



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

I Erkennung

- I Identifizierung von charakteristischen Mustern
- I z. B. DROID, Tika, file, Siegfried

I Validierung

- I Norm- bzw. Spezifikationskonformität
- I z.B. KOST-VAL, JHOVE, W3C



Formaterkennung – wie?

I Extension / Dateiendung

- I fehlt, falsch, nicht eindeutig
 - I 37 PDF-Varianten → .pdf
 - I JPEG → .jpe, .jpeg, .jpg, ...

I MIME-Type

- I zu ungenau
 - I „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

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	10	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F			
000:	3C	21	44	4F	43	54	59	50	45	20	48	54	4D	4C	20	50	55	<	!	D	O	C	T	E	>	H	T	M	L		P	U			
011:	42	4C	49	43	20	22	2D	2F	2F	57	33	43	2F	2F	44	54	44	B	L	I	C		-	-	/	/	W	3	C	/	/	D	T		
022:	20	48	54	4D	4C	20	34	2E	30	31	2F	2F	45	4E	22	20	22	H	T	M	L		4	.	0	1	/	/	E	N	"	"			
033:	68	74	74	70	3A	2F	2F	77	77	77	2E	77	33	2E	6F	72	67	h	t	t	p	:	/	/	w	w	w	.	w	3	.	o	r	g	
044:	2F	54	52	2F	68	74	6D	6C	34	2F	73	74	72	69	63	74	2E	/	T	R	/	h	t	m	l	4	/	s	t	r	i	c	t	.	
055:	64	74	64	22	3E	0D	0A	3C	68	74	6D	6C	20	6C	61	6E	67	d	t	d	"	>	..	<	h	t	m		l	a	n	g	"	"	
066:	3D	22	64	65	22	3E	0D	0A	20	3C	68	65	61	64	3E	0D	0A	=	"	d	e	"	>	..	<	h	e	a	d	>	..				
077:	20	20	3C	6D	65	74	61	20	68	74	74	70	2D	65	71	75	69	<	m	e	t	a		h	t	t	p	-	e	q	u	i	v	"	"

File extension: htm
File extension: html
Name HTML 4.01
Description With DOCTYPE declaration of 4.01. Single byte sequence. BOF character sequence: {0-1024}<!(DOCTYPE doctype) (HTML html) (PUBLIC/public) "-//{1-16}///(DTD dtd) {0-64}(HTML html) 4.01

Signaturen am Beispiel

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

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	10	0123456789ABCDEF0
000:	3C	21	44	4F	43	54	59	50	45	20	48	54	4D	4C	20	50	55	<!DOCTYPE HTML PU
011:	42	4C	49	43	20	22	2D	2F	2F	57	33	43	2F	2F	44	54	44	BLIC "-//W3C//DTD
022:	20	48	54	4D	4C	20	34	2E	30	31	2F	2F	45	4E	22	20	22	HTML 4.01//EN" "
033:	68	74	74	70	3A	2F	2F	77	77	77	2E	77	33	2E	6F	72	67	http://www.w3.org
044:	2F	54	52	2F	68	74	6D	6C	34	2F	73	74	72	69	63	74	2E	/TR/html4/strict.
055:	64	74	64	22	3E	0D	0A	3C	68	74	6D	6C	20	6C	61	6E	67	dtd">..<html lang
066:	3D	22	64	65	22	3E	0D	0A	20	3C	68	65	61	64	3E	0D	0A	="de">.. <head>..
077:	20	20	3C	6D	65	74	61	20	68	74	74	70	2D	65	71	75	69	<meta http-equi

File extension: htm
File extension: html
Name HTML 4.01
Description With DOCTYPE declaration of 4.01. Single byte sequence. BOF character sequence: {0-1024}<!(DOCTYPE doctype) (HTML html) (PUBLIC/public) "-//{1-16}//(DTD dtd) {0-64}(HTML html) 4.01

Signaturen am Beispiel

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

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F																	
000:	3C	21	44	4F	43	54	59	50	45	20	48	54	4D	4C	20	50	55	<	!	(D	O	C	T	E		H	T	M	L		P	U																
011:	42	4C	49	43	20	22	2D	2F	2F	57	33	43	2F	2F	44	54	42	B	L	I	C		"	-	/	/	W	3	C	/	/	D	T	D															
022:	20	48	54	4D	4C	20	34	2E	30	31	2F	2F	45	4E	22	20	22	H	T	M	L		"	4	.	0	1	/	/	E	N	"																	
033:	68	74	74	70	3A	2F	2F	77	77	77	2E	77	33	2E	6F	72	67	h	t	t	p		/	/	w	w	w	.	w	3	.	o	r	g															
044:	2F	54	52	2F	68	74	6D	6C	34	2F	73	74	72	69	62	74	2E	/	T	/	h	t	m	l	4	/	s	t	r	i	c	t	.																
055:	64	74	64	22	3E	0D	0A	3C	68	74	6D	6C	20	6C	61	6E	67	a	t	d	"	>	..	<	h	t	m		l	a	n	g																	
066:	3D	22	64	65	22	3E	0D	0A	20	3C	68	65	61	64	3E	0D	0A	=	"	d	e	"	>	..	<	h	e	a	d	>	..	<	h	e	a	d	>	..	<	h	o	d	y	>	<	h	1	>
077:	20	20	3C	6D	65	74	61	20	68	74	74	70	20	65	71	75	69	<	m	e	t	a		h	t	t	p	-	e	q	u	i	v	=	"	c	o	n	t	e	n	t	-	t	y	p	e	"	

File extension: htm
File extension: html

Name	HTML 4.01
Description	With DOCTYPE declaration of 4.01. Single byte sequence. BOF character sequence: {0-1024}<!(DOCTYPE doctype) (HTML html) (PUBLIC/public) "-//{1-16}///(DTD dtd) {0-64}(HTML html) 4.01

Signaturen am Beispiel

- maschinenlesbar

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
000:	3C	21	44	4F	43	54	59	50	45	20	48	54	4D	4C	20	50	55	<!	DOCTYPE	HTML	PU											
011:	42	4C	49	43	20	22	2D	2F	2F	57	33	43	2F	2F	44	54	44	BLIC	"-//	W3C//	DTD											
022:	20	48	54	4D	4C	20	34	2E	30	31	2F	2F	45	4E	22	20	22	HTML	4.01	//	EN"	"										
033:	68	74	74	70	3A	2F	2F	77	77	77	2E	77	33	2E	6F	72	67	http://	www.w3.org													
044:	2F	54	52	2F	68	74	6D	6C	34	2F	73	74	72	69	63	74	2E	/	TR/html4/	strict.												
055:	64	74	64	22	3E	0D	0A	3C	68	74	6D	6C	20	6C	61	6E	67	dtd">	..	<	html	lang										
066:	3D	22	64	65	22	3E	0D	0A	20	3C	68	65	61	64	3E	0D	0A	=	"de">	..	<	head>	..									
077:	20	20	3C	6D	65	74	61	20	68	74	74	70	2D	65	71	75	69	<	meta	http-equi												
088:	76	3D	22	63	6F	6E	74	65	6E	74	2D	74	79	70	65	22	20	v="	content-type"													
099:	63	6F	6E	74	65	6F	74	3D	22	74	65	78	74	2E	68	74	6D	content="	text/h	tm												
0AA:	6C	3B	20	63	68	Value												3C21	(444F4354595045	646F6374797065)20												
0BB:	3E	0D	0A	20	3C													(48544D4C 68746D6C)20														
0CC:	64	79	3E	0D	0A													(5055424C4943 7075626C6963)20	222D2F2F{1-16}2F2F													
0DD:	57	69	6C	6C	6B													(445444 647464)20	{0-64}(48544D4C 68746D6C)													
0EE:	72	20	41	55	64													20342E3031														
0FF:	0A	0D	0A	0D	0A																											
110:	68	74	6D	6C	3E																											

Signaturen am Beispiel

- maschinenlesbar

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	10	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
000:	3C	21	44	4F	43	54	59	50	45	20	48	54	4D	4C	20	50	55	<!DOCTYPE HTML PU														
011:	42	4C	49	43	20	22	2D	2F	2F	57	33	43	2F	2F	44	54	44	BLIC "-//W3C//DTD														
022:	20	48	54	4D	4C	20	34	2E	30	31	2F	2F	45	4E	22	20	22	HTML 4.01//EN" "														
033:	68	74	74	70	3A	2F	2F	77	77	77	2E	77	33	2E	6F	72	67	http://www.w3.org														
044:	2F	54	52	2F	68	74	6D	6C	34	2F	73	74	72	69	63	74	2E	/TR/html4/strict.														
055:	64	74	64	22	3E	0D	0A	3C	68	74	6D	6C	20	6C	61	6E	67	dtd">..<html lang														
066:	3D	22	64	65	22	3E	0D	0A	20	3C	68	65	61	64	3E	0D	0A	="de">.. <head>..														
077:	20	20	3C	6D	65	74	61	20	68	74	74	70	2D	65	71	75	69	<meta http-equi														
088:	76	3D	22	63	6F	6E	74	65	6E	74	2D	74	79	70	65	22	20	v="content-type"														
099:	63	6F	6E	74	65	6F	74	3D	22	74	65	78	74	2E	68	74	6D	content="text/h														
0AA:	6C	3B	20	63	68	Value	3C21(444F4354595045	646F6374797065)20																								
0BB:	3E	0D	0A	20	3C		(48544D4C 68746D6C)20																									
0CC:	64	79	3E	0D	0A		(5055424C4943 7075626C6963)20222D2F2F{1-16}2F2F																									
0DD:	57	69	6C	6C	6B		(445444 647464)20{0-64}(48544D4C 68746D6C)																									
0EE:	72	20	41	55	64		20342E3031																									
0FF:	0A	0D	0A	0D	0A																											
110:	68	74	6D	6C	3E																											

