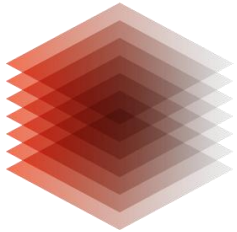
LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK



TIB

wohlgeformt und valide – Formate und Formatvalidierung

Merle Friedrichsen
Frankfurt, 16. Mai 2017
nestor for newbies

Agenda

- Formate
- Formatidentifizierung
- Formatvalidierung
- Ergebnisse der Formatvalidierung
- Demonstration

Format, ist bey den Buchdruckern die Grösse derer Columnen, und bey denen Buchbindern die äusserliche Gestalt und Grösse eines Buchs, was seine Länge und Breite anlanget.

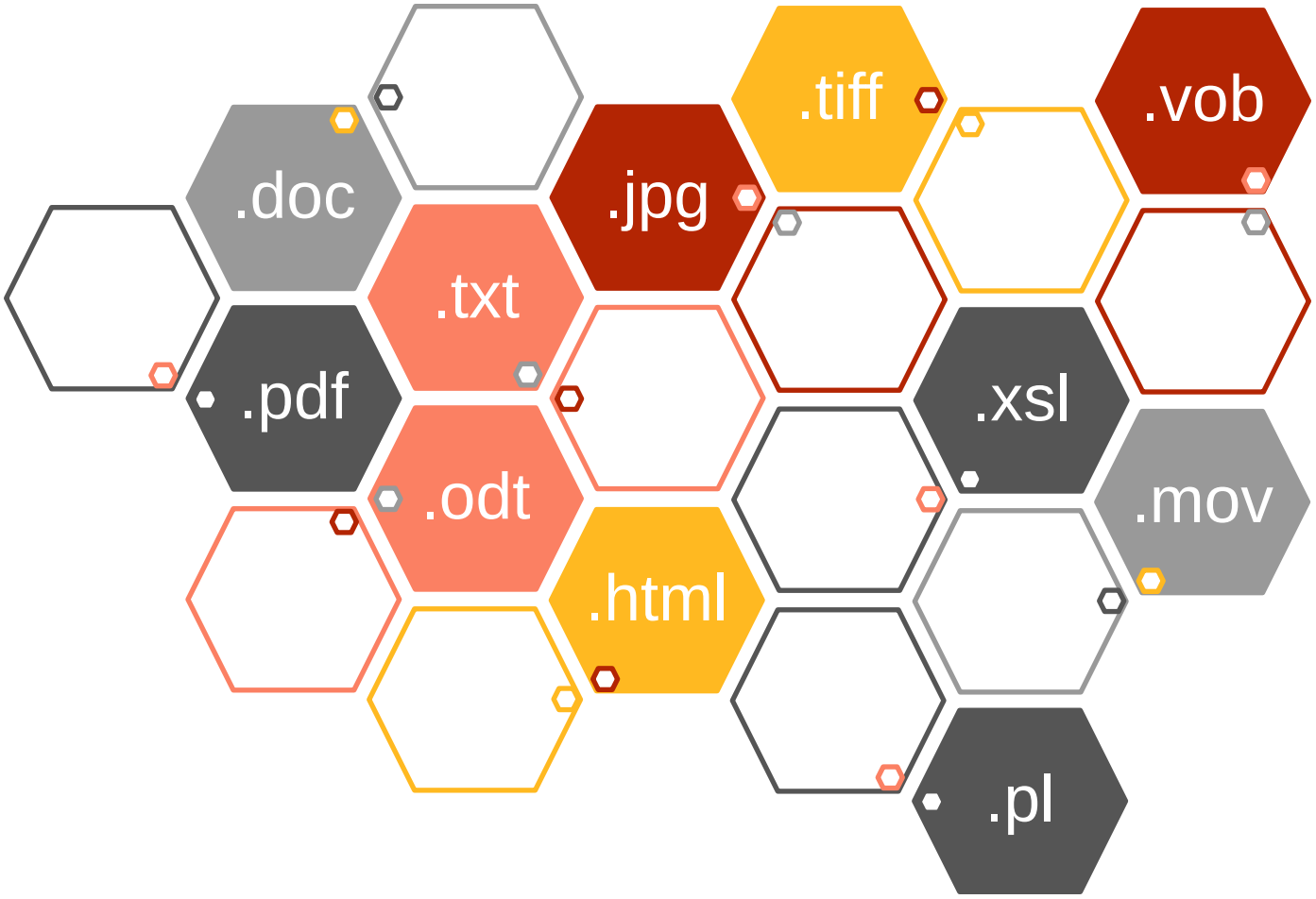
Johann Heinrich Zedler, Grosses vollständiges Universal-Lexicon, 1731 -1754

