SLUB
Wir führen Wissen.

Langzeitarchivfähige Dateiformate

SLUB Dresden

Version 1.3, 2018-01-08

Release Note 1.3 vom 2018-01-08

- PRONOM-Id für SVG gefixt
- Verweis auf TIFF Handreichung
- Hinweis auf Matroska/FFV1
- Ergänzung Pronom IDs
- Klarstellung HTML

Einleitung

Digitale Langzeitarchivierung stellt besondere Anforderungen an die verwendeten Dateiformate. Digitale Master müssen in offen spezifizierten Dateiformaten vorliegen, die eine leichte Konvertierbarkeit erlauben, um dem Formatwandel über die Jahre Rechnung zu tragen. Gleichzeitig muss dabei der semantische Inhalt auch über mehrere Konvertierungen hinweg erhalten bleiben. Proprietäre Datenformate halten diesen Anforderungen oft nicht stand. Deswegen ist es wichtig, vor der permanenten Archivierung eine Menge an Dateiformaten festzulegen, die für langzeitarchivfähig gehalten wird und deren Einspeisung in das Langzeitarchiv deshalb erlaubt werden darf.

Dabei sollten Dateiformate in der digitalen Langzeitarchivierung folgenden Kriterien genügen:

- offen standardisiert (notfalls offen spezifiziert); kein proprietäres Format
- weit verbreitet
- geringe Komplexität
- ohne Zugriffsschutzmechanismen wie Kopierschutz, Verschlüsselung, DRM
- selbstdokumentierend
- robust
- keine Abhängigkeiten zu anderen Dateiformaten
- lizenzfrei
- validierbar

Die offene Standardisierung von Dateiformaten erlaubt es im Fall der Formatobsoleszenz, Lese- und Konvertierungsprogramme neu zu erstellen. Weitverbreitete Dateiformate bringen es mit sich, dass das Formatwissen über diese stärker verbreitet und für die Nachwelt eher dokumentiert ist. Hinzu kommt, dass die Verfügbarkeit von geeigneter Software mit hoher Wahrscheinlichkeit länger gegeben ist. Die geringe Komplexität eines Dateiformates geht oft mit einer besseren Verständlichkeit und einer weniger fehleranfälligen Implementierung einher. Unter selbstdokumentierenden Dateiformaten versteht man jene, die über eine bestimmte Signatur (magic byte, meist am Dateianfang) eine einfache Identifikation des Dateiformates erlauben. Dateiformate sind dann robust, wenn sich einzelne Bitfehler auf den Inhalt der Datei nur gering oder gar nicht auswirken. Aus diesem Grund wird oft von komprimierten Datenformaten abgeraten. Wenn, wie im Falle des Videodatenformates Matroska/FFV1, verlustfreie Datenkompression mit dem gezielten Hinzufügen von Redundanz, zB. durch CRC, kombiniert wird,

kann die Forderung nach Robustheit erfüllt werden. Sinnvoll ist es, wenn Dateiformate vollständig spezifiziert sind und nicht auf andere Dokumente angewiesen sind. Dies erhöht die Chance Lese- und Konvertierungsprogramme nur anhand der vorhandenen Spezifikation zu erstellen. Lizenzbehaftete, proprietäre Spezifikationen erschweren den Erhalt, da eine Sicherung der Spezifikationsdokumente notwendig ist und Lizenzgeber über die Zeit nicht mehr verfügbar sein können. Dateiformate sollten aus dem selben Grund auch keine Zugriffsschutzmechanismen enthalten, da der Zugriff auf den Inhalt über die Zeit verloren gehen kann (oder gar nicht möglich ist). Sinnvoll ist ebenso, dass für die Dateiformate in der digitalen Langzeitarchivierung geeignete Validatoren oder zumindest Referenzimplementierungen zur Verfügung stehen, die eine Beurteilung ermöglichen, inwieweit Dateien von Programmen interpretiert werden können.

Dieses Dokument listet Dateiformate auf, die nach aktuellem Stand der Bearbeitung durch die SLUB grundsätzlich für die dauerhafte Aufnahme in das Langzeitarchiv der SLUB zugelassen sind.