In der Praxis

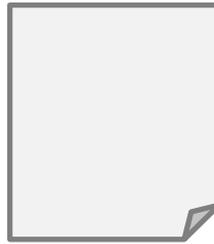
Datenbank



PRONOM



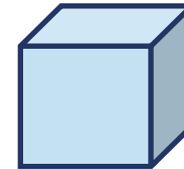
Schnittstelle



Signature-File



Werkzeug



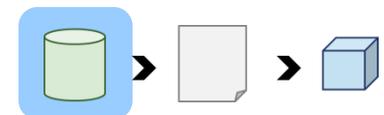
DROID

PRONOM

The National Archives

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

PUID Schema:
fmt/18



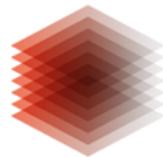
PRONOM

A 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



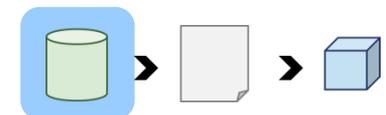
TIB LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK



Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY
OF NEW ZEALAND



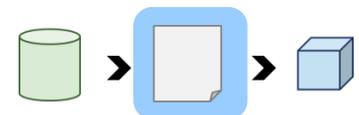
ETH zürich



(DROID) Signature-File

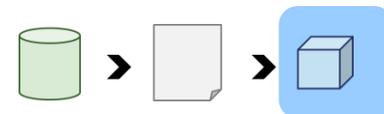
- aus Pronom-Datenbank generiert
 - Relativ regelmäßig, mehrmals im Jahr
- Aufbau
 - Menschenlesbare Auflistung der Formate („fmt“)
 - technischer Abschnitt zur Erkennung der Signaturen

```
1      <?xml version="1.0" encoding="UTF-8"?>
2      [ ] <FFSignatureFile DateCreated="2016-09-27T15:37:53" Version="88" xmlns="
3          <InternalSignatureCollection>
30969  <FileFormatCollection>
38274  </FFSignatureFile>
38275
```



DROID – was und wie

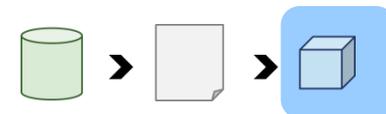
- l Digital Record Object Identification
- l Entwickler TNA, Start 2005
- l 2017: v6.3
- l plattformunabhängig, Java-Umgebung
- l GUI oder Kommandozeile
- l weltweite Verbreitung
- l frei, Open Source



DROID – was und wie

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

Resource	Extension	Ids	Method	PUID	Format	Version	Mime type	Hash
△ Resource								
📁 C:\Users\tksuser\Desktop\A								
📁 B								
📄 Test1.txt	txt	⚙️	Extension	x-fmt/111	Plain Text File		text/plain	30aba6f8
📄 Test2.txt		🌐						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	3af968at
📄 C:\Users\tksuser\Desktop\Test6.jpg	jpg	⚙️	Signature	fmt/645	Exchangeable Image File Format (Compressed)	2.2.1	image/jpeg	8af1a77c
📄 C:\Users\tksuser\Desktop\Test7.lck	lck	(2)	Extension	"fmt/335", "fmt/218"	"Dreamweaver Lock File", "Microsoft FrontPage"			7b60bf4e



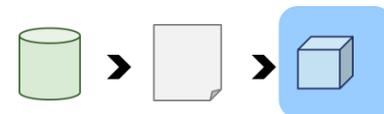


Grenzen von DROID & PRONOM



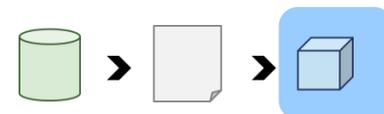
DROID – Signature sticht

1. Überprüfung auf Signatures
2. Überprüfung auf Extensions



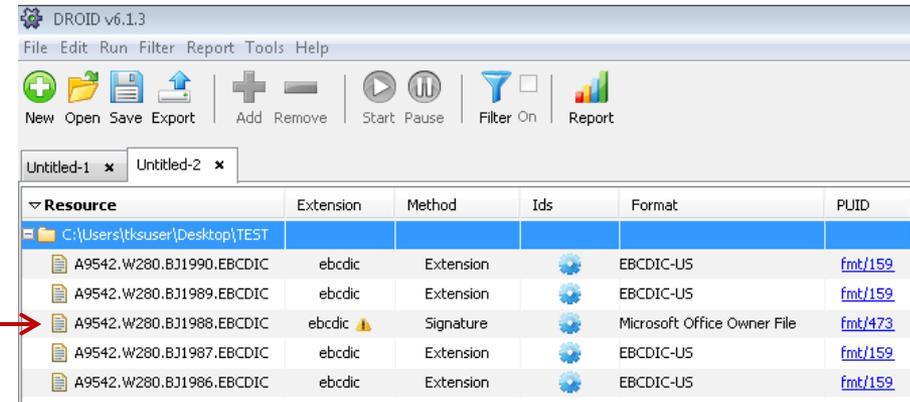
DROID – Signature sticht

1. Überprüfung auf Signatures
 2. Überprüfung auf Extensions
- █ Echtdaten in Plain Text / EBCDIC
- erwartet: fmt/159 (EBCDIC-US)

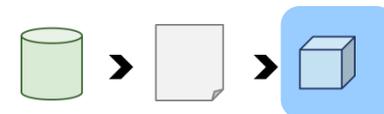


DROID – Signature sticht

1. Überprüfung auf Signatures
 2. Überprüfung auf Extensions
- █ Echtdaten in Plain Text / EBCDIC
- erwartet: fmt/159 (EBCDIC-US)
- Ergebnis: fmt/473 (MS Owner File)



Resource	Extension	Method	Ids	Format	PUID
C:\Users\tkuser\Desktop\TEST					
A9542.W280.B31990.EBCDIC	ebcdic	Extension		EBCDIC-US	fmt/159
A9542.W280.B31989.EBCDIC	ebcdic	Extension		EBCDIC-US	fmt/159
A9542.W280.B31988.EBCDIC	ebcdic	Signature		Microsoft Office Owner File	fmt/473
A9542.W280.B31987.EBCDIC	ebcdic	Extension		EBCDIC-US	fmt/159
A9542.W280.B31986.EBCDIC	ebcdic	Extension		EBCDIC-US	fmt/159



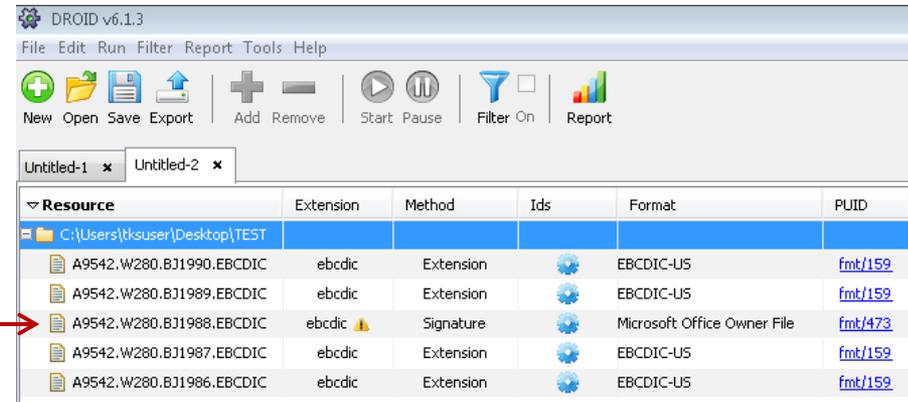
DROID – Signature sticht

1. Überprüfung auf Signatures
2. Überprüfung auf Extensions

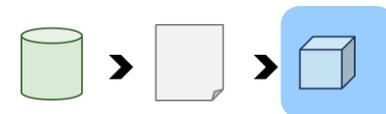
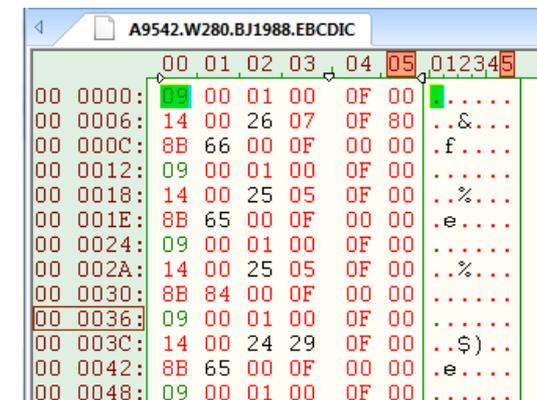
I Echtdaten in Plain Text / EBCDIC