Formate

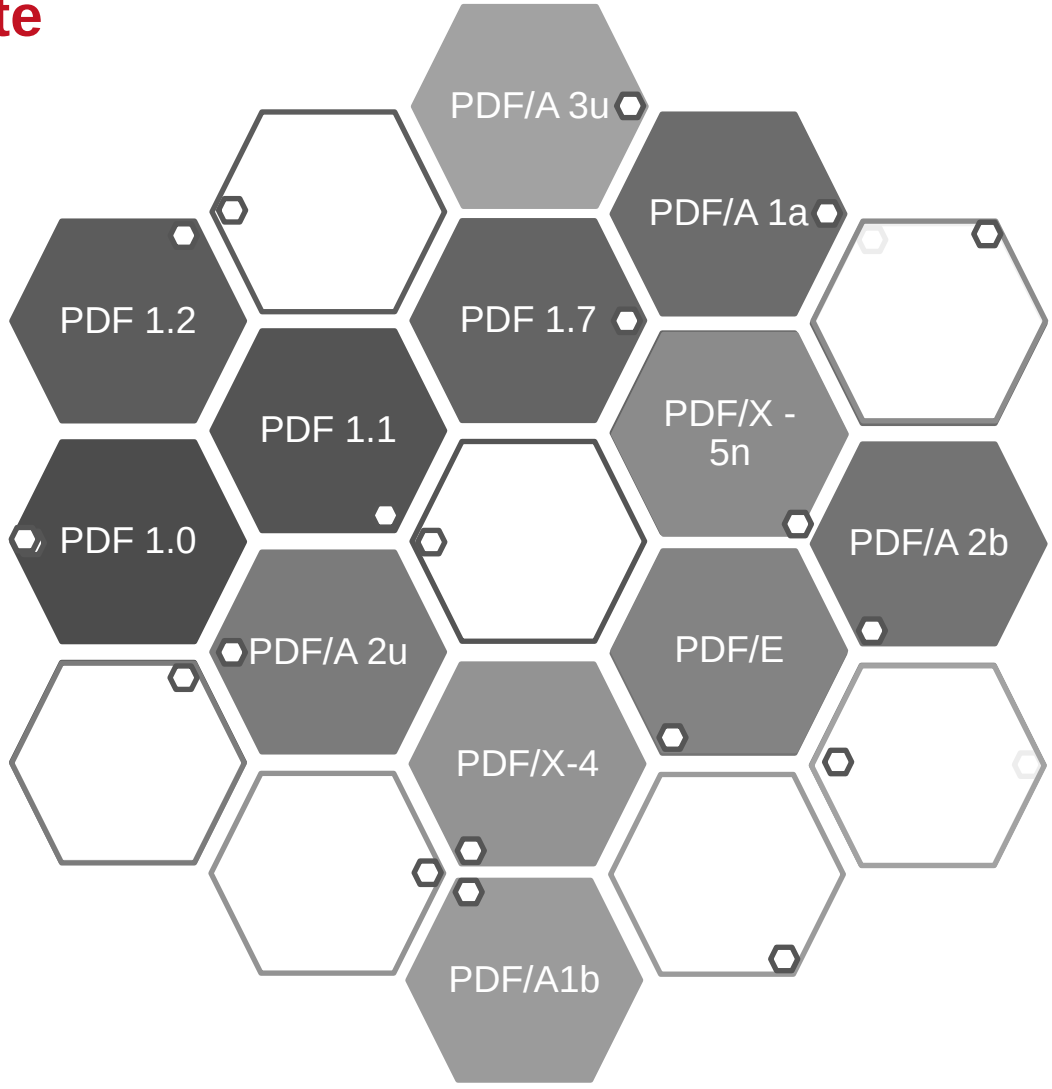
In einer **Formatspezifikation** ist festgelegt, wie die Datei aufgebaut sein soll, damit Programme die Datei richtig darstellen können.

