SÄCHSISCHES STAATSARCHIV | Freistaat SACHSEN

Nutzen und Grenzen der Formaterkennung

(Zu-) Fälle bei Pronom und Droid

Stephanie Kortyla, Christian Treu

21. AUdS in Basel 2017

Nutzen und Grenzen der Formaterkennung

1. Einleitung
2. Wofür?
3. Wie?
4. Praxis mit PRONOM & DROID
5. Grenzen von PRONOM & DROID
6. Fazit & Ausblick

Formaterkennung – wofür?

- Analyse
  - Übersicht, Fehler, Filter

  → technische Metadaten
- Bereitstellung von Tools (Darstellung und Veränderung)
- Preservation Planning / Action (OAIS)
- Statistik

- Risikomanagement
- …

Erkennung vs. Validierung

- **Erkennung**
  - Identifizierung von charakteristischen Mustern
  - z. B. DROID, Tika, file, Siegfried
- **Validierung**
  - Norm- bzw. Spezifikationskonformität
  - z.B. KOST-VAL, JHOVE, W3C

Formaterkennung – wie?

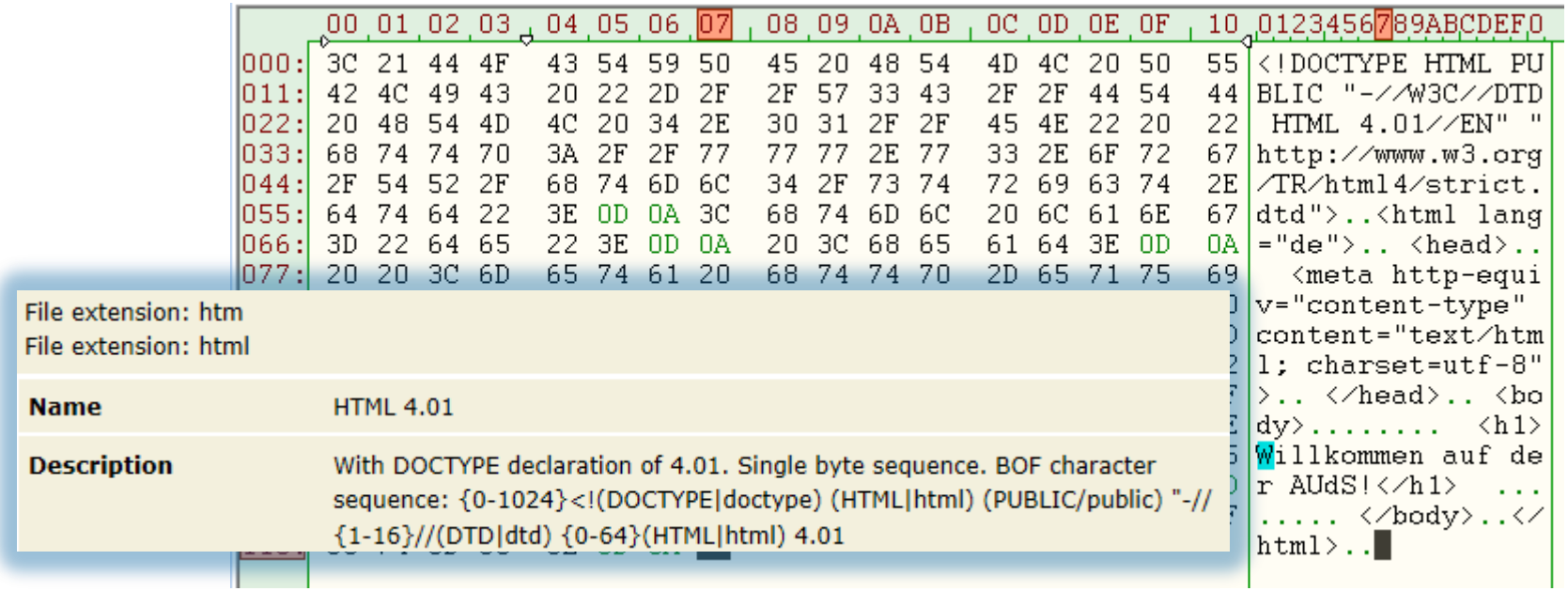
- **Extension / Dateiendung**
  - fehlt, falsch, nicht eindeutig
    - 37 PDF-Varianten → .pdf
    - JPEG → .jpe, .jpeg, .jpg, …
- **MIME-Type**
  - zu ungenau
    - „application/pdf“, „text/html“

Formaterkennung – wie?

- **Signature / Magic Number**
  - spezielles Bitmuster
    - an festen Stellen (z. B. BOF, EOF)
    - ggf. an variablen Stellen
  - Abtasten → hohe Wahrscheinlichkeit

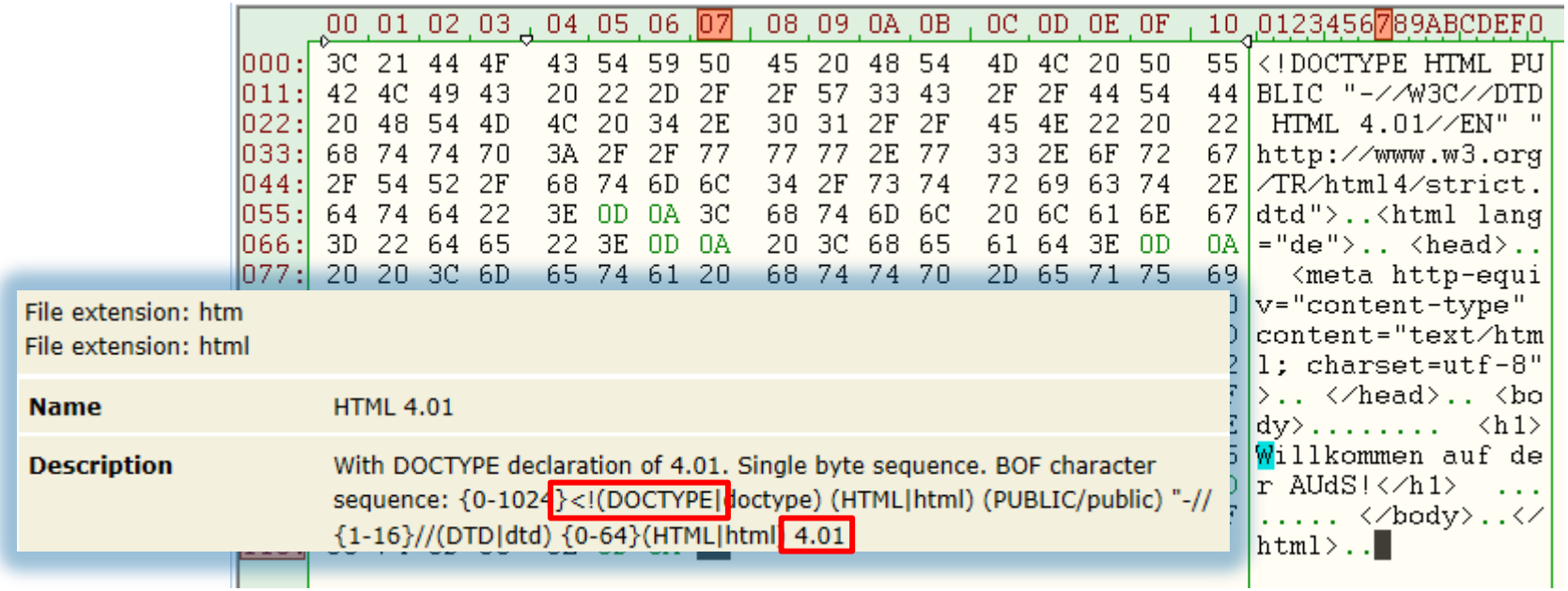
Signaturen am Beispiel

- Formaterkennung basiert auf typischen, wiederkehrenden Mustern, die spezifisch für ein Format sein sollten
- menschenlesbar



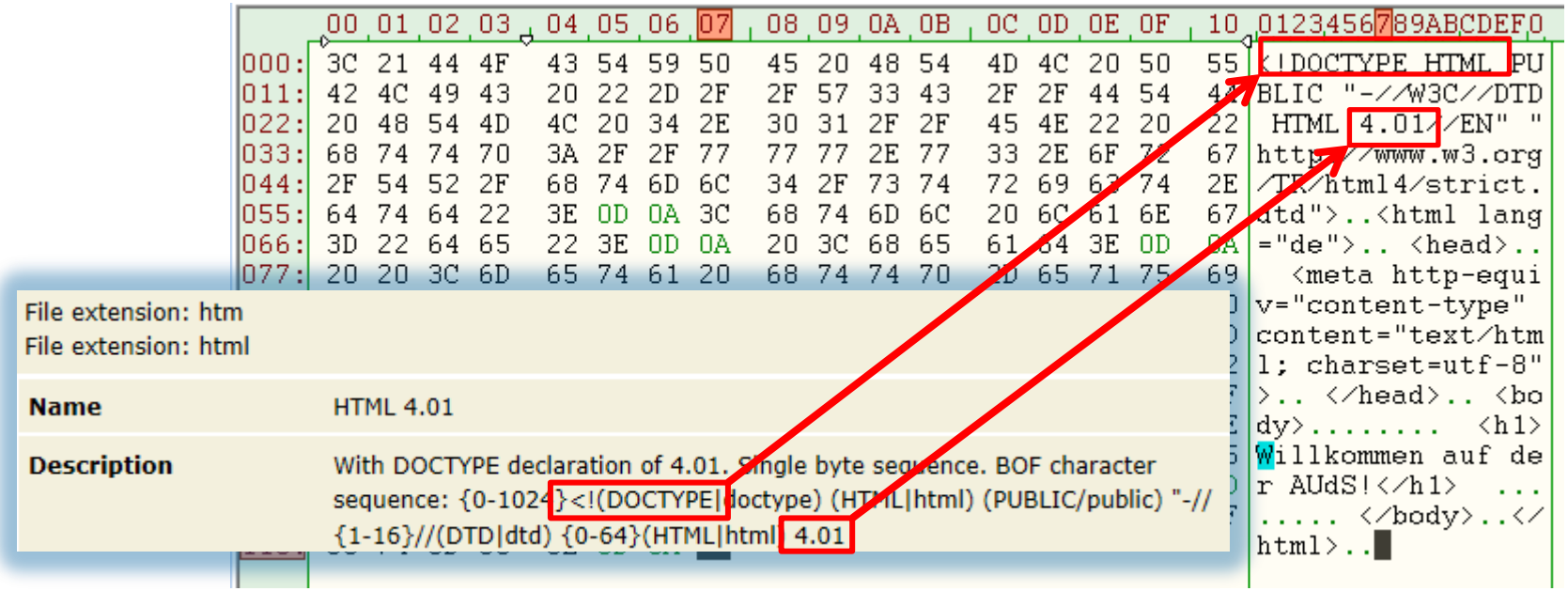
Signaturen am Beispiel

- Formaterkennung basiert auf typischen, wiederkehrenden Mustern, die spezifisch für ein Format sein sollten
- menschenlesbar