→ 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



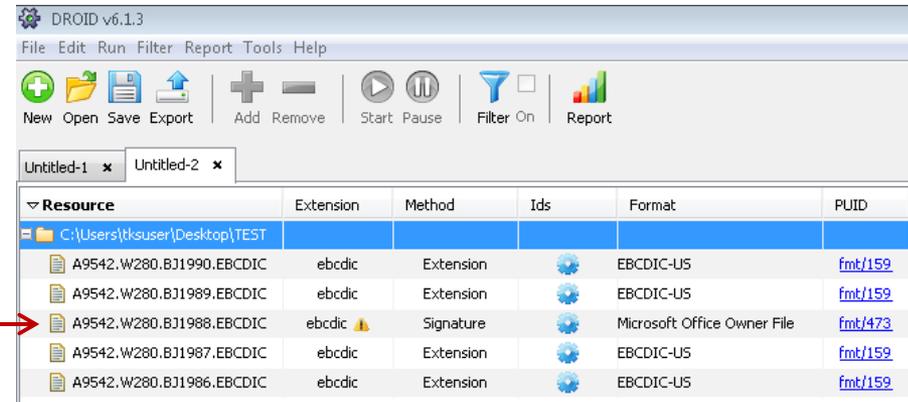
DROID – Signature sticht

1. Überprüfung auf Signatures
2. Überprüfung auf Extensions

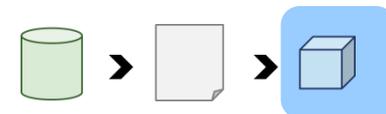
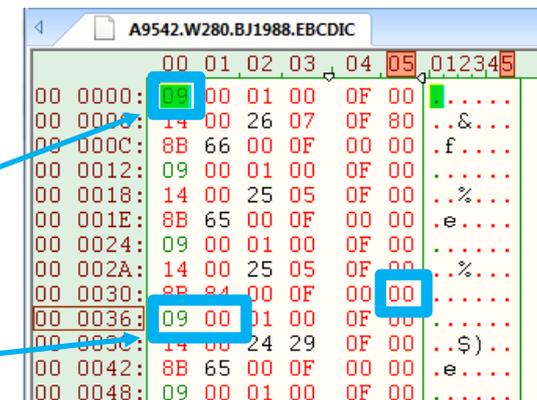
I Echtdaten in Plain Text / EBCDIC

→ erwartet: fmt/159 (EBCDIC-US)

→ Ergebnis: fmt/473 (MS Owner File)

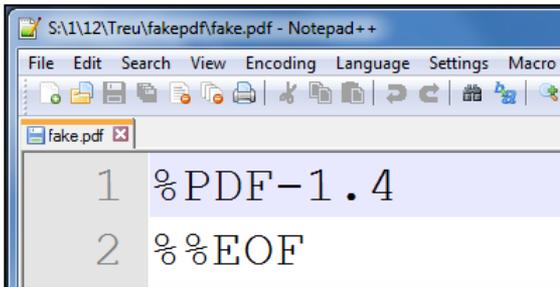


Name	MS Owner file v4	
Description	1st byte is within the range of 06 to 0F. Next bytes are name of registered user. 55th byte is identical to 1st. 56th byte is 00	
Byte sequences	Position type	Absolute from BOF
	Offset	0
	Byte order	
	Value	<u>09{52}000900</u>

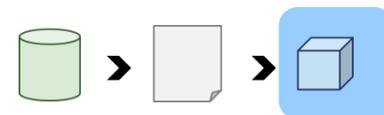


DROID auf dem Glatteis

I nicht funktionsfähige Datei

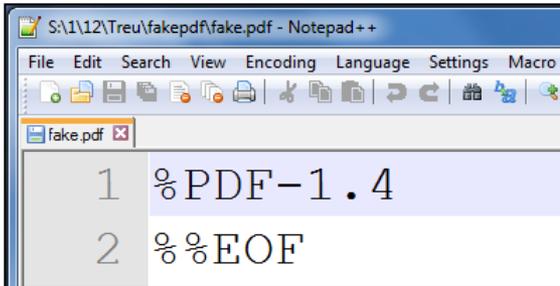


```
S:\1\12\Treu\fakepdf\fake.pdf - Notepad++
File Edit Search View Encoding Language Settings Macro
fake.pdf
1 %PDF-1.4
2 %%EOF
```



DROID auf dem Glatteis

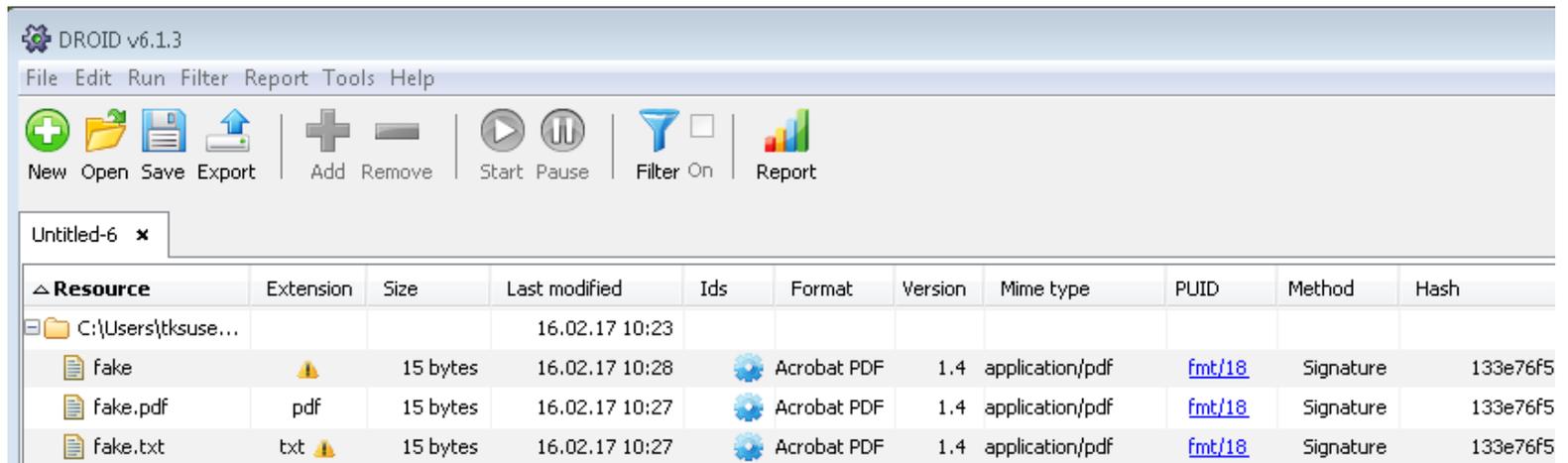
I nicht funktionsfähige Datei



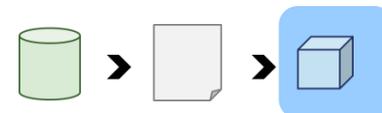
```
S:\1\12\Treu\fakepdf\fake.pdf - Notepad++
File Edit Search View Encoding Language Settings Macro
fake.pdf
1 %PDF-1.4
2 %%EOF
```

→ DROID lässt sich einfach täuschen

→ arbeitet aber erwartungsgemäß / zuverlässig



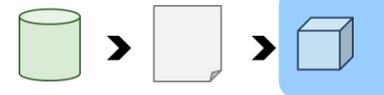
Resource	Extension	Size	Last modified	Ids	Format	Version	Mime type	PUID	Method	Hash
C:\Users\tksuse...			16.02.17 10:23							
fake	⚠	15 bytes	16.02.17 10:28	⚙	Acrobat PDF	1.4	application/pdf	fmt/18	Signature	133e76f5
fake.pdf	pdf	15 bytes	16.02.17 10:27	⚙	Acrobat PDF	1.4	application/pdf	fmt/18	Signature	133e76f5
fake.txt	txt ⚠	15 bytes	16.02.17 10:27	⚙	Acrobat PDF	1.4	application/pdf	fmt/18	Signature	133e76f5



Ergebnisdifferenzen

Echtdaten

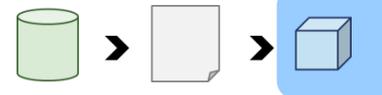
Datei	DROID v6.1.3 (SignFile V79, MBS)	DROID v6.3 (SignFile V88, full scan)	Siegfried (1.6.7)	Formatname
A1.jpg	fmt/41	fmt/645	fmt/645	Raw JPEG Stream ↔ EXIF 2.2.1
B1.pdf	fmt/18	fmt/95	fmt/95	PDF 1.4 ↔ PDF/A-1a
B2.pdf	fmt/276	fmt/478	fmt/478	PDF 1.7 ↔ PDF/A-2u



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



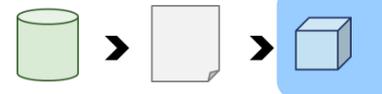
DROID – Scan-Modi

BOF

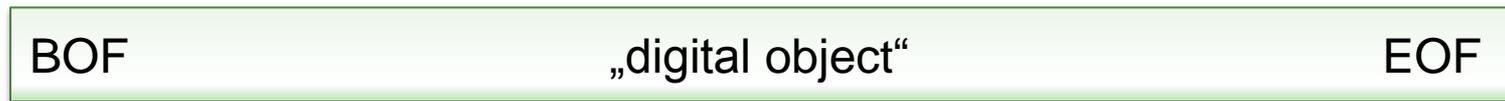
„digital object“

EOF

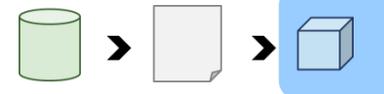
Nach Jay Gattuso (NZNL)