Die Formatspezifikation ist nicht immer öffentlich zugänglich.

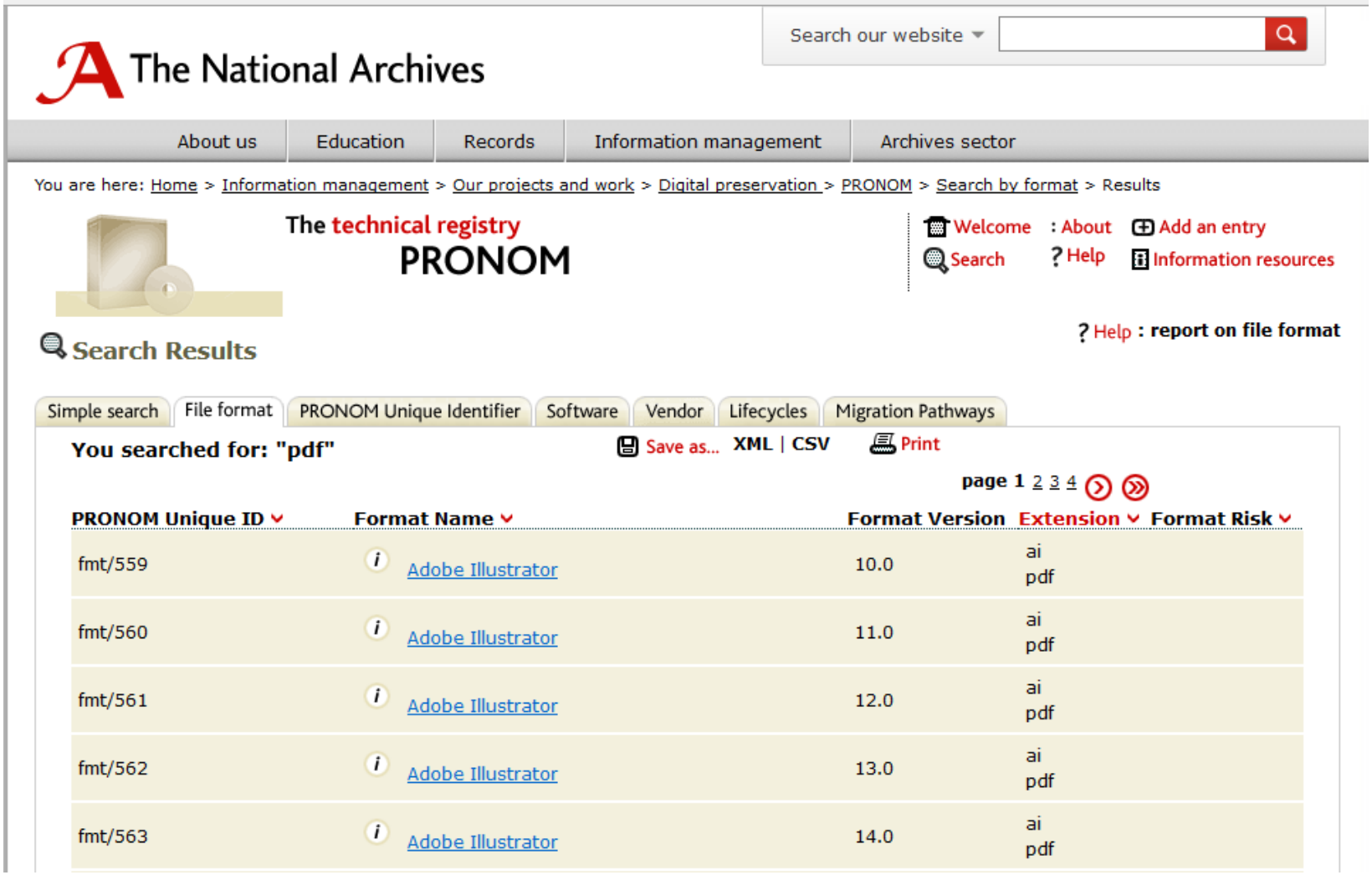
Formate



Formate

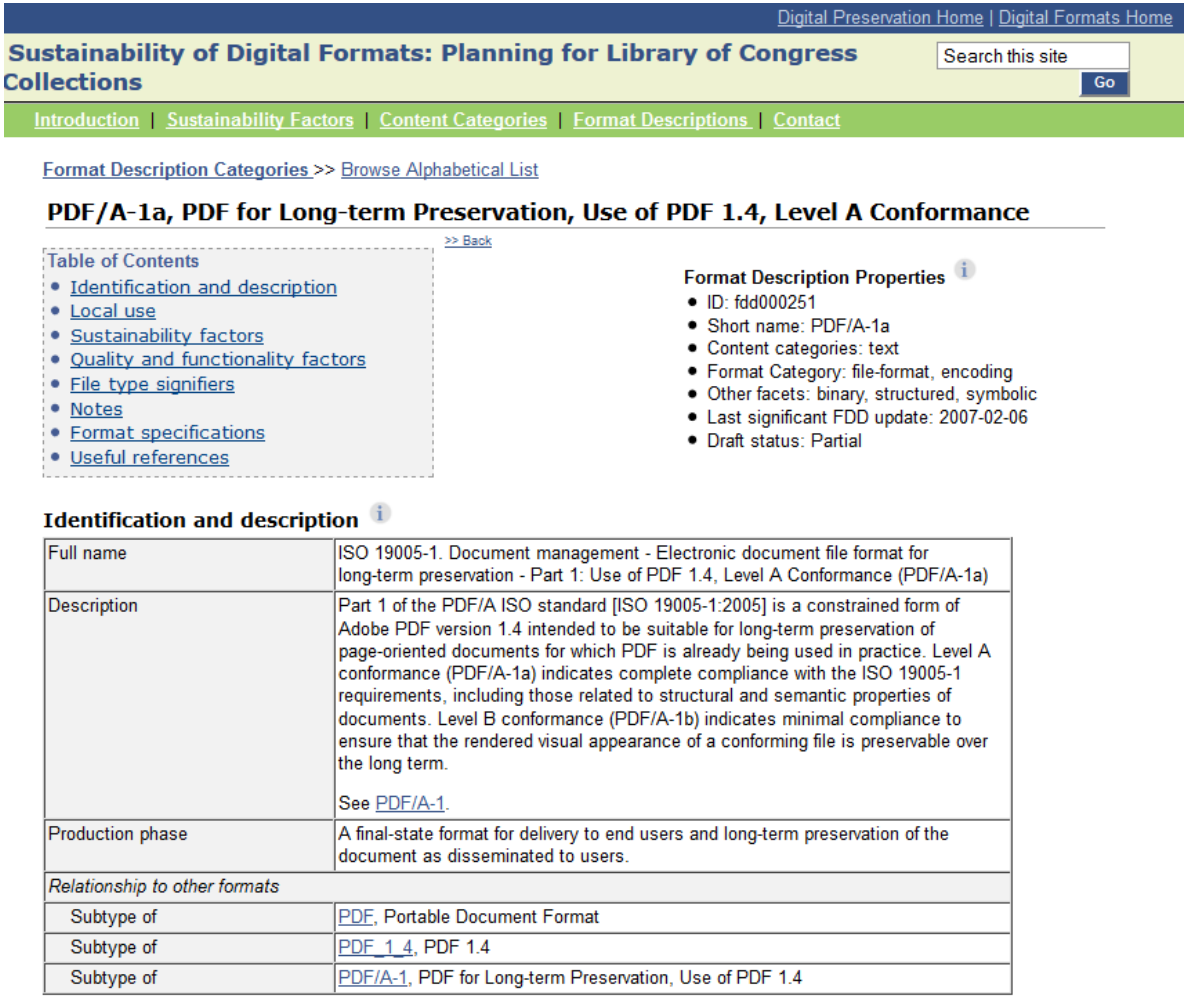

PDF/A 3u
PDF/A 1a
PDF 1.2
PDF 1.7
PDF 1.1
PDF/X - 5n
PDF 1.0
PDF/A 2b
PDF/A 2u
PDF/E
PDF/X-4
PDF/A1b

Formate Tools



http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

Formate - Tools



Digital Preservation Home | Digital Formats Home

Sustainability of Digital Formats: Planning for Library of Congress Collections

Search this site Go

Introduction | Sustainability Factors | Content Categories | Format Descriptions | Contact

Format Description Categories >> Browse Alphabetical List

PDF/A-1a, PDF for Long-term Preservation, Use of PDF 1.4, Level A Conformance

>> Back

Table of Contents