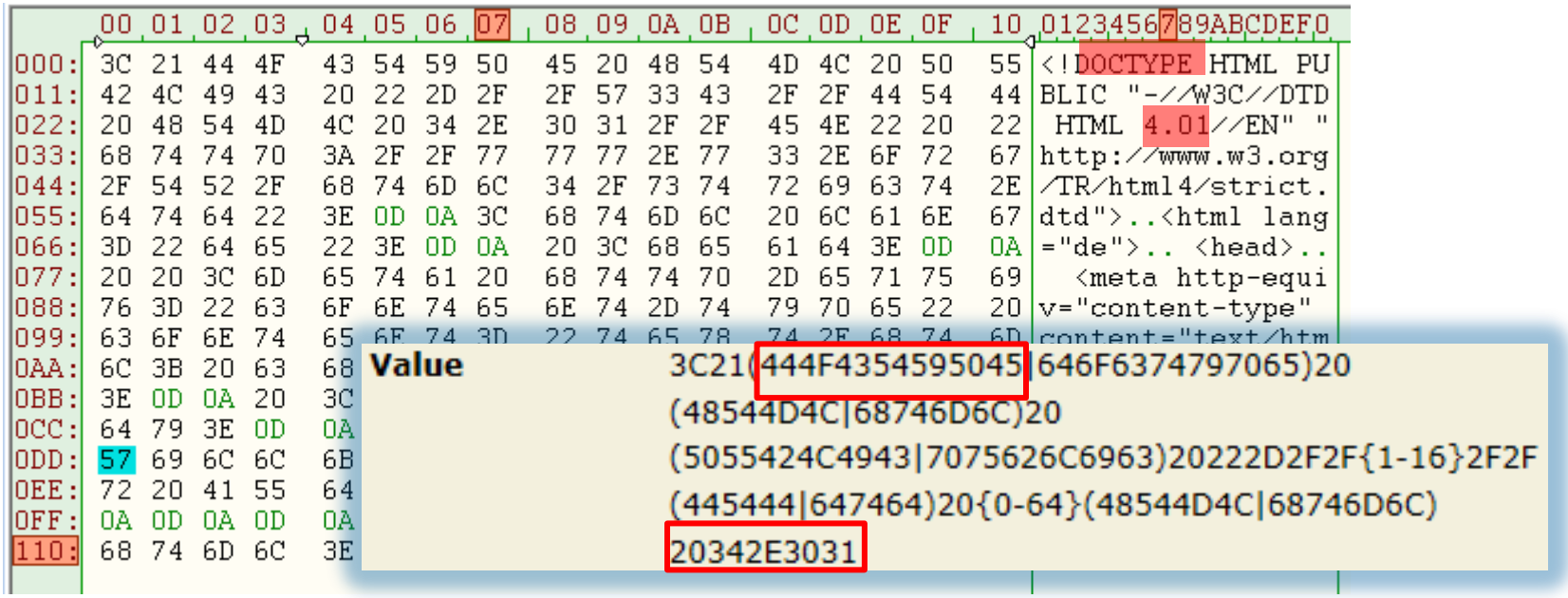
Signaturen am Beispiel

- Formaterkennung basiert auf typischen, wiederkehrenden Mustern, die spezifisch für ein Format sein sollten
- menschenlesbar



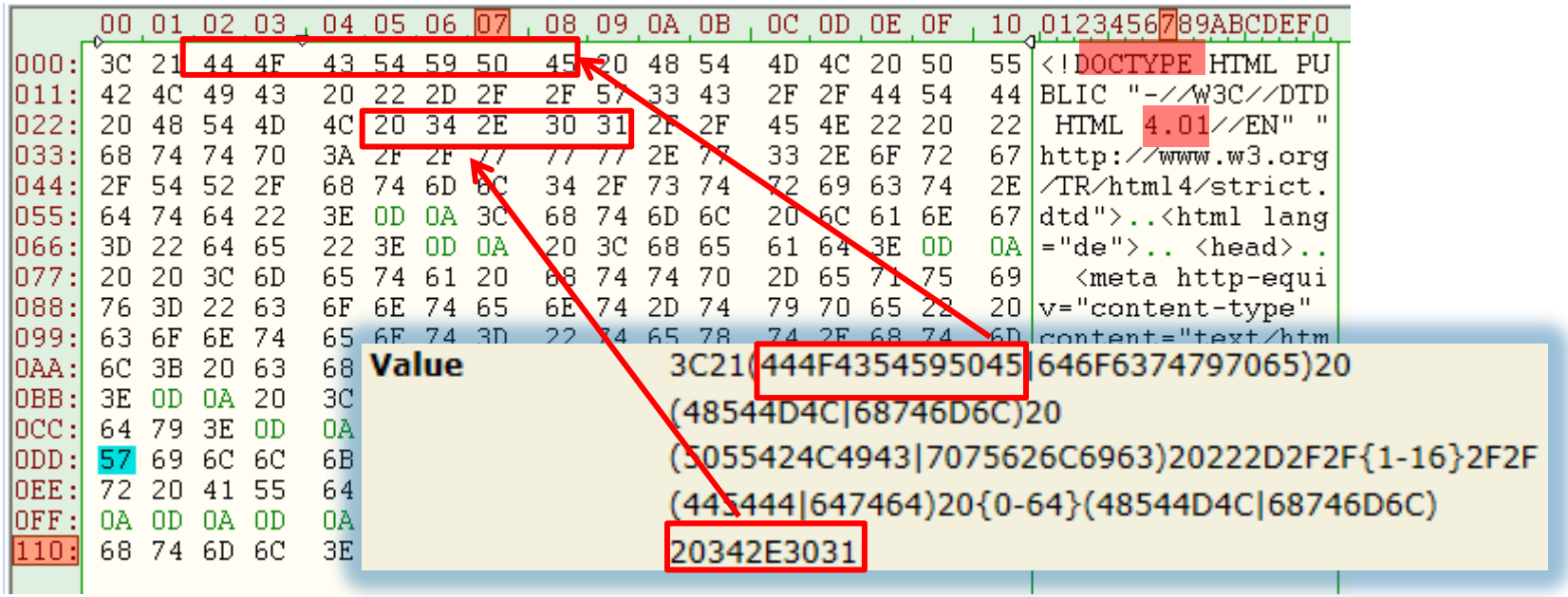
Signaturen am Beispiel

- maschinenlesbar

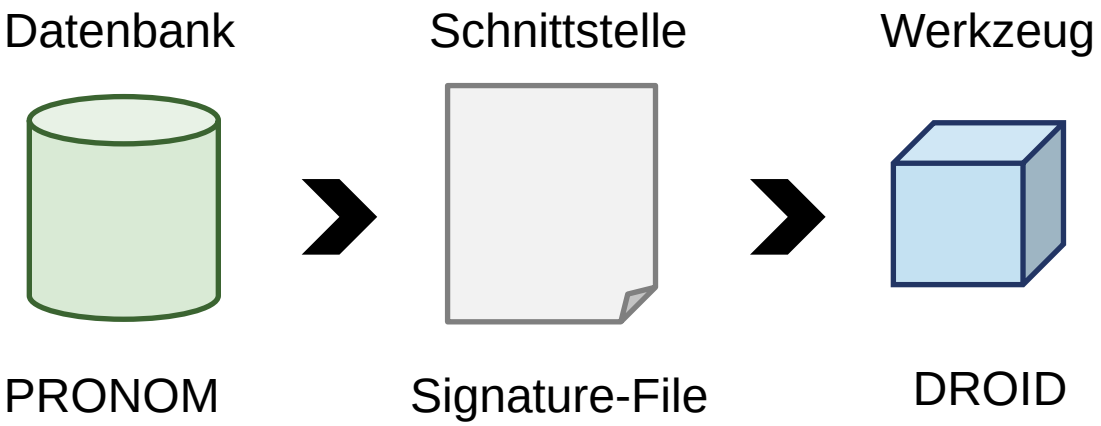


Signaturen am Beispiel

- maschinenlesbar



In der Praxis

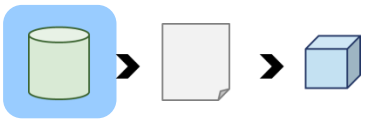


PRONOM

The National Archives

- Start 2002
- Webbasierte Datenbank (kostenfrei)
- eindeutige Identifizierungsmöglichkeit für Formate (PUID)
- Ergänzungs-/ Verbesserungsvorschläge (Community)

PUID Schema:
fmt/18



PRONOM

The National Archives

- Start 2002
- Webbasierte Datenbank (kostenfrei)
- eindeutige Identifizierungsmöglichkeit für Formate (PUID)
- Ergänzungs-/ Verbesserungsvorschläge (Community)

PUID Schema: fmt/18

BIBLIOTHÈQUE CANTONALE ET UNIVERSITAIRE BCU LAUSANNE

TIB LEIBNIZ-INFORMATIONSZENTRUM TECHNIK UND NATURWISSENSCHAFTEN UNIVERSITÄTSBIBLIOTHEK

KUNGL. BIBLIOTEKET National Library of Sweden

Te Puna Mātauranga o Aotearoa NATIONAL LIBRARY OF NEW ZEALAND

Hull History Centre

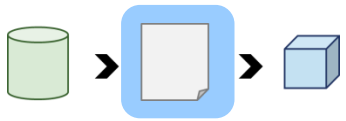
ETH zürich

(DROID) Signature-File

- aus Pronom-Datenbank generiert
  - Relativ regelmäßig, mehrmals im Jahr

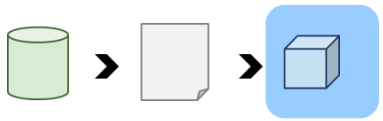
- Aufbau
  - Menschenlesbare Auflistung der Formate („fmt“)
  - technischer Abschnitt zur Erkennung der Signaturen

```
    1   <?xml version="1.0" encoding="UTF-8"?>
    2   <FFSignatureFile DateCreated="2016-09-27T15:37:53" Version="88" xmlns="
        http://www.nationalarchives.gov.uk/pronom/SignatureFile">
    3     <InternalSignatureCollection>
30969     <FileFormatCollection>
38274   </FFSignatureFile>
38275
```

DROID – was und wie

- Digital Record Object Identification
- Entwickler TNA, Start 2005
- 2017: v6.3
- plattformunabhängig, Java-Umgebung
- GUI oder Kommandozeile
- weltweite Verbreitung
- frei, Open Source



DROID – was und wie

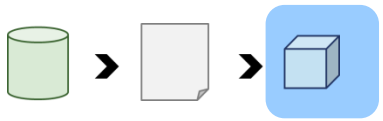
- Datei-Scan, Informationssammlung → Profil → optionaler Export
- Stapelverfahren, über Verzeichnisse hinweg
- Typen: Dateien, Verzeichnisse, „Archiv“-Dateien
- Methoden: Signatures, Extension, (Container)