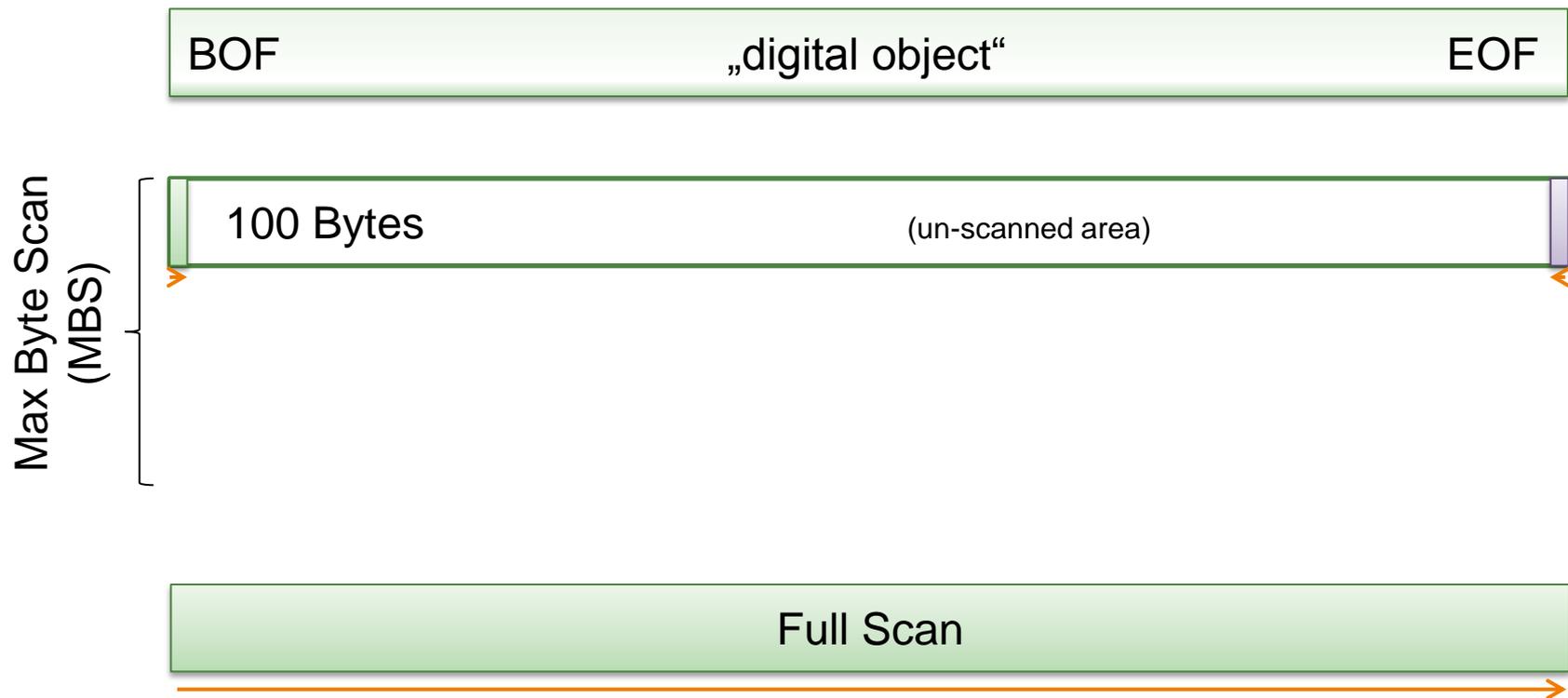
DROID – Scan-Modi



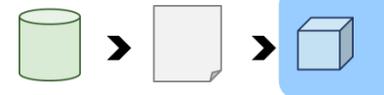
Nach Jay Gattuso (NZNL)



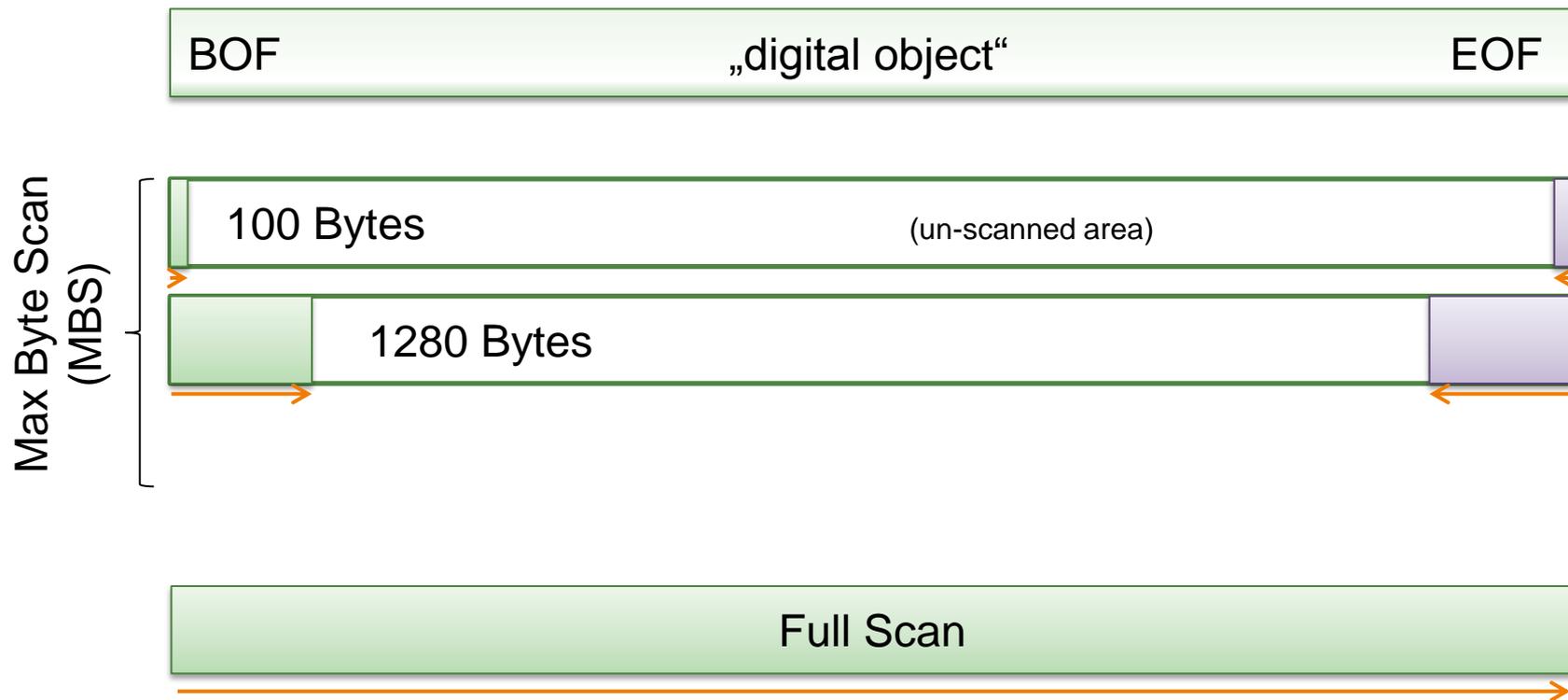
DROID – Scan-Modi



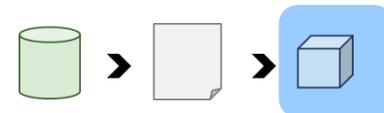
Nach Jay Gattuso (NZNL)