- Identification and description
- Local use
- Sustainability factors
- Quality and functionality factors
- File type signifiers
- Notes
- Format specifications
- Useful references

Format Description Properties
- ID: fdd000251
- Short name: PDF/A-1a
- Content categories: text
- Format Category: file-format, encoding
- Other facets: binary, structured, symbolic
- Last significant FDD update: 2007-02-06
- Draft status: Partial

Identification and description

Full name	ISO 19005-1. Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4, Level A Conformance (PDF/A-1a)
Description	Part 1 of the PDF/A ISO standard [ISO 19005-1:2005] is a constrained form of Adobe PDF version 1.4 intended to be suitable for long-term preservation of page-oriented documents for which PDF is already being used in practice. Level A conformance (PDF/A-1a) indicates complete compliance with the ISO 19005-1 requirements, including those related to structural and semantic properties of documents. Level B conformance (PDF/A-1b) indicates minimal compliance to ensure that the rendered visual appearance of a conforming file is preservable over the long term. See PDF/A-1.
Production phase	A final-state format for delivery to end users and long-term preservation of the document as disseminated to users.
Relationship to other formats	
Subtype of	PDF, Portable Document Format
Subtype of	PDF_1_4, PDF 1.4
Subtype of	PDF/A-1, PDF for Long-term Preservation, Use of PDF 1.4

https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml

Das Wesentliche ist für die Augen unsichtbar.

