immer dargestellt und ausgelesen werden können.
- Für den gesamten Text muss die verwendete Sprache eingestellt sein. Sprachwechsel müssen gekennzeichnet sein.
- Alle Überschriften werden als solche getaggt.
  - Dabei wird die hierarchische Gliederung durch die Tags <H1> bis <H6> abgebildet.
  - Überschriften folgen einer numerischen Reihenfolge.
  - Es dürfen keine Hierarchien übersprungen werden.
  - Es können zusätzliche Ebenen <H7> bis <Hn> erstellt werden, diese dürfen nur als arbasiche Ziffern verwendet werden.
- Ligaturen und besondere Glyphen müssen als solche gekennzeichnet werden. Sie brauchen einen erklärenden Alternativtext.
- Aufzählungen müssen als Liste gekennzeichnet sein. Jeder Listeneintrag muss als einzelner Listeneintrag gekennzeichnet sein. Nummerierte Listen werden imListeneintrag als solche gekennzeichnet.
- Fuß- bzw.Endnoten und Referenzen müssen als solches durch eine Note-Tag gekennzeichnet werden.
- Mathematsiche Ausdrücke müssen durch ein Formula-Tag umschlossen sein, der einen erklärenden Alternativtext besitzt.
- Alle Verzeichnisse,Inhaltsverzeichnis, Tabellen- uns Abbildungsverzeichnis müssen verlinkt sein. Die internen Links in allen Verzeichnissen müssen barrierefrei sein, also einen Alternativtext haben.

Zulässige PDF/A Formate

Als Erweiterung zu dem PDF/A Standard existiert der PDF/UA-Standard ("Universal Accessibility"), der Menschen mit Sehbehinderung PDF Dokumente barrierefrei zugänglich macht und somit Teihabe und Inklusion ermöglicht.

Für die Ablieferung von Publikationen und Pflichtexemplaren in das Langzeitarchiv der SLUB (SLUBArchiv) werden folgende Standards als langzeitarchivfähige Dateiformate festgelegt:

- PDF/A-1[b] in den Konformitätsstufen A und B als Dateiformat für die Langzeitverfügbarkeit auf Basis der PDF Spezifikation PDF 1.4[5]
- PDF/A-2[c] in den Konformitätsstufen A und B als Dateiformat für die Langzeitverfügbarkeit auf Basis der PDF Spezifikation PDF 1.7[a].
- PDF/UA-1[e] als barrierefreies und langzeitverfügbares Dateiformat auf Basis der PDF

Spezifikation PDF 1.7[a] und den Web Content Accessibility Guidlines {WCAG 2.0}

Der Standard für PDF/A-3[d] kommt nicht zur Anwendung, da er nach Auffassung des SLUBArchiv Eigenschaften[6] mitbringt, die seine Eignung als langzeitarchivfähiges Dateiformat in Frage stellen.

Die Abbildung der detaillierten veraPDF-Regeln auf die Spezifikationen von PDF und PDF/A wird in einer der kommenden Veröffentlichungen ergänzt.

Quellenverweise

- [a] ISO 32000-1:2008, Document management — Portable document format — Part 1: PDF 1.7
- [b] ISO 19005-1 (First Edition: 2005-10-01), Document management — Electronic document file format for long-term preservation — Part 1: Use of PDF 1.4 (PDF/A-1)
- [c] ISO 19005-2 (First Edition: 2011-07-01), Document management — Electronic document file format for long-term preservation — Part 2: Use of ISO 32000-1 (PDF/A-2)
- [d] ISO 19005-3:2012, Document management — Electronic document file format for long-term preservation — Part 3: Use of ISO 32000-1 with support for embedded files (PDF/A-3)
- [e] ISO 14289-1:2016-12, Document management applications - Electronic document file format enhancement for accessibility - Part1: Use of ISO 32000-1 (PDF/UA-1)
- [f] XMP (2004)
- [g] XMP (2005)
- [h] ICC.1:1998-09 File Format for Color Profiles (Version 2.2.0)
- [i] ICC.1:2001-12 File Format for Color Profiles (Version 4.0.0)
- [j] ICC.1:2003-09 File Format for Color Profiles (Version 4.1.0)
- [k] Specification ICC.1:2010 (Profile version 4.3.0.0) Image technology colour management — Architecture, profile format, and data structure
- [l] ISO 15076-1:2010 Image technology colour management — Architecture, profile format and data structure — Part 1: Based on ICC.1:2010[k]
- [m] ISO/IEC 15444-2:2004 Information technology — JPEG 2000 image coding system: Extensions — Part 2

[1] z.B. Bildschirm "Adobe RGB(1998)"
[2] Type1 und TrueType Fonts
[3] z.B. mit anderem Farbraum, anderer Auflösung
[4] z.B. für das Ersetzen niedrig aufgelöster Bilder für die Layoutgestaltung durch hoch aufgelöste Bilder aus dem Dateisystem während der Ausgabe
[5] z.B. via Adobe Acrobat 5
[6] hauptsächlich die Einbettung beliebiger Dateiformate