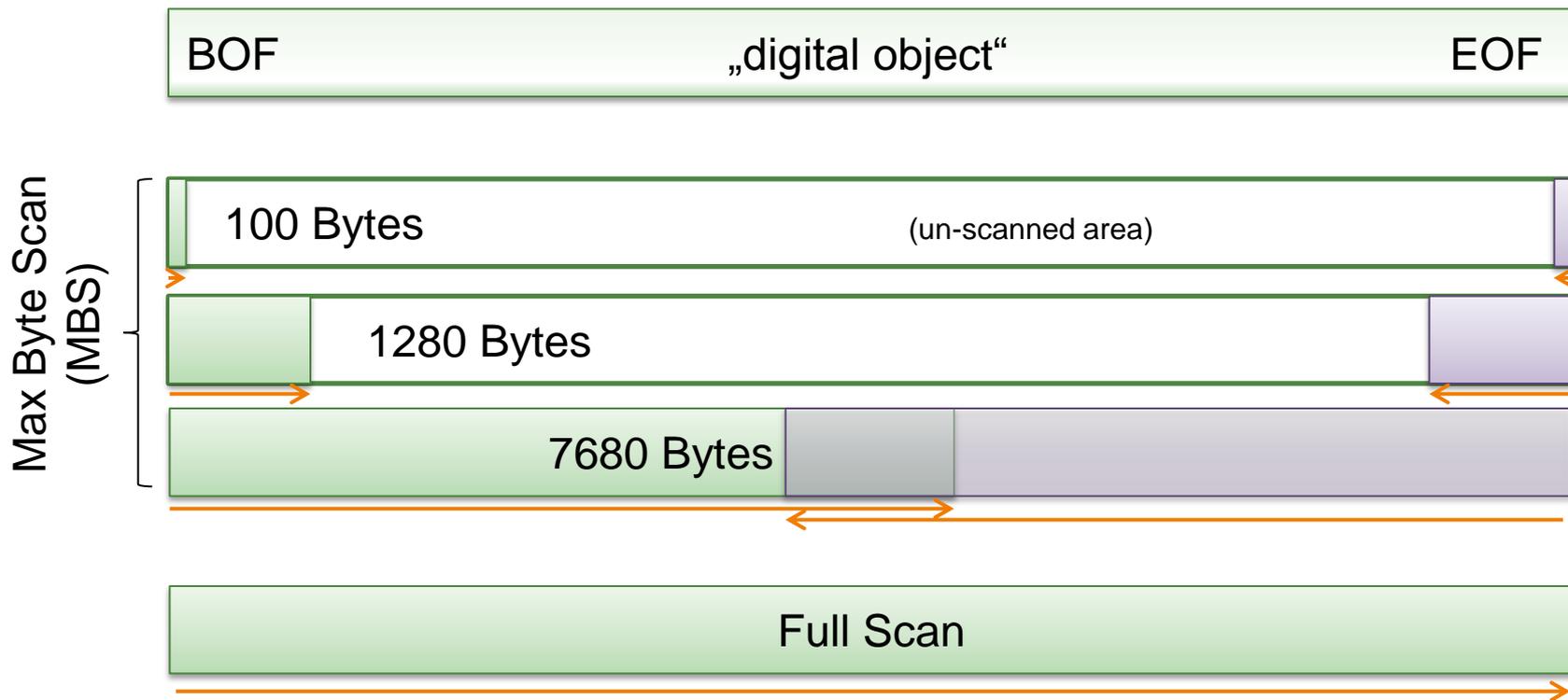
DROID – Scan-Modi



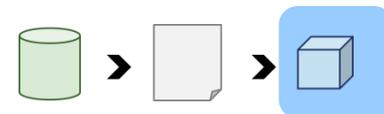
Nach Jay Gattuso (NZNL)



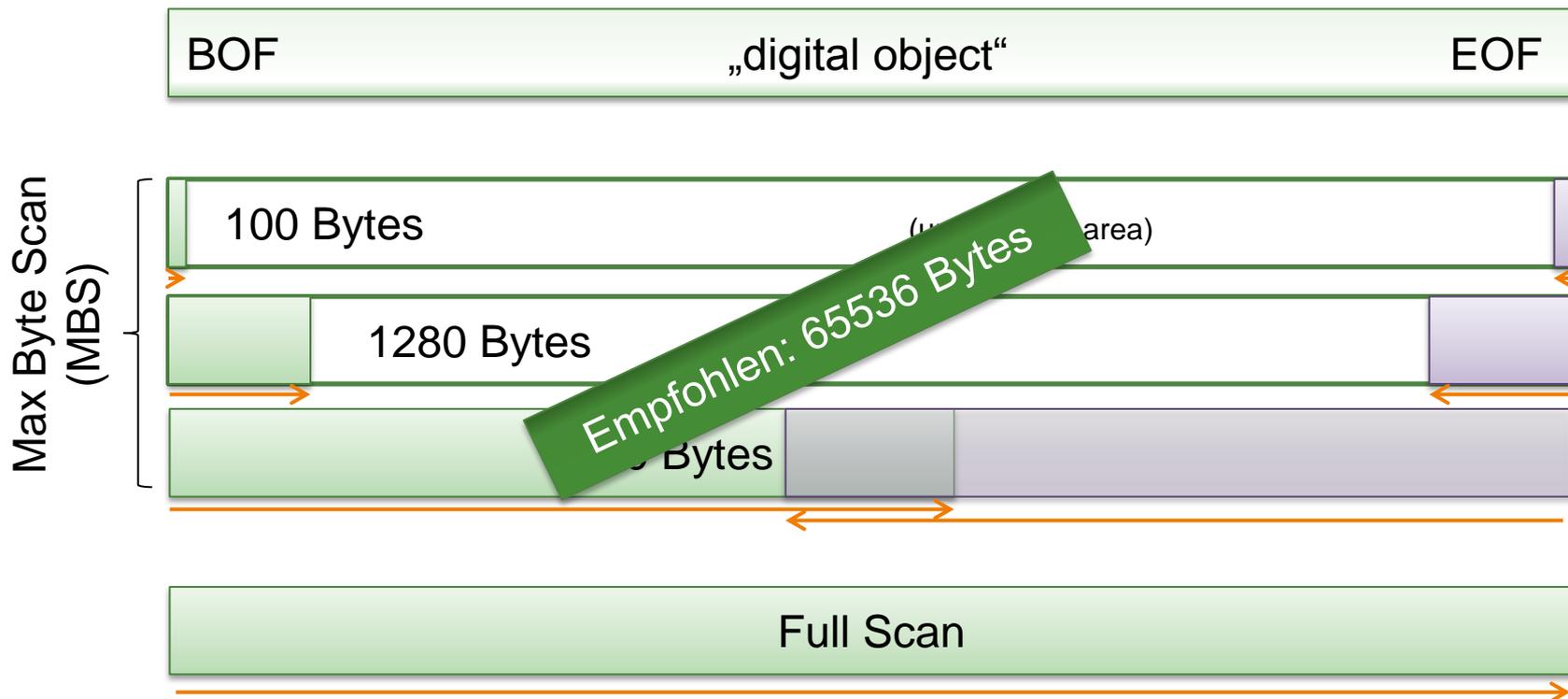
DROID – Scan-Modi



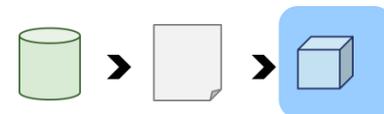
Nach Jay Gattuso (NZNL)



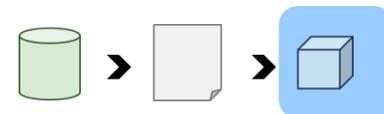
DROID – Scan-Modi



Nach Jay Gattuso (NZNL)



Scan-Beispiel



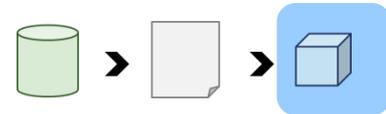
Scan-Beispiel

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>  
=  
786D6C6E733A7064666169643D(22(27)687474703A2F2F7777772E6169696D2E6F72672F706466612F  
6E732F69647064666169643A70617274(3D22(3D27(3E)31(22(27(3C2F7064666169643A706172743E)(0  
11)7064666169643A636F6E666F726D616E6365(3E(3D22(3D27)41(22(27(3C2F7064666169643A636F6E  
666F726D616E63653E)
```

%%EOF
=
2525454F46



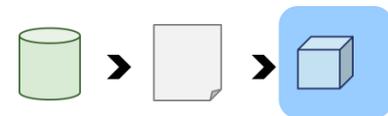
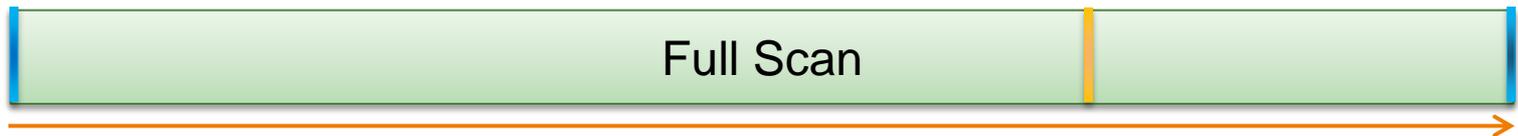
Scan-Beispiel

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>  
=  
798D6C6E733A7064666169643D(22(27)687474703A2F2F7777772E6169696D2E6F72672F706466612F  
6E732F69647064666169643A70617274(3D22(3D27(3E)31(22(27(3C2F7064666169643A706172743E)0  
11)7064666169643A636F6E666F726D616E6365(3E)3D22(3D27)41(22(27(3C2F7064666169643A636F6E  
666F726D616E63653E)
```

%%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>
=
786D6C6E733A7064666169643D(22(27)687474703A2F2F7777772E6169696D2E6F72672F706466612F
6E732F69647064666169643A70617274(3D22(3D27(3E)31(22(27(3C2F7064666169643A706172743E)0
11)7064666169643A636F6E666F726D616E6365(3E)3D22(3D27)41(22(27(3C2F7064666169643A636F6E
666F726D616E63653E)

%%EOF
=
2525454F46



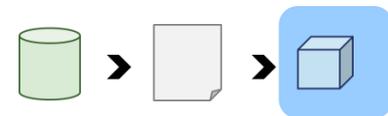
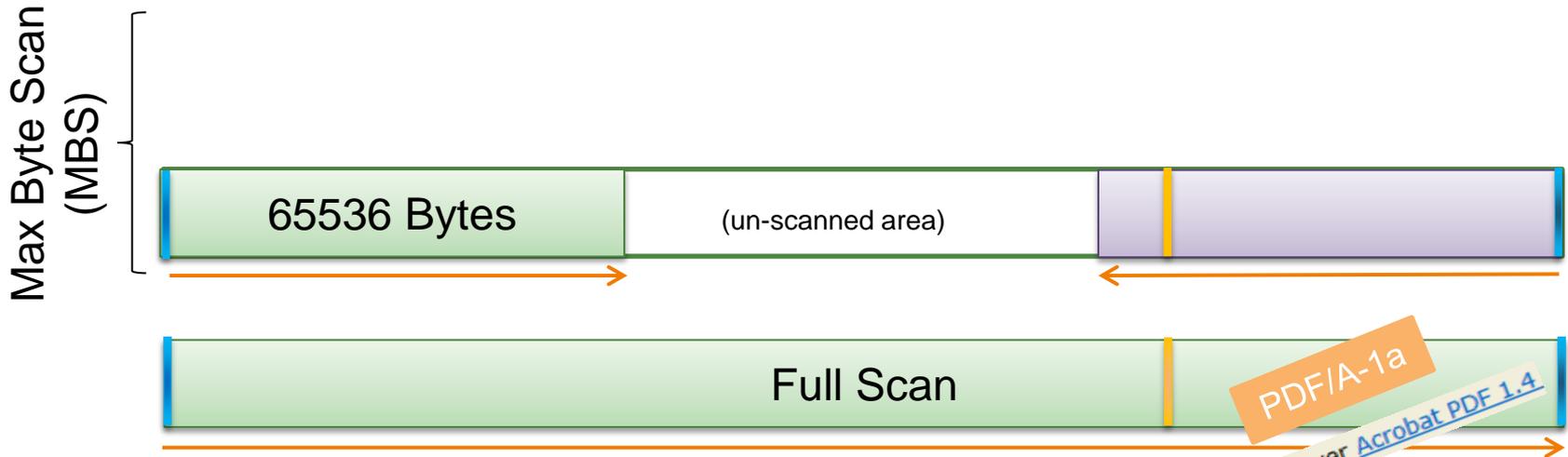
Scan-Beispiel