Resource	Extension	Ids	Method	PUID	Format	Version	Mime type	Hash
C:\Users\tksuser\Desktop\A								
B								
Test1.txt	txt		Extension	x-fmt/111	Plain Text File		text/plain	30aba6f8
Test2txt								30aba6f8
C:\Users\tksuser\Desktop\Test3.xls	xls		Container	fmt/61	Microsoft Excel 97 Workbook (xls)	8	application/vnd.ms-excel	c768bdbb
C:\Users\tksuser\Desktop\Test4.csv	csv		Extension	x-fmt/18	Comma Separated Values		text/csv	34ee5f50
C:\Users\tksuser\Desktop\Test5.xls	xls		Container	fmt/61	Microsoft Excel 97 Workbook (xls)	8	application/vnd.ms-excel	3af968ab
C:\Users\tksuser\Desktop\Test6.jpg	jpg		Signature	fmt/645	Exchangeable Image File Format (Compressed)	2.2.1	image/jpeg	8af1a77(
C:\Users\tksuser\Desktop\Test7.lck	lck	(2)	Extension	"fmt/335", "fmt/218"	"Dreamweaver Lock File", "Microsoft FrontPage"			7b60bf45

Grenzen von DROID & PRONOM

DROID – Signature sticht

1. Überprüfung auf Signatures
2. Überprüfung auf Extensions

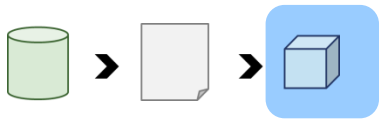


DROID – Signature sticht

1. Überprüfung auf Signatures
2. Überprüfung auf Extensions

- Echtdaten in Plain Text / EBCDIC

→ erwartet: fmt/159 (EBCDIC-US)



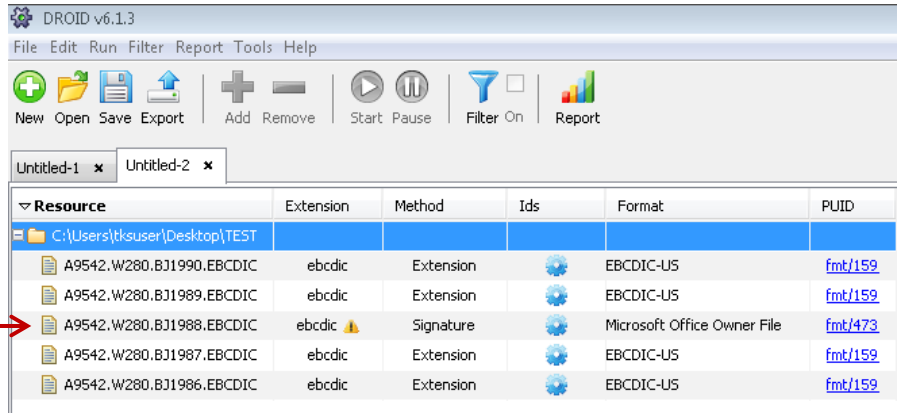
DROID – Signature sticht

1. Überprüfung auf Signatures
2. Überprüfung auf Extensions

- Echtdaten in Plain Text / EBCDIC

→ erwartet: fmt/159 (EBCDIC-US)

→ Ergebnis: fmt/473 (MS Owner File)



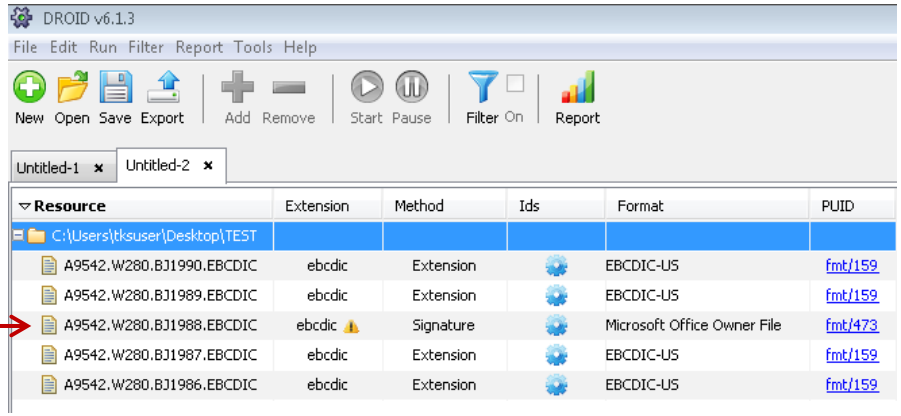
DROID – Signature sticht

1. Überprüfung auf Signatures
2. Überprüfung auf Extensions

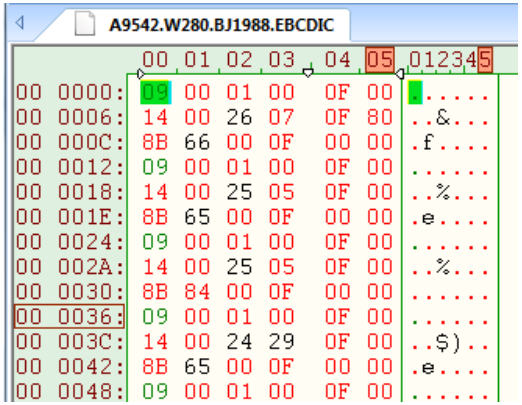
- Echtdaten in Plain Text / EBCDIC

→ erwartet: fmt/159 (EBCDIC-US)

→ Ergebnis: fmt/473 (MS Owner File)



Name	MS Owner file v4	
Description	1st byte is within the range of 06 to 0F. Next bytes are name of registered user. 55th byte is identical to 1st. 56th byte is 00	
Byte sequences	**Position type**	Absolute from BOF
	Offset	0
	Byte order	
	Value	09{52}000900



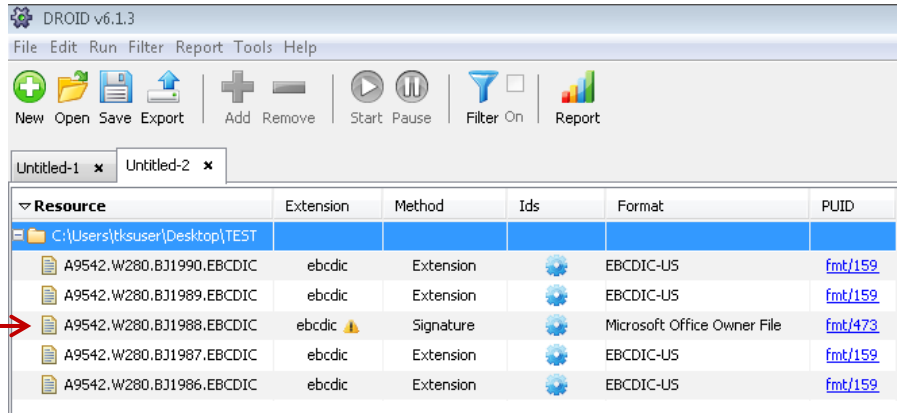
DROID – Signature sticht

1. Überprüfung auf Signatures
2. Überprüfung auf Extensions

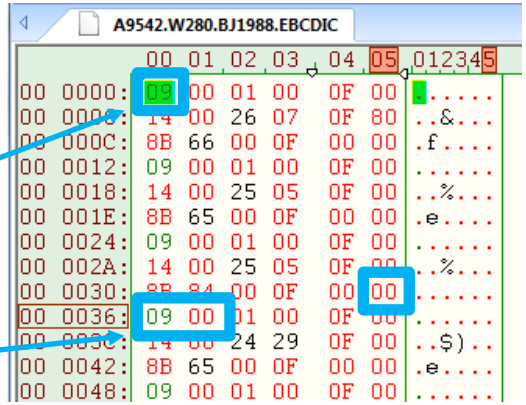
- Echtdaten in Plain Text / EBCDIC

→ erwartet: fmt/159 (EBCDIC-US)

→ Ergebnis: fmt/473 (MS Owner File)

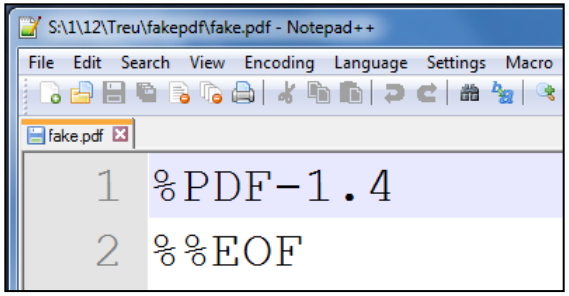


Name	MS Owner file v4	
Description	1st byte is within the range of 06 to 0F. Next bytes are name of registered user. 55th byte is identical to 1st. 56th byte is 00	
Byte sequences	**Position type**	Absolute from BOF
	Offset	0
	Byte order	
	Value	09{52}000900



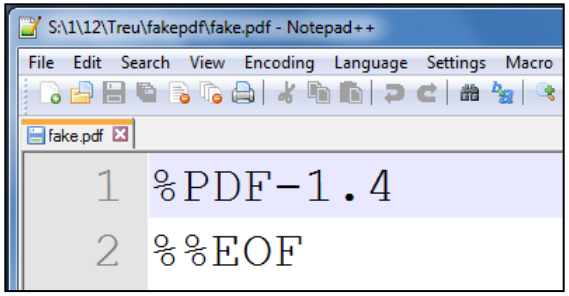
DROID auf dem Glatteis

- **nicht** funktionsfähige Datei



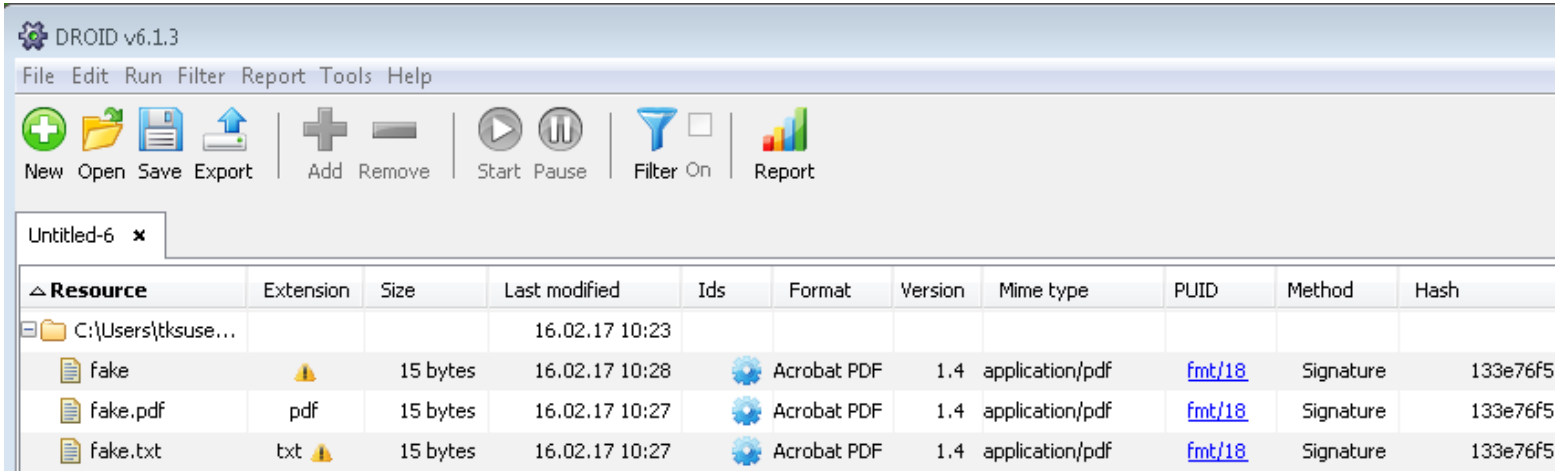
DROID auf dem Glatteis

- **nicht** funktionsfähige Datei



→ DROID lässt sich einfach täuschen

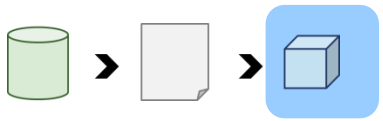
→arbeitet aber erwartungsgemäß / zuverlässig



