SLUB
Wir führen Wissen.

Handreichung Publikationen und Pflichtexemplare

SLUB Dresden

Version 1.0, 2019-08-09

Release Note 1.0 vom 2019-08-09

- Initial

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.

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[1]
- PDF/A-2[c] in den Konformitätsstufen A und B als Dateiformat für die Langzeitverfügbarkeit auf Basis der PDF Spezifikation PDF 1.7[a].

Der Standard für PDF/A-3[d] kommt nicht zur Anwendung, da er nach Auffassung des SLUBArchiv Eigenschaften[2] mitbringt, die seine Eignung als langzeitarchivfähiges Dateiformat in Frage stellen.

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.

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)

Regelsatz für Publikationen, Dissertationen, E-Journals und elektronische Pflichtexemplare

Der Regelsatz des SLUBArchivs baut auf dem Dateiformat PDF/A und dessen Spezifikationen auf. Entsprechend fordert er deren Einhaltung. Zusätzlich beinhaltet er SLUBArchiv-spezifische Einschränkungen sowie Klarstellungen hinsichtlich der Funktionalität von PDF-Dokumenten.

Für den Regelsatz des SLUBArchivs existiert eine technische Implementierung in Form eines Profils für das PDF/A-Validierungswerkzeug veraPDF [https://verapdf.org/software/]. Es kommt für die maschinelle Prüfung der Archivfähigkeit von PDF/A Dokumenten zu Einsatz.

Metadaten

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 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.

Datenstruktur

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.

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

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.

Farben und Farbräume

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 [g], ICC.1:2001-12[h], ICC.1:2003-09[i] oder ISO 15076-1:2010[k]. Der Ausgabefarbraum ist mittels "OutputIntent" anzugeben[3]. Der Einsatz von Prozess- und Schmuckfarben sollte vermieden werden.

Schriftarten (Fonts)

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.
- Es sollten nur Fonts benutzt werden, die für eine unbegrenzte, universelle Wiedergabe frei in das PDF eingebettet werden können und dürfen.
- Ein eingebetteter Font soll 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[4]. Hier müssen alle im Font erlaubten Glyphen innerhalb des PDF Dokuments abgelegt werden (sh. PDF/A-2[c], NOTE 2 und NOTE 3).

Glyphen

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

Bilder

Für im PDF Dokument verwendete Bilder dürfen keine Alternativbilder im gleichen Dokument[5] referenziert werden. Für Bilder dürfen keine OPI Informationen[6] verwendet werden. Für die Kompression ist JPEG2000[l] als Baseline JPX zu verwenden (M 9.2).

Eigenständige grafische Objekte (XObject)

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

Die Abbildung der detaillierten veraPDF-Regeln auf die Spezifikationen von PDF und PDF/A, sowie die Liste der extrahierten technischen Metadaten 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] XMP (2004)
- [f] XMP (2005)
- [g] ICC.1:1998-09 File Format for Color Profiles (Version 2.2.0)
- [h] ICC.1:2001-12 File Format for Color Profiles (Version 4.0.0)
- [i] ICC.1:2003-09 File Format for Color Profiles (Version 4.1.0)
- [j] Specification ICC.1:2010 (Profile version 4.3.0.0) Image technology colour management — Architecture, profile format, and data structure
- [k] ISO 15076-1:2010 Image technology colour management — Architecture, profile format and data structure — Part 1: Based on ICC.1:2010[j]
- [l] ISO/IEC 15444-2:2004 Information technology — JPEG 2000 image coding system: Extensions — Part 2

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