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>
=
798D6C6E733A7064666169643D(22(27(687474703A2F2F7777772E6169696D2E6F72672F706466612F
6E732F69647064666169643A70617274(3D22(3D27(3E)31(22(27(3C2F7064666169643A706172743E)0
11)7064666169643A636F6E666F726D616E6365(3E(3D22(3D27)41(22(27(3C2F7064666169643A636F6E
666F726D616E63653E)

%%EOF
=
2525454F46



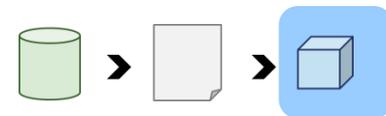
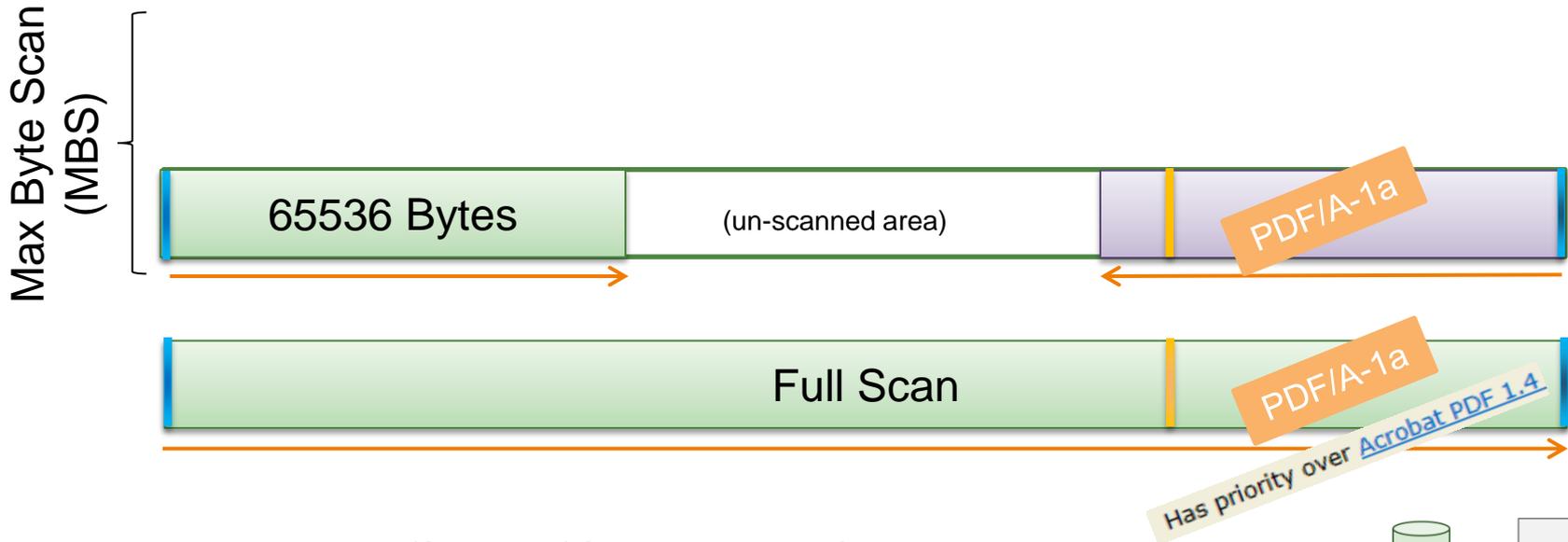
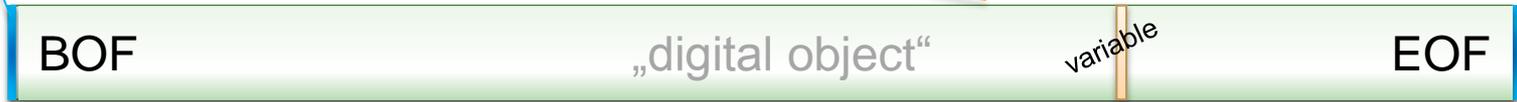
Scan-Beispiel

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>  
=  
798D6C6E733A7064666169643D(22(27(687474703A2F2F7777772E6169696D2E6F72672F706466612F  
6E732F69647064666169643A70617274(3D22(3D27(3E)31(22(27(3C2F7064666169643A706172743E)0  
11)7064666169643A636F6E666F726D616E6365(3E)3D22(3D27)41(22(27(3C2F7064666169643A636F6E  
666F726D616E63653E)
```

%%EOF
=
2525454F46



Scan-Beispiel

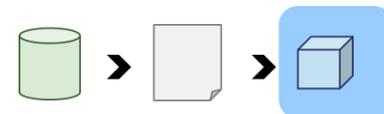
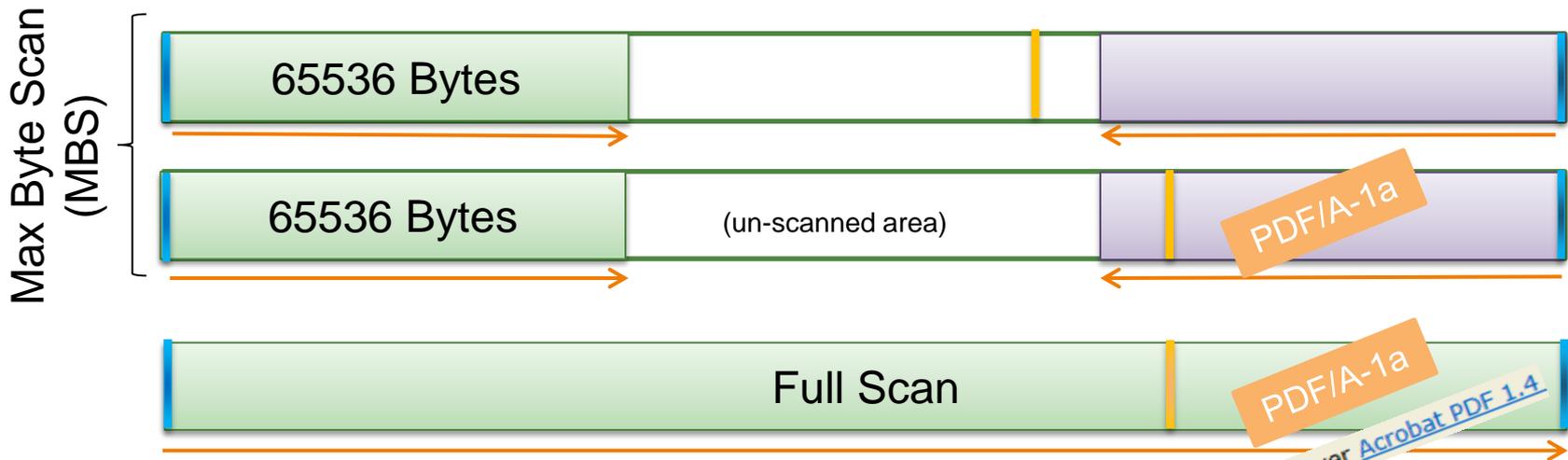
PDF/A-1a= fmt/95

%PDF-1.4
=
255044462D312E34

```

xmlns:pdfaid="http://www.aiim.org/pdfa/ns/id/
<pdfaid:part>1<pdfaid:part>
<pdfaid:conformance>A<pdfaid:conformance>
=
798D6C6E733A7064666169643D(22(27(687474703A2F2F7777772E6169696D2E6F72672F706466612F
6E732F69647064666169643A70617274(3D22(3D27(3E)31(22(27(3C2F7064666169643A706172743E)0
11)7064666169643A636F6E666F726D616E6365(3E)3D22(3D27)41(22(27(3C2F7064666169643A636F6E
666F726D616E63653E)
    