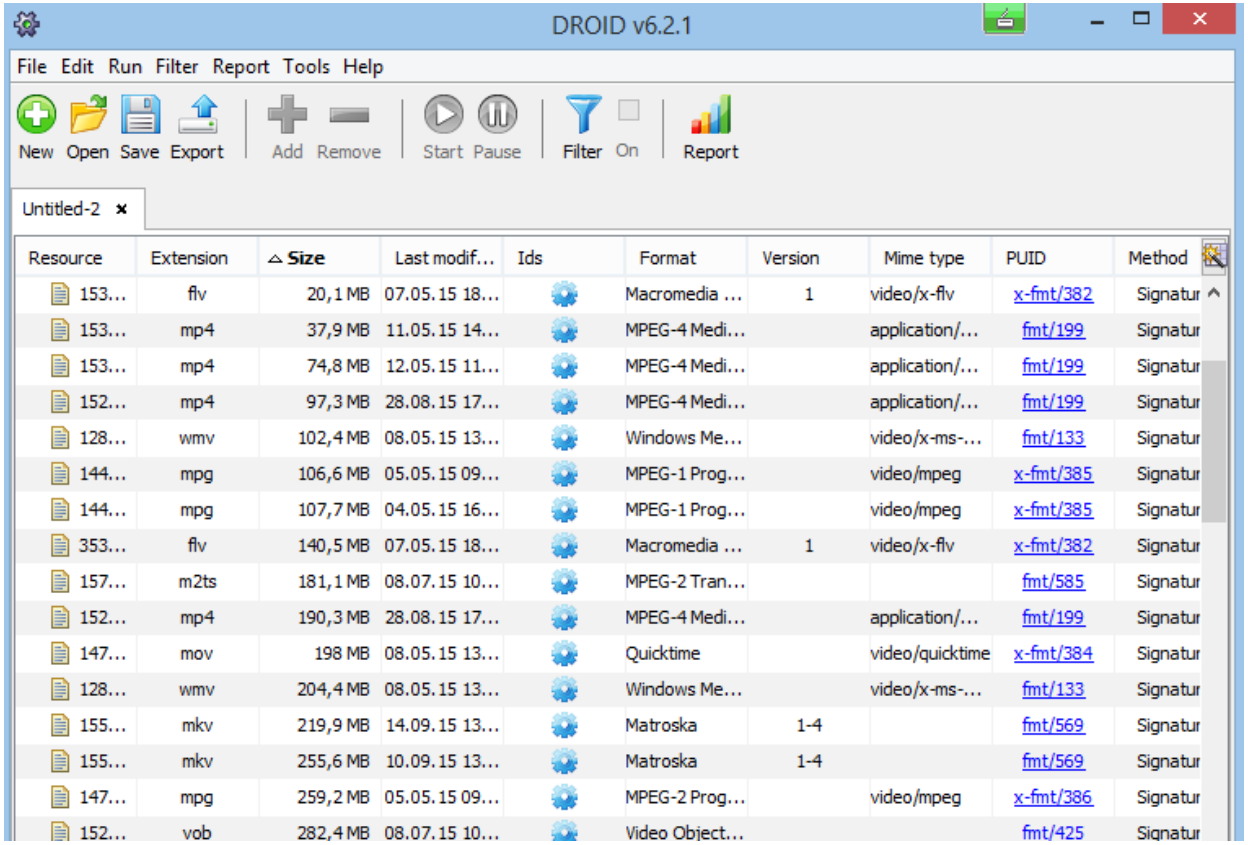
Antoine de Saint-Exupery, Der kleine Prinz, 1942

Formatidentifizierung - Tools

Formatidentifizierungstools
prüfen Formate anhand von Dateiaufbau und/oder Dateiendung

- DROID
- siegfried
- TrID
- FIDO
- …

Formatidentifizierung - Tools



Resource	Extension	Size	Last modif...	Ids	Format	Version	Mime type	PUID	Method
153...	flv	20,1 MB	07.05.15 18...		Macromedia ...	1	video/x-flv	x-fmt/382	Signatur
153...	mp4	37,9 MB	11.05.15 14...		MPEG-4 Medi...		application/...	fmt/199	Signatur
153...	mp4	74,8 MB	12.05.15 11...		MPEG-4 Medi...		application/...	fmt/199	Signatur
152...	mp4	97,3 MB	28.08.15 17...		MPEG-4 Medi...		application/...	fmt/199	Signatur
128...	wmv	102,4 MB	08.05.15 13...		Windows Me...		video/x-ms-...	fmt/133	Signatur
144...	mpg	106,6 MB	05.05.15 09...		MPEG-1 Prog...		video/mpeg	x-fmt/385	Signatur
144...	mpg	107,7 MB	04.05.15 16...		MPEG-1 Prog...		video/mpeg	x-fmt/385	Signatur
353...	flv	140,5 MB	07.05.15 18...		Macromedia ...	1	video/x-flv	x-fmt/382	Signatur
157...	m2ts	181,1 MB	08.07.15 10...		MPEG-2 Tran...			fmt/585	Signatur
152...	mp4	190,3 MB	28.08.15 17...		MPEG-4 Medi...		application/...	fmt/199	Signatur
147...	mov	198 MB	08.05.15 13...		Quicktime		video/quicktime	x-fmt/384	Signatur
128...	wmv	204,4 MB	08.05.15 13...		Windows Me...		video/x-ms-...	fmt/133	Signatur
155...	mkv	219,9 MB	14.09.15 13...		Matroska	1-4		fmt/569	Signatur
155...	mkv	255,6 MB	10.09.15 13...		Matroska	1-4		fmt/569	Signatur
147...	mpg	259,2 MB	05.05.15 09...		MPEG-2 Prog...		video/mpeg	x-fmt/386	Signatur
152...	vob	282,4 MB	08.07.15 10...		Video Object...			fmt/425	Signatur

http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

Anything that can go wrong *will* go wrong.