Ergebnisdifferenzen

Echtdaten

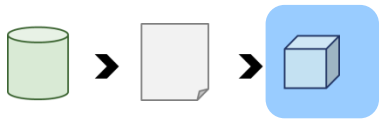
Datei	**DROID v6.1.3** (SignFile V79, MBS)	**DROID v6.3** (SignFile V88, full scan)	**Siegfried (1.6.7)**	**Formatname**
A1.jpg	fmt/41	fmt/645	fmt/645	Raw JPEG Stream ←→ EXIF 2.2.1
B1.pdf	fmt/18	fmt/95	fmt/95	PDF 1.4 ←→ PDF/A-1a
B2.pdf	fmt/276	fmt/478	fmt/478	PDF 1.7 ←→ PDF/A-2u



Ergebnisdifferenzen

Echtdaten

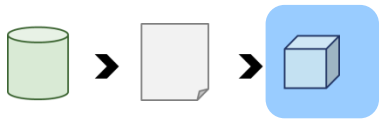
Datei	**DROID v6.1.3** (SignFile V79, MBS)	**DROID v6.3** (SignFile V88, full scan)	**Siegfried (1.6.7)**	**Formatname**
A1.jpg	fmt/41	fmt/645	fmt/645	Raw JPEG Stream ←→ EXIF 2.2.1
B1.pdf	fmt/18	fmt/95	fmt/95	PDF 1.4 ←→ PDF/A-1a
B2.pdf	fmt/276	fmt/478	fmt/478	PDF 1.7 ←→ PDF/A-2u



DROID – Scan-Modi

BOF	„digital object“	EOF

Nach Jay Gattuso (NZNL)

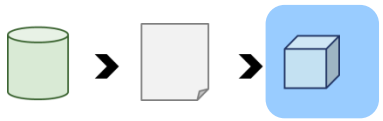


DROID – Scan-Modi

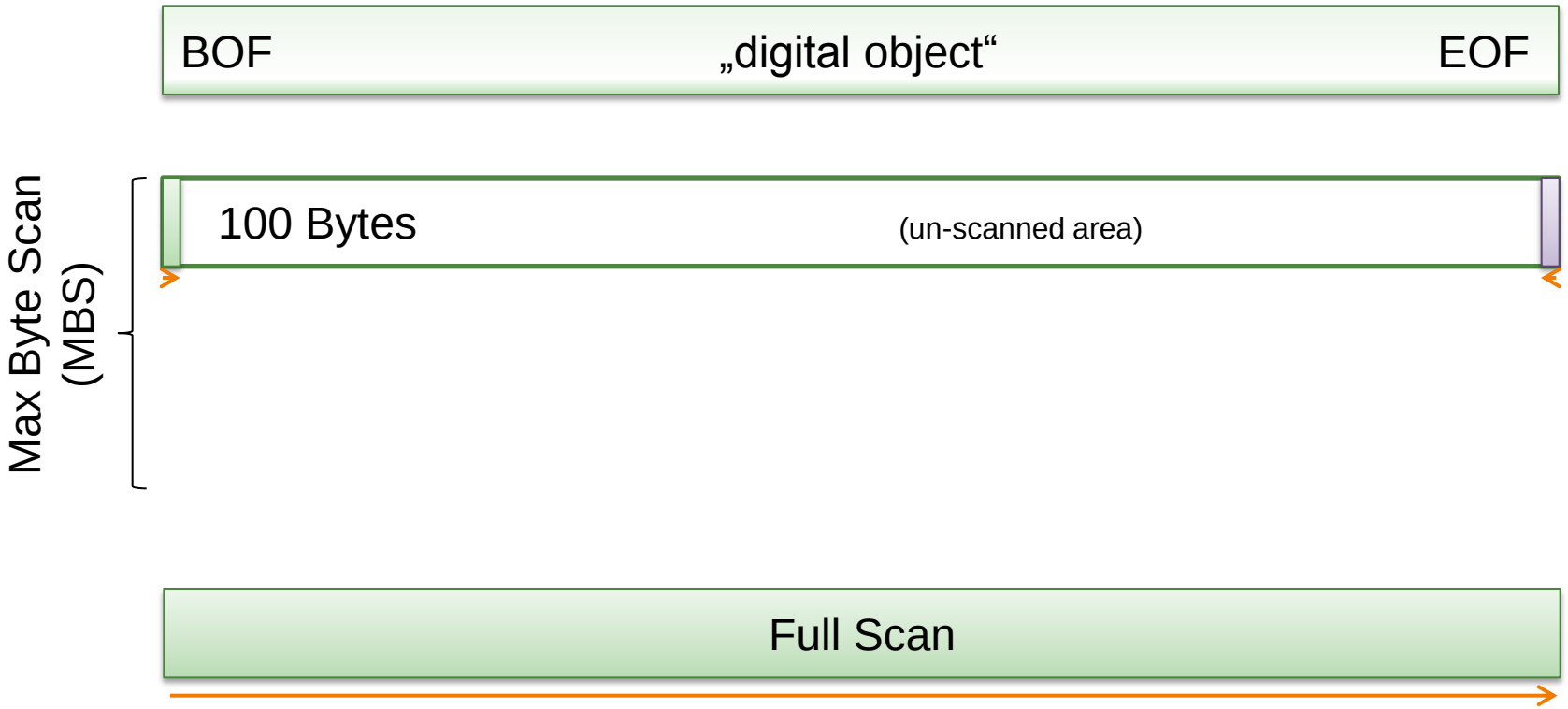
BOF „digital object“ EOF

Full Scan

Nach Jay Gattuso (NZNL)