```

%%EOF
=
2525454F46



Scan-Beispiel

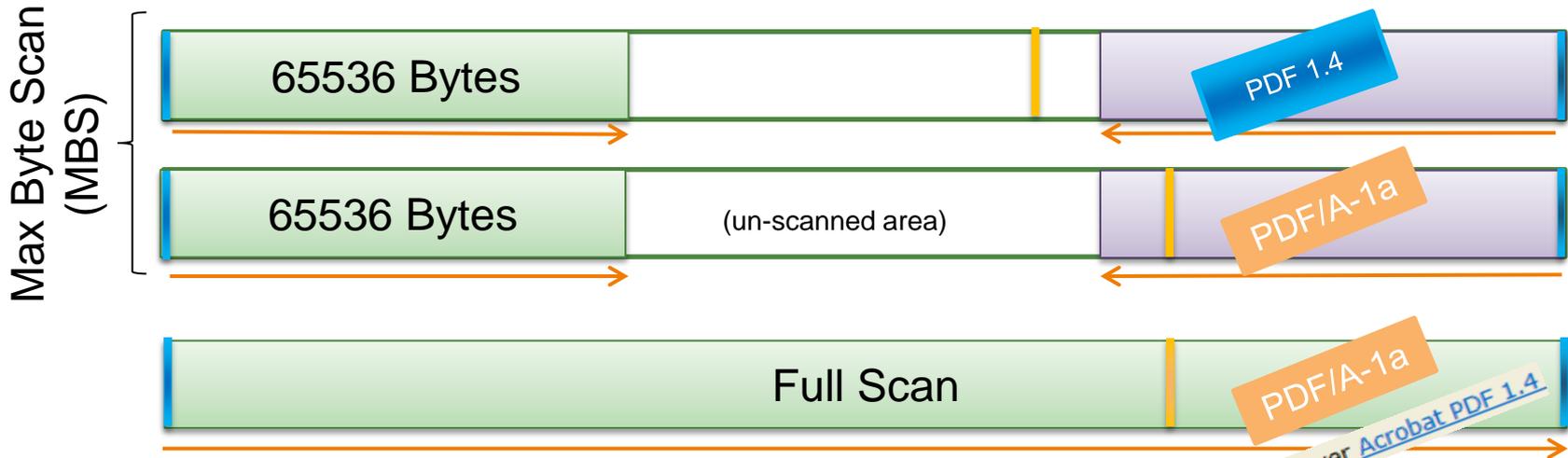
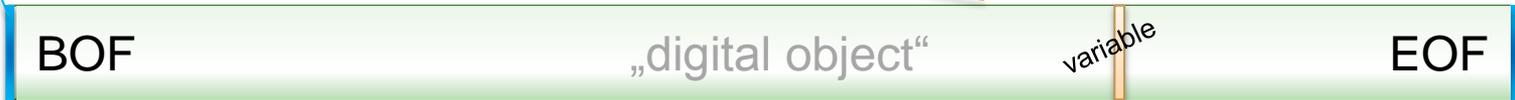
PDF/A-1a= fmt/95

%PDF-1.4
=
255044462D312E34

```

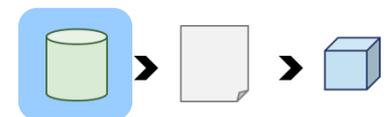
xmlns:pdfaid="http://www.aiim.org/pdfa/ns/id/
<pdfaid:part>1<pdfaid:part>
<pdfaid:conformance>A<pdfaid:conformance>
=
798D6C6E733A7064666169643D(22(27(687474703A2F2F7777772E6169696D2E6F72672F706466612F
6E732F69647064666169643A70617274(3D22(3D27(3E)31(22(27(3C2F7064666169643A706172743E)0
11)7064666169643A636F6E666F726D616E6365(3E)3D22(3D27)41(22(27(3C2F7064666169643A636F6E
666F726D616E63653E)
    
```

%%EOF
=
2525454F46

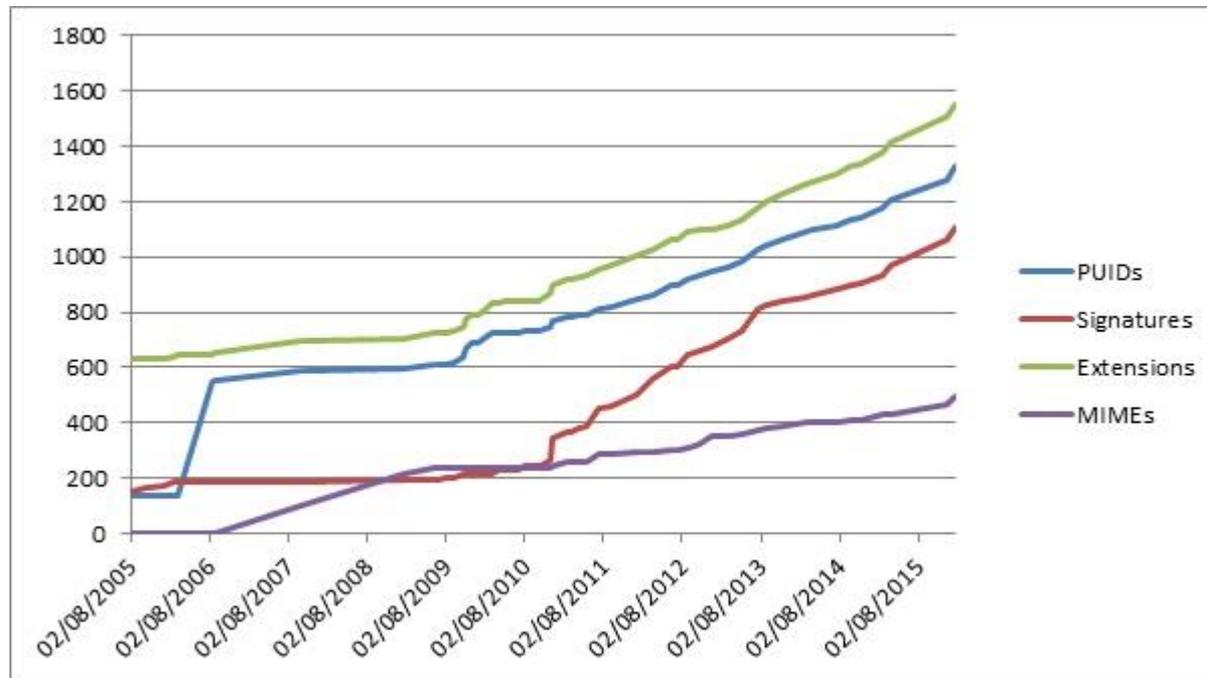




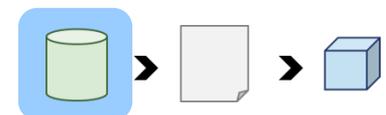
Weiterentwicklung der Datenbank



PRONOM-Entwicklung (2005 – 2016)



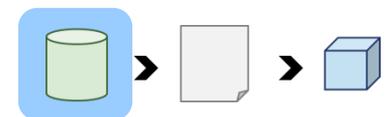
Quelle: Paul Young (<http://blog.nationalarchives.gov.uk/blog/identifying-digital-file-formats-collaborative-effort/>)



Weiterentwicklung von PRONOM

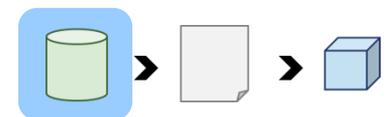
- Neue Einträge
- Vorläufige Einträge
- Zusammenlegung von Einträgen
- Überholte Einträge - „Deprecated“

- → Persistenz?!



Ehemals vorläufige Einträge

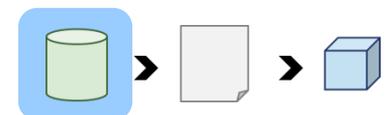
Name	Acrobat PDF 1.4	Comma Separated Values
Version	1.4	
Other names	PDF (1.4)	
Identifiers	MIME: application/pdf <u>PUID: fmt/18</u>	MIME: text/csv <u>PUID: x-fmt/18</u>



Ehemals vorläufige Einträge

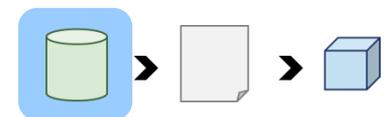
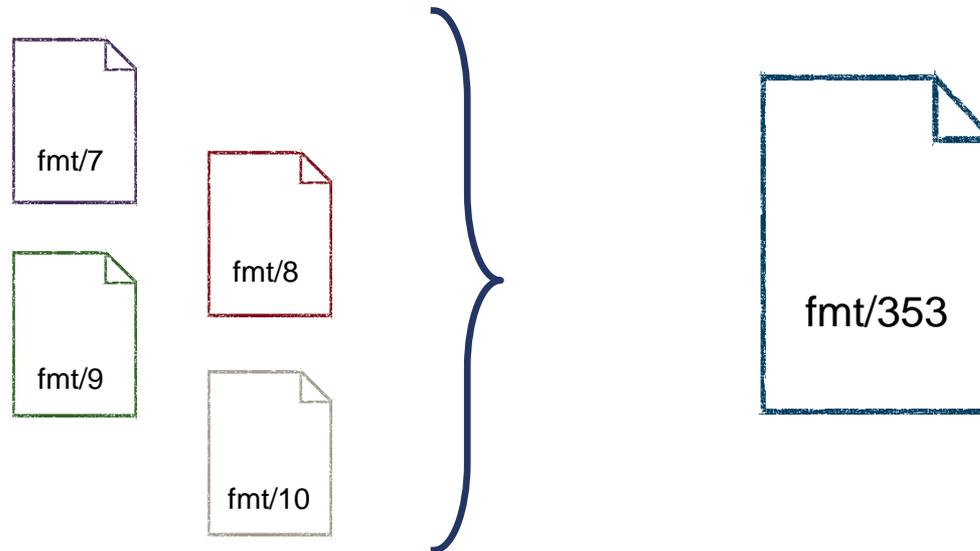
Name	Acrobat PDF 1.4	Comma Separated Values
Version	1.4	
Other names	PDF (1.4)	
Identifiers	MIME: application/pdf <u>PUID: fmt/18</u>	MIME: text/csv <u>PUID: x-fmt/18</u>

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



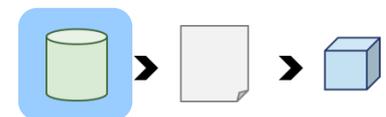
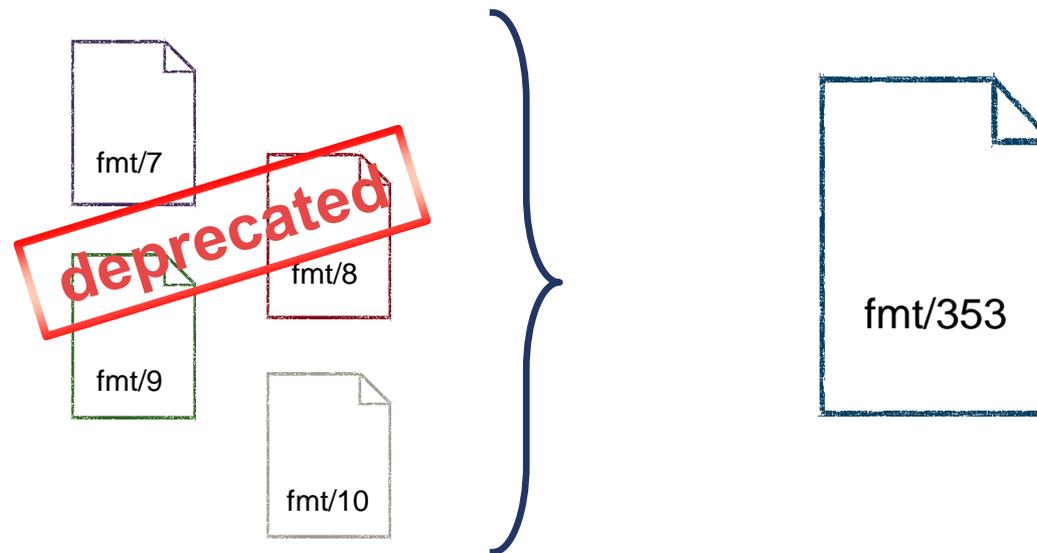
Zusammenlegung von Einträgen

Beispiel TIFF



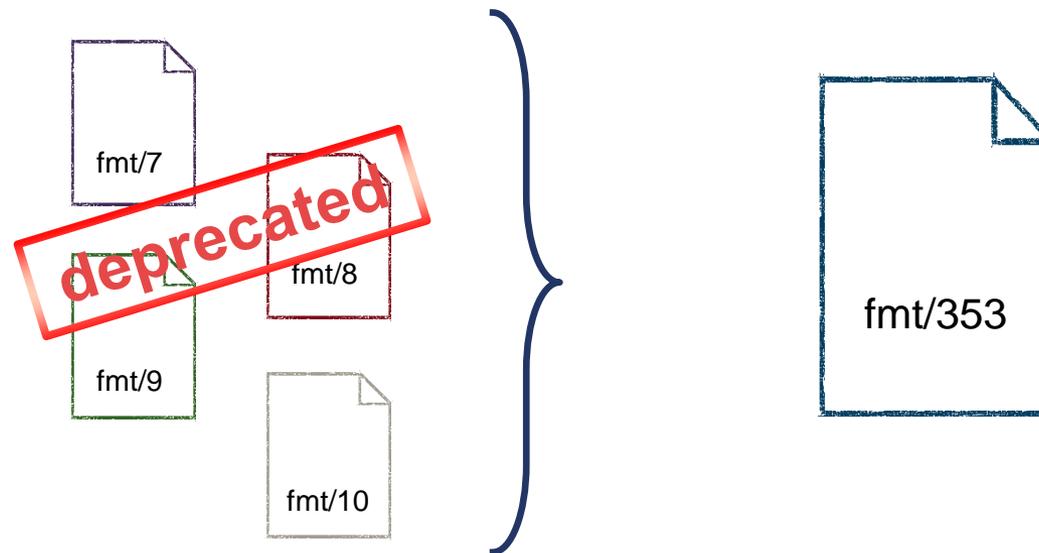
Zusammenlegung von Einträgen

Beispiel TIFF

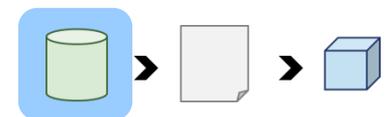


Zusammenlegung von Einträgen

Beispiel TIFF

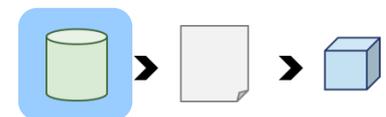


64 deprecated PUIDs
(Stand 2017)

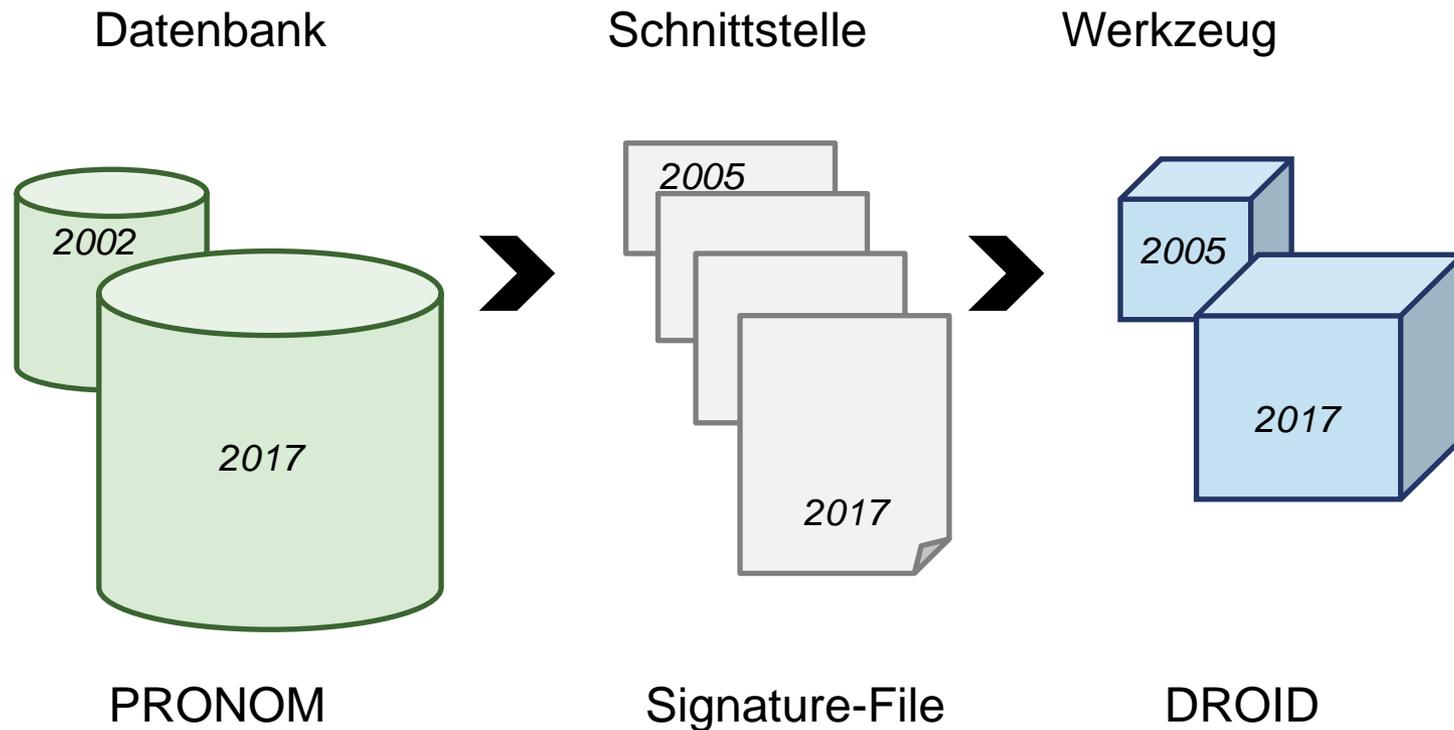




Keine PUID. Und dann?

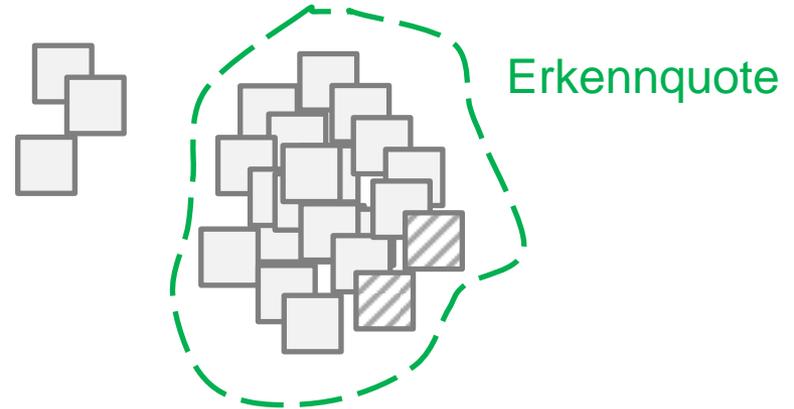


Fazit