Murphys Law

Formatvalidierung

Entspricht die Datei der Formatspezifikation?

Wohlgeformt: Aufbau / syntaktisch

Valide: Inhalt / semantisch

Formatvalidierung

Entspricht die Datei der Formatspezifikation?

Wohlgeformt: Aufbau / syntaktisch

Valide: Inhalt / semantisch

```
<Adresse>
        <Straße>  Welfengarten 1b   </Straße>
        <Ort>     Hannover          </Ort>
        <Land>    Deutschland       </Land>
</Adresse>
```

Formatvalidierung - Tools

- Verschiedenen Tools für verschiedene Formate
  - nicht für alle Formate gibt es Tools
- Tools prüfen auf unterschiedlichen Ebenen
- Tools kommen (selten) auf unterschiedliche Ergebnisse

Beispiele für PDF-Validatoren
JHOVE
- prüft die Syntax und die Struktur einer PDF-Datei
- http://openpreservation.org/technology/products/jhove/

veraPDF
- prüft, ob eine PDF-Datei den Anforderungen von PDF/A entspricht
- http://verapdf.org/home/

Die guten ins Töpfchen, die schlechten ...

Jacob und Wilhelm Grimm, Aschenputtel, 1813

Was mache ich, wenn Dateien nicht valide sind?

Wenn es die Möglichkeit gibt, wohlgeformte und valide Dateien zu erhalten, dann sollte man dies ausnutzen!

- neue Version vom Datenproduzenten anfordern
- Software testen
- Dienstleister verpflichten
- Dateien reparieren

Wenn es keine Möglichkeit gibt

- Ins Archiv aufnehmen, Ergebnisse der Formatidentifizierung und Validierung im Archiv dokumentieren
- Validierungstools updaten und Dateien revalidieren
- Preservation Watch: vielleicht gibt es später die Möglichkeit, die Dateien zu reparieren?

Digital Perservation is always an environment of paranoia.

Dave Rice, 2013, dericed.com

Wohlgeformt und valide – Hurra!

- Ins Archiv aufnehmen, Ergebnisse der Formatidentifizierung und Validierung im Archiv dokumentieren
- Validierungstools updaten und Dateien revalidieren
- Preservation Watch: bislang unbekannte Fehler können entdeckt werden
  - Bei Software, die bei der Erstellung genutzt wurde
  - Bei Formaten

Zusammenfassung

Damit wir eine Datei nutzbar erhalten können:

- In welchem Format und in welcher Version liegt die Datei vor?
- Wie ist das Format spezifiziert?
- Entspricht die Datei der Spezifikation?

Mit diesen Informationen kann eine Datei (wieder) dargestellt werden, und auch migriert werden.

Demonstration

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

MEHR INFORMATIONEN

www.tib.eu

Kontaktdaten
Merle Friedrichsen
T 0511 762-14274, merle.friedrichsen@tib.eu

Quellen

PRONOM
http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

Sustainability of Digital Formats
https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml

KOST – Preservation Watch:
http://kost-ceco.ch/cms/index.php?preservation_de

Tools

Formatidentifizierung

- DROID: http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/
- Siegfried: https://github.com/richardlehane/siegfried
- FIDO: http://openpreservation.org/technology/products/fido/

Formatvalidierung

- JHOVE: http://openpreservation.org/technology/products/jhove/
- Checkit_tiff: https://github.com/SLUB-digitalpreservation/checkit_tiff
- KOST-Val: http://kost-ceco.ch/cms/index.php?kost_val_de
- veraPDF: http://verapdf.org/software/
- MediaConch: https://mediaarea.net/MediaConch/download.html

Und viele weitere:
COPTR - Community Owned digital Preservation Tool Registry
http://coptr.digipres.org/Category:Validation