Für bestimmte Dateiformate hat die SLUB Handreichungen herausgegeben, die die Kriterien für die Aufnahme in das SLUBArchiv genauer spezifizieren.

Formate

PDF/A & PDF/UA (Portable Documents Format)

- **archivfähige Formatversionen:** PDF/A-1 (ISO 19005-1:2005), PDF/A-2 (ISO 19005-2:2011)
- **PRONOM-Formate:** fmt/354, fmt/95
- **Randbedingungen:** (keine)

GIF (Graphics Interchange Format)

- **archivfähige Formatversionen:** GIF87a (nicht: GIF89a)
- **PRONOM-Formate:** fmt/3
- **Randbedingungen:**
  - Bilddaten ausschließlich unkomprimiert

HTML, HTM (HyperText Markup Language)

- **archivfähige Formatversionen:** HTML 3.2 (nur Bestandsdaten), HTML 4.01 strict, HTML 5
- **PRONOM-Formate:** fmt/98 (HTML 3.2), fmt/100 (HTML 4.01), fmt/471 (HTML 5)
- **Randbedingungen:**
  - nur statische Seiten ohne ausführbaren Code
  - Archivierung nur mit zugehörigem CSS (wenn vorhanden)
  - Archivierung vorzugsweise als WARC (fmt/289)

SVG (Scalable Vector Graphics)

- **archivfähige Formatversionen:** SVG 1.0, SVG 1.1, SVG 1.2 Tiny
- **PRONOM-Formate:** fmt/91, fmt/92, fmt/413
- **Randbedingungen:** (keine)

TIF (Tagged Image File Format)

- **archivfähige Formatversionen:** TIFF Rev. 6.0 Part 1 (Baseline TIFF), GeoTIFF, Exif
- **PRONOM-Formate:** fmt/10, fmt/155, x-fmt/387
- **Randbedingungen:**
  - Bilddaten ausschließlich unkomprimiert (TIFF-Tag 259, Datentyp SHORT, Wert 1 (“no compression”))
  - maßgebend ist die *Handreichung TIFF* des SLUB Archiv

WARC (WebARChive)

- **archivfähige Formatversionen:** WARC (ISO 28500)
- **PRONOM-Formate:** fmt/289
- **ACHTUNG:** Die weitere Behandlung dieses Formates ist noch nicht abschließend geklärt (z. B.: Was passiert mit den eingeschlossenen und mglw. nicht archivfähigen Bilddateien?)

XML (eXtensible Markup Language)

- **archivfähige Formatversionen:** (alle) (u. a. Alto-XML, METS, MODS, DC)
- **PRONOM-Formate:** fmt/101
- **Randbedingungen:**
  - UTF-8-codiert
  - well-formed
  - validierbar gegen ein frei zugängliches standardisiertes Schema (RelaxNG, DTD, XML-Schema)

DOCX, XLSX, PPTX (OfficeOpen XML)

- **archivfähige Formatversionen:** docx 2013, xlsx 2013, pptx 2013
- **PRONOM-Formate:** ExL-Fmt-62, fmt/189
- **Randbedingungen:**
  - mit MS Office 2013 oder neuerer Version erstellt (ältere Formate sind nicht standardkonform)

ODF (OASIS Open Document Format for Office Applications)

- **archivfähige Formatversionen:** odt (OpenDocument-Text), ods (OpenDocument-Spreadsheet), odp (OpenDocument-Präsentation), odg (OpenDocument-Grafik)
- **PRONOM-Formate:** fmt/136, fmt/290, fmt/291, x-fmt/3, fmt/137, fmt/294, fmt/295, fmt/138, fmt/292, fmt/293, fmt/139, fmt/296, fmt/297
- **Randbedingungen:** (keine)

PNG (Portable Network Graphics)

- **archivfähige Formatversionen:** png
- **PRONOM-Formate:** fmt/11, fmt/12, fmt/13
- **Randbedingungen:** (keine)

Matroska Multimediacontainer

- **archivfähige Formatversionen:** mkv (FFV1 v3), mka (linear PCM)
- **PRONOM-Formate:** fmt/569
- **Randbedingungen:**
  - mit eingeschaltetem CRC32
  - maßgebend ist die *Handreichung retrodigitalisierter Film und Video* des SLUB Archiv