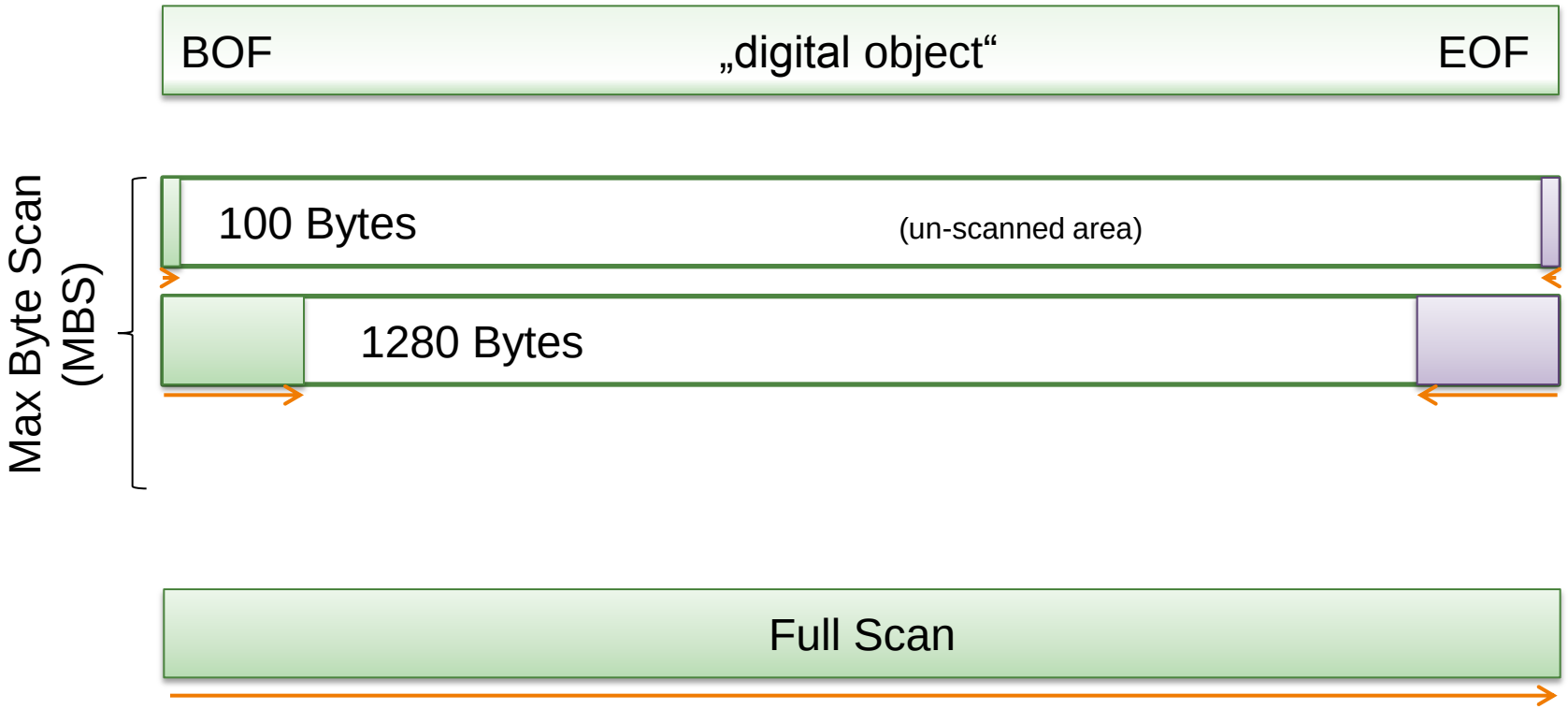


DROID – Scan-Modi



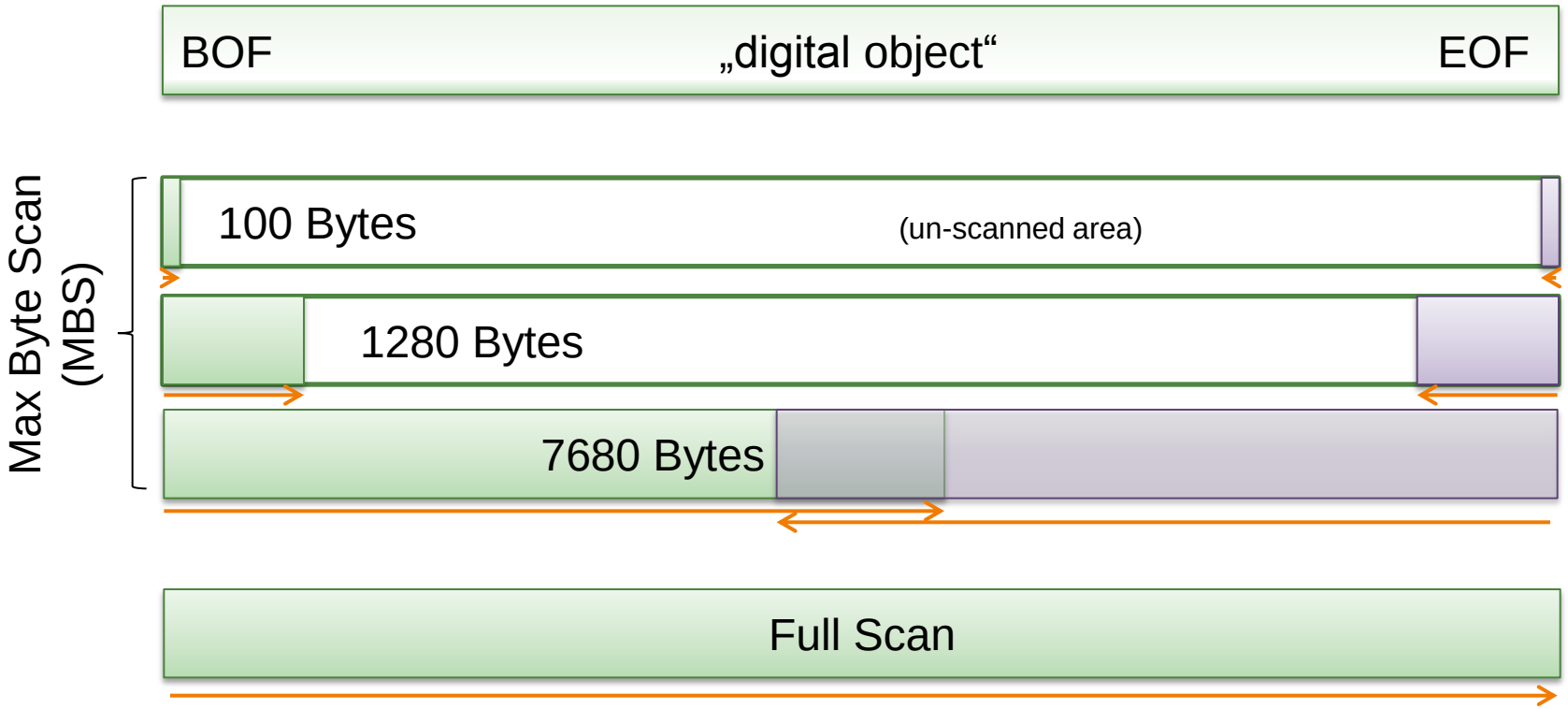
Nach Jay Gattuso (NZNL)

DROID – Scan-Modi



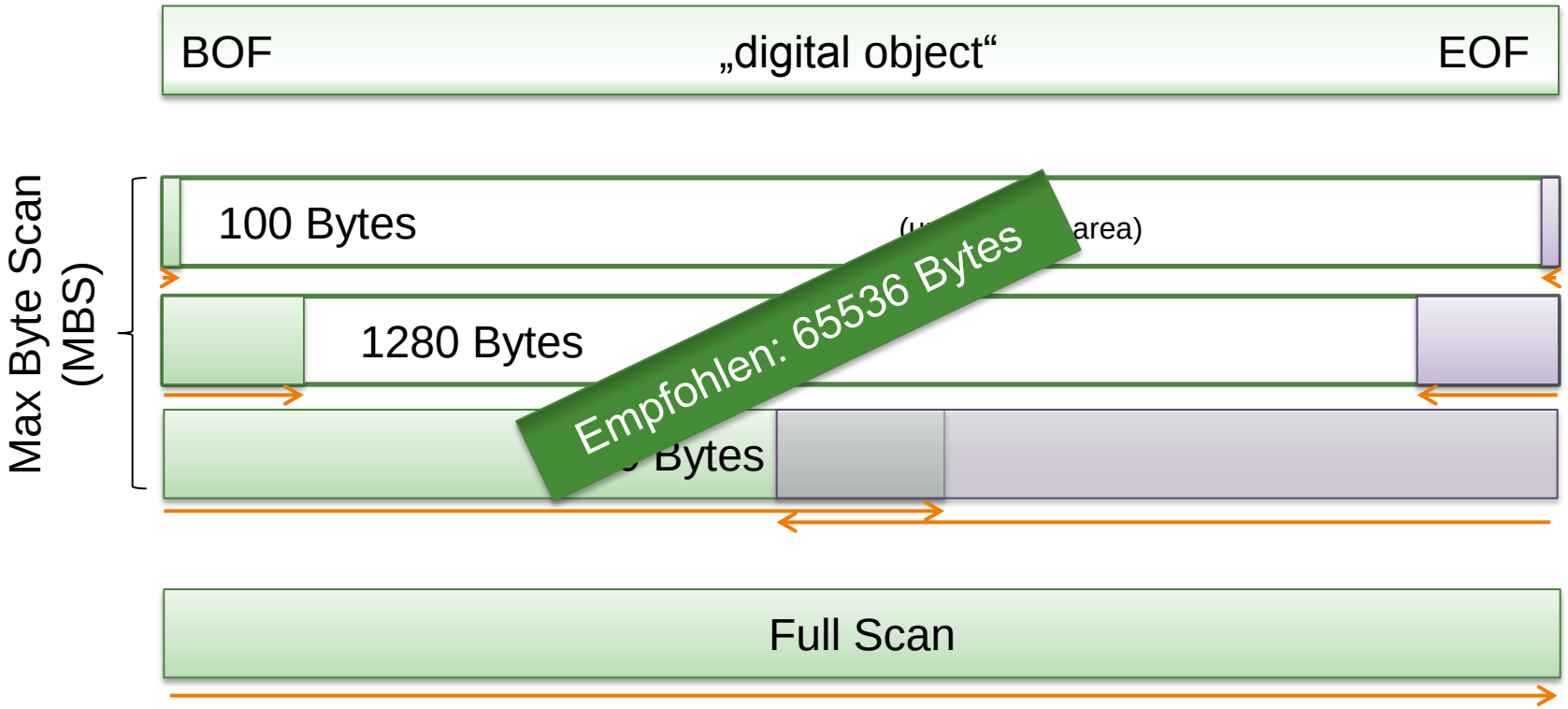
Nach Jay Gattuso (NZNL)

DROID – Scan-Modi



Nach Jay Gattuso (NZNL)

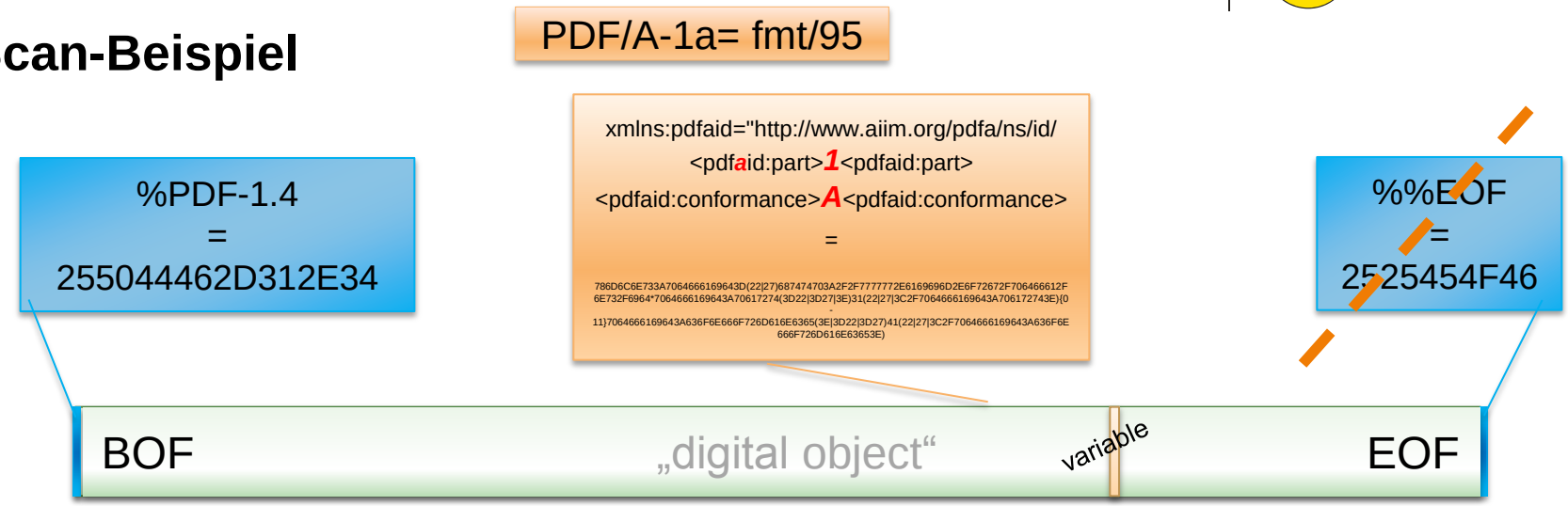
DROID – Scan-Modi



Nach Jay Gattuso (NZNL)

Scan-Beispiel

Scan-Beispiel



Scan-Beispiel

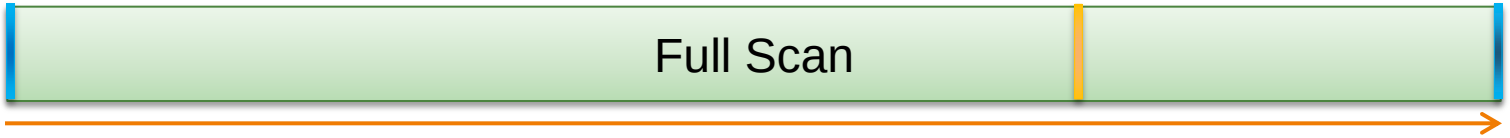
PDF/A-1a= fmt/95

%PDF-1.4
=
255044462D312E34

xmlns:pdfaid="http://www.aiim.org/pdfa/ns/id/
<pdfaid:part>1<pdfaid:part>
<pdfaid:conformance>A<pdfaid:conformance>
=

786D6C6E733A70646666169643D(22|27)687474703A2F2F7777772E6169696D2E6F72672F706466612F6E732F6964*70646666169643A70617274(3D22|3D27|3E)31(22|27|3C2F70646666169643A706172743E){0,11}70646666169643A636F6E666F726D616E6365(3E|3D22|3D27)41(22|27|3C2F70646666169643A636F6E666F726D616E63653E)

%%EOF
=
2525454F46



Scan-Beispiel

PDF/A-1a= fmt/95

%PDF-1.4
=
255044462D312E34

xmlns:pdfaid="http://www.aiim.org/pdfa/ns/id/
<pdfaid:part>1<pdfaid:part>
<pdfaid:conformance>A<pdfaid:conformance>
=

786D6C6E733A70646666169643D(22|27)687474703A2F2F7777772E6169696D2E6F72672F706466612F6E732F6964*70646666169643A70617274(3D22|3D27|3E)31(22|27|3C2F70646666169643A706172743E){0,11}70646666169643A636F6E666F726D616E6365(3E|3D22|3D27)41(22|27|3C2F70646666169643A636F6E666F726D616E63653E)

%%EOF
=
2525454F46

BOF „digital object“ variable EOF

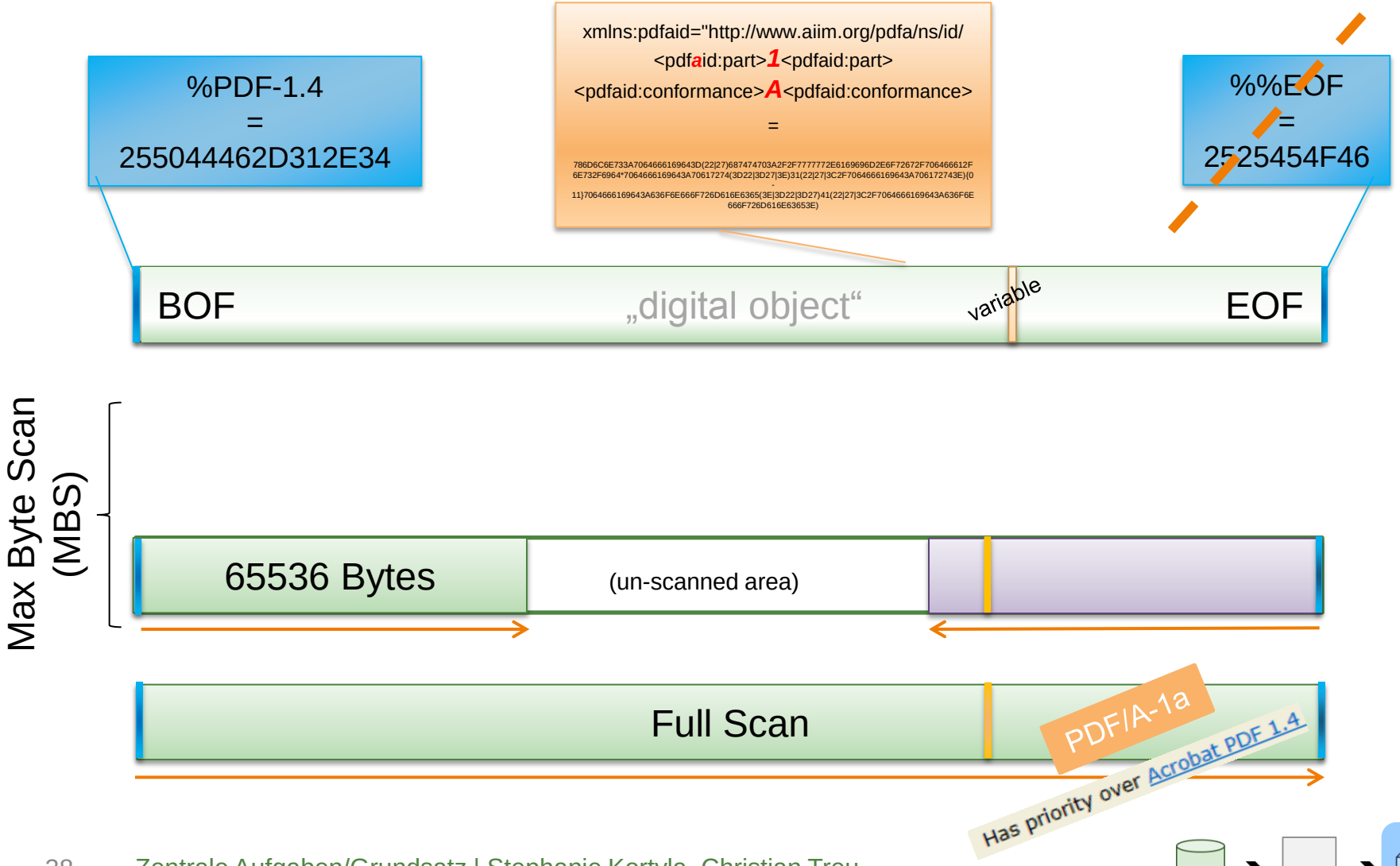
Full Scan

PDF/A-1a

Has priority over Acrobat PDF 1.4

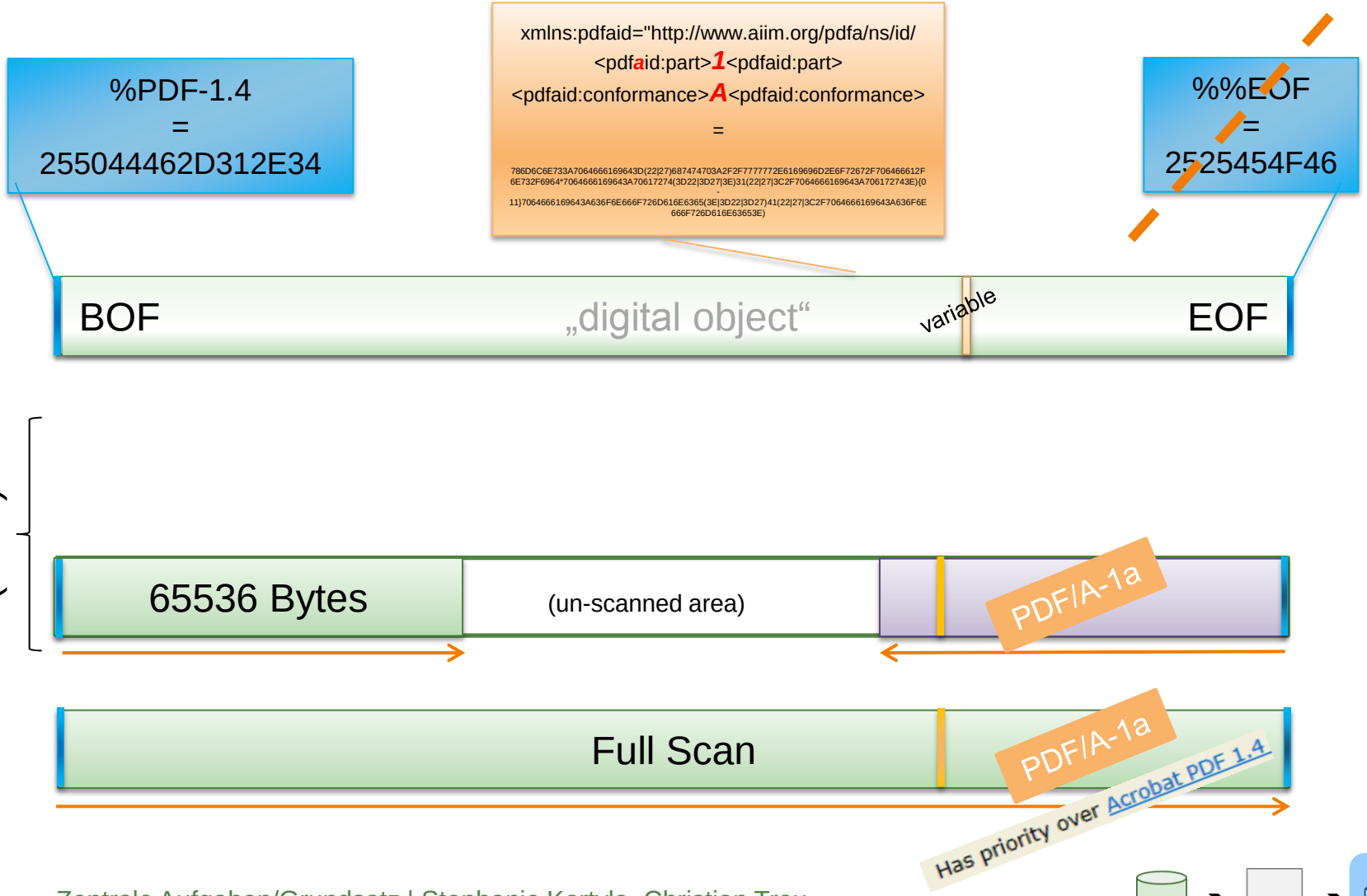
Scan-Beispiel

PDF/A-1a= fmt/95



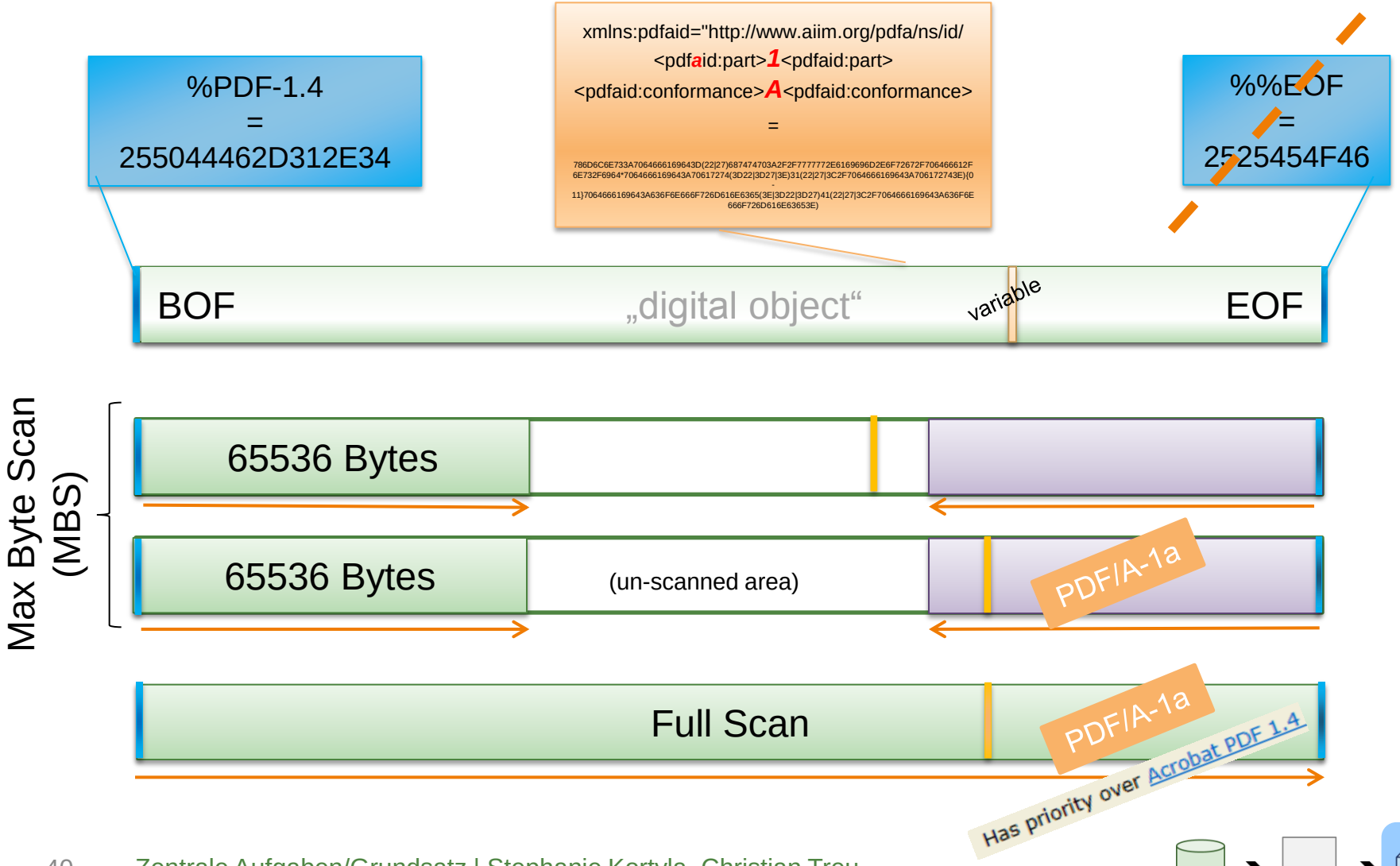
Scan-Beispiel

PDF/A-1a= fmt/95



Scan-Beispiel

PDF/A-1a= fmt/95



Scan-Beispiel

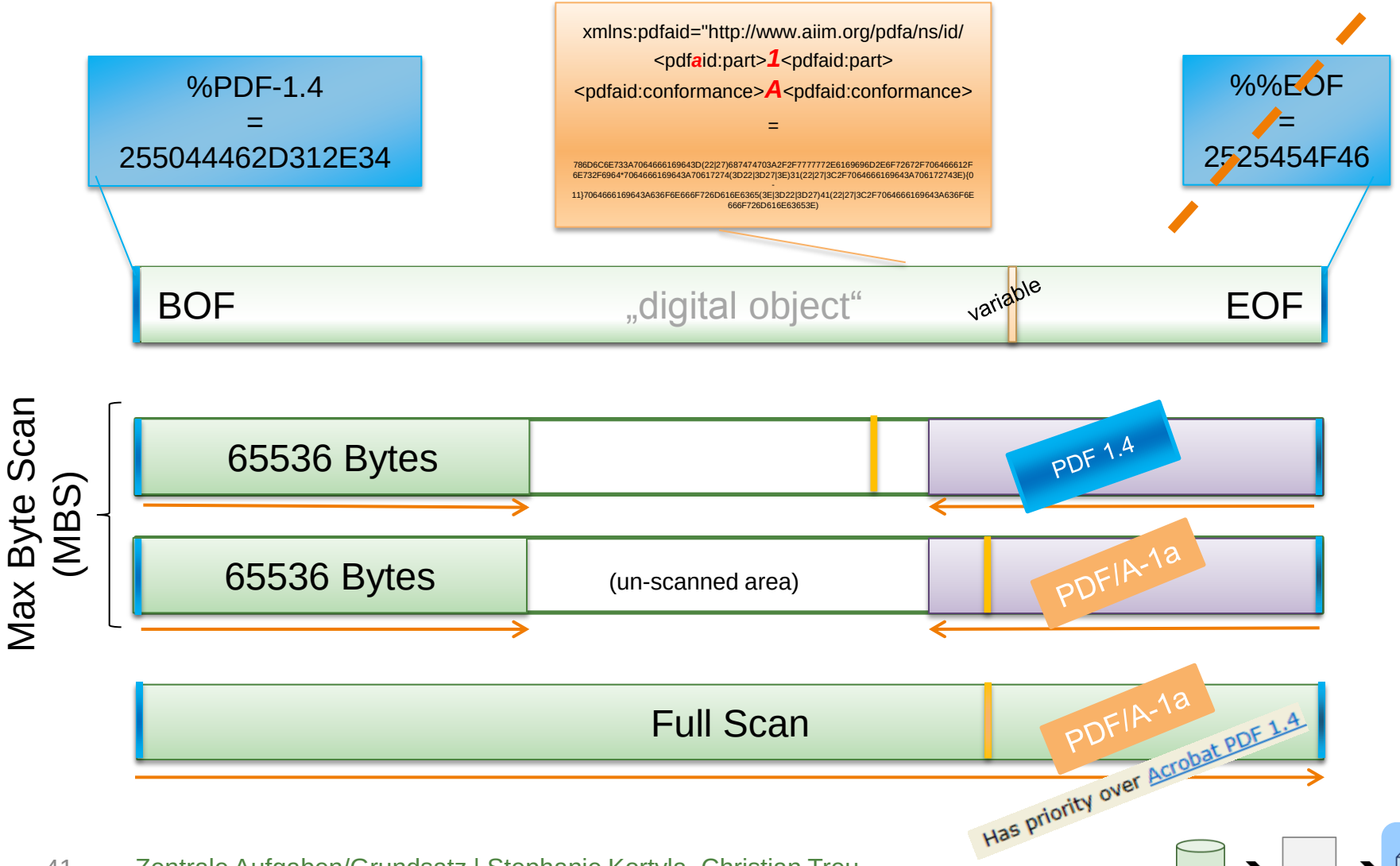
PDF/A-1a= fmt/95

%PDF-1.4
=
255044462D312E34

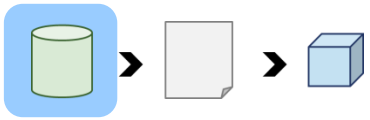
xmlns:pdfaid="http://www.aiim.org/pdfa/ns/id/
<pdfaid:part>1<pdfaid:part>
<pdfaid:conformance>A<pdfaid:conformance>
=

786D6C6E733A70646666169643D(22|27)687474703A2F2F7777772E6169696D2E6F72672F706466612F6E732F6964*70646666169643A70617274(3D22|3D27|3E)31(22|27|3C2F70646666169643A706172743E){0,11}70646666169643A636F6E666F726D616E6365(3E|3D22|3D27)41(22|27|3C2F70646666169643A636F6E666F726D616E63653E)

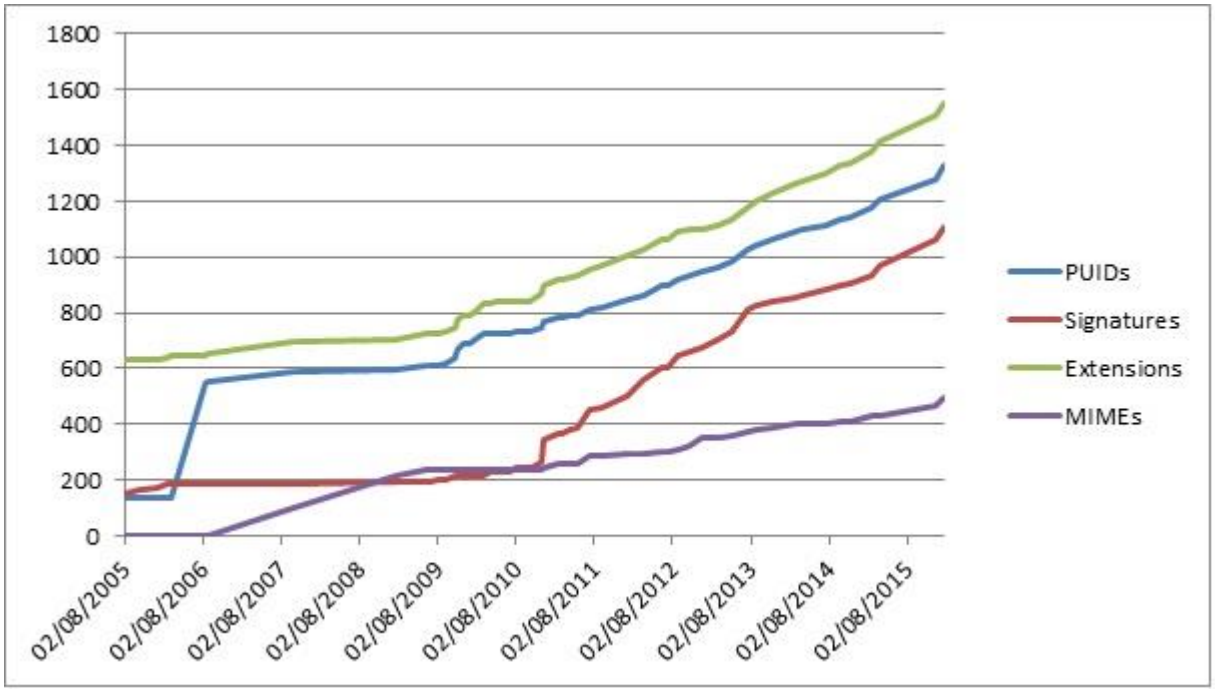
%%EOF
=
2525454F46



Weiterentwicklung der Datenbank



PRONOM-Entwicklung (2005 – 2016)

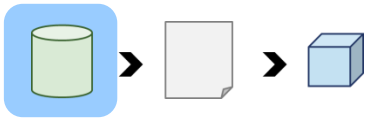


Quelle: Paul Young (http://blog.nationalarchives.gov.uk/blog/identifying-digital-file-formats-collaborative-effort/)

Weiterentwicklung von PRONOM

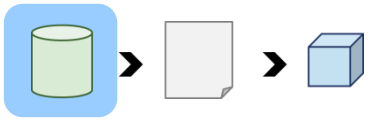
- Neue Einträge
- Vorläufige Einträge
- Zusammenlegung von Einträgen
- Überholte Einträge - „Deprecated“

- → Persistenz?!



Ehemals vorläufige Einträge

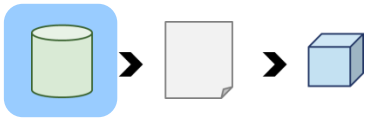
Name	Acrobat PDF 1.4	Comma Separated Values
Version	1.4	
Other names	PDF (1.4)	
Identifiers	MIME: application/pdf PUID: fmt/18	MIME: text/csv PUID: x-fmt/18



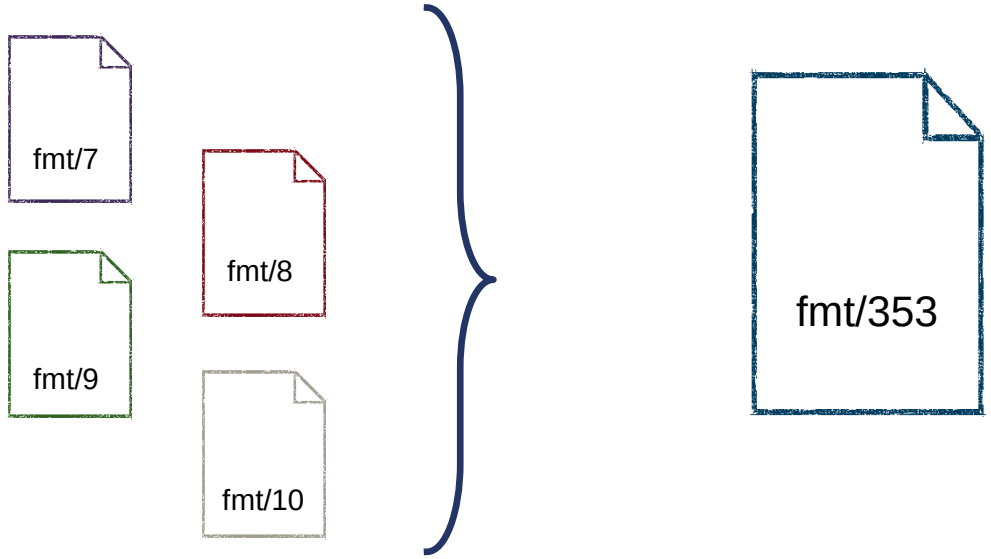
Ehemals vorläufige Einträge

Name	Acrobat PDF 1.4	Comma Separated Values
Version	1.4	
Other names	PDF (1.4)	
Identifiers	MIME: application/pdf PUID: fmt/18	MIME: text/csv PUID: x-fmt/18

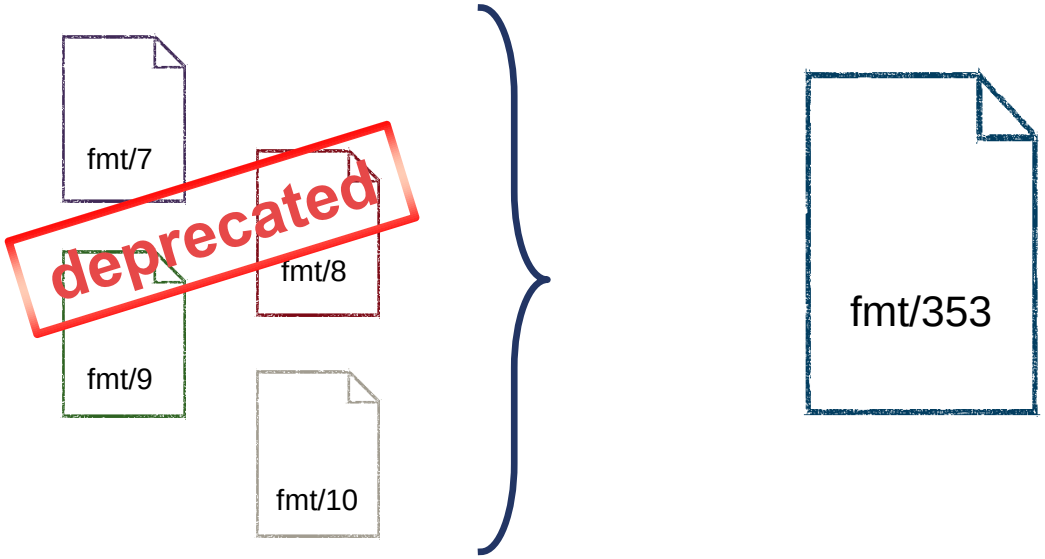
- fmt ≠ x-fmt
- Insgesamt: ca. 450 „doppelt“ vergebene Signaturen



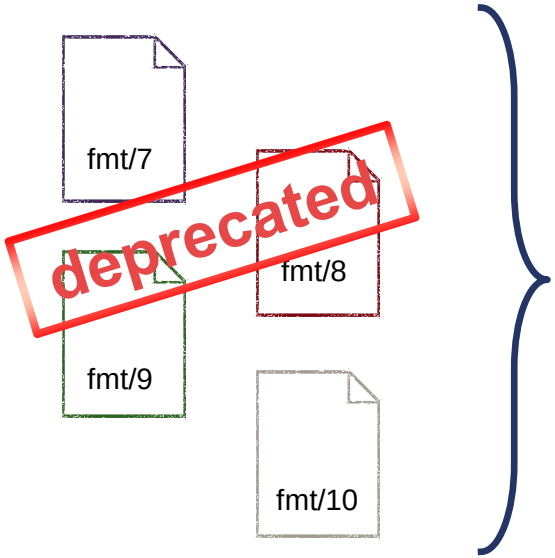
Zusammenlegung von Einträgen Beispiel TIFF



Zusammenlegung von Einträgen Beispiel TIFF

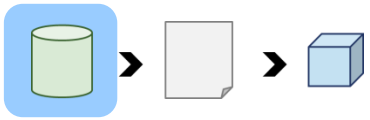


Zusammenlegung von Einträgen Beispiel TIFF

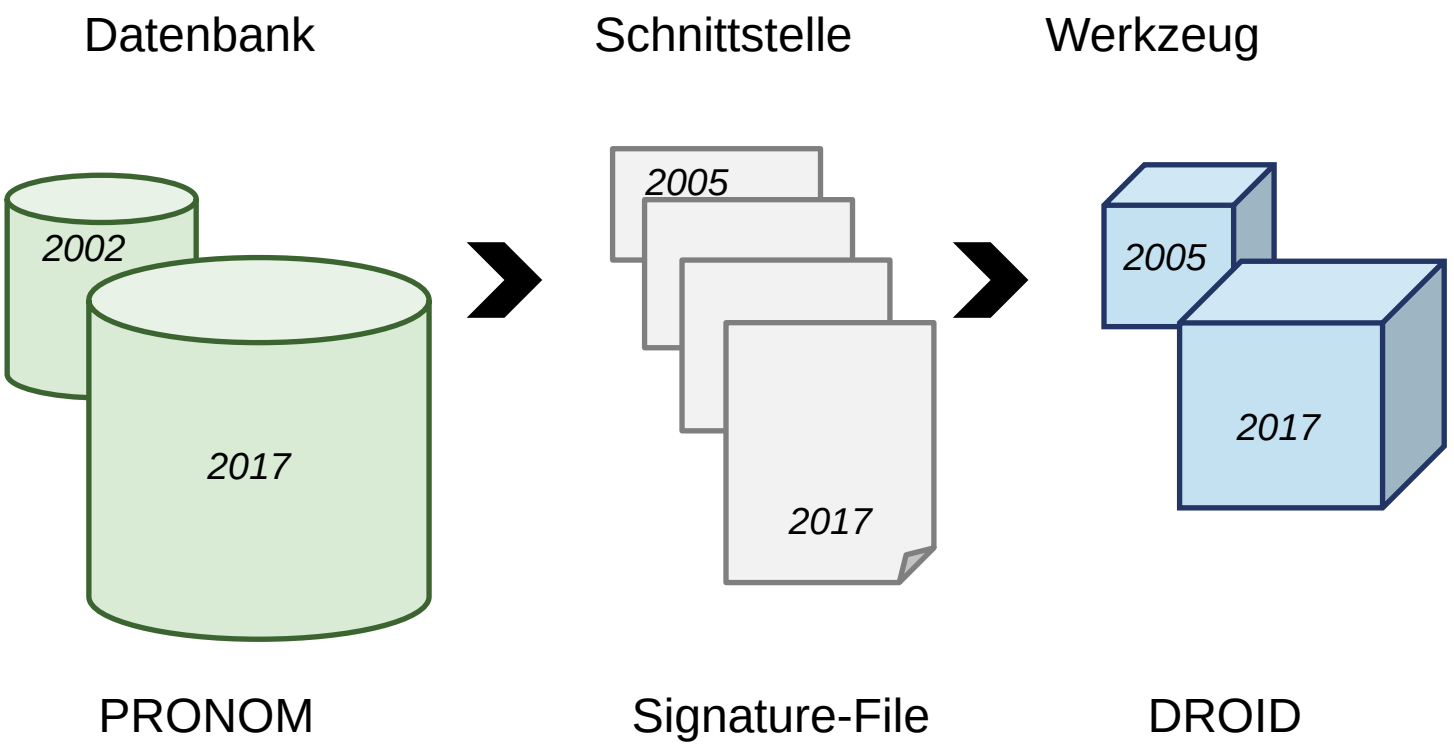


64 deprecated PUIDs
(Stand 2017)

Keine PUID. Und dann?

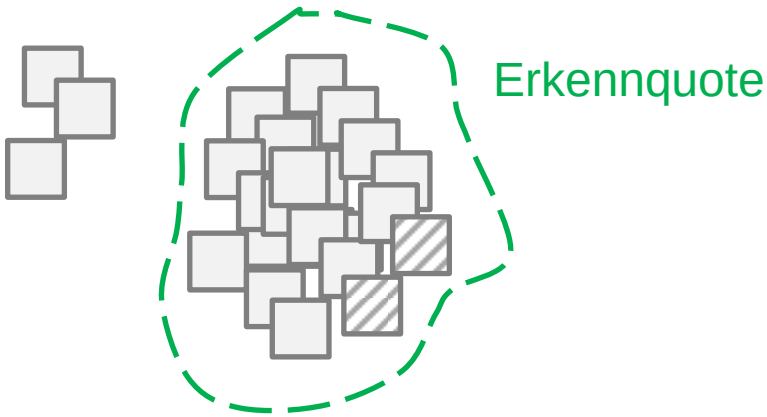


Fazit


Datenbank
Schnittstelle
Werkzeug
2002
2017
2005
2017
2005
2017
PRONOM
Signature-File
DROID

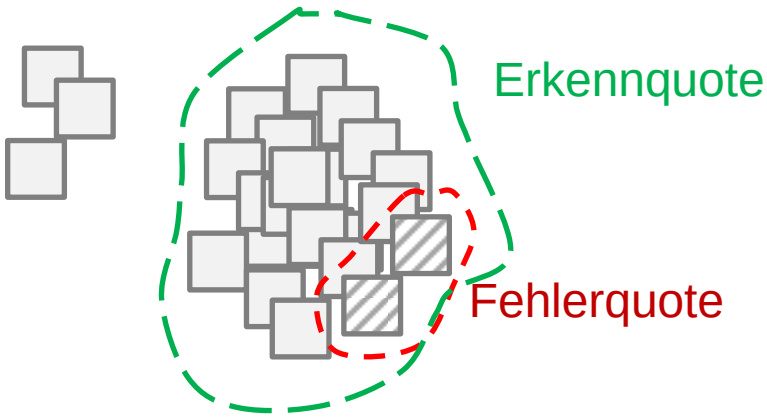
Fazit

- trotz Weiterentwicklung



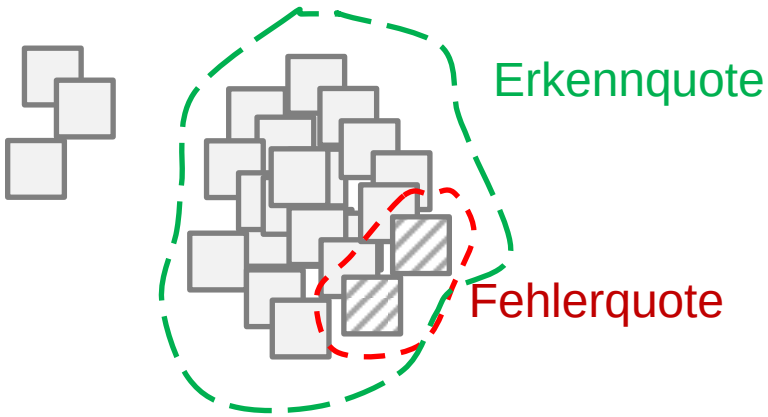
Fazit

- trotz Weiterentwicklung



Fazit

- trotz Weiterentwicklung



Formaterkennung an sich…

- “(…) it’s a **work in progress**, and therefore could not be considered complete at this time.”
- “Perhaps it is only possible to have a **relative ‘truth’** (…)”

(Jay Gattuso, NZNL, 2012)

Was können wir machen?

- kurzfristig:
  - andere Tools hinzuziehen (z.B. Siegfried, R. Lehane)
- langfristig:
  - Stärkere Vernetzung (auf Arbeitsebene)
  - Plattformausbau?
    - nestor (AGs, Praktikertage, Workshops, Publikationskanäle, …)
    - Vorbild KOST, LWL
    - Anwendertreffen, Fachtagungen

Vielen Dank!

el_sta@sta.smi.sachsen.de

Vielen Dank!

el_sta@sta.smi.sachsen.de

	00	01	02	03
0:	8B	AD	F0	0D
4:	BE	EF	CA	CE
8:	CA	FE	BA	BE

Vielen Dank!

el_sta@sta.smi.sachsen.de

	00	01	02	03
0:	8B	AD	F0	0D
4:	BE	EF	CA	CE
8:	CA	FE	BA	BE

→ 8 bad food
beef cake
cafe babe

Diskussion?

- Ähnliche Erfahrungen?
- Kein PUID. Und dann?
- **Rerun?**
- Ist PRONOM genug?
- Offenes <-> restriktives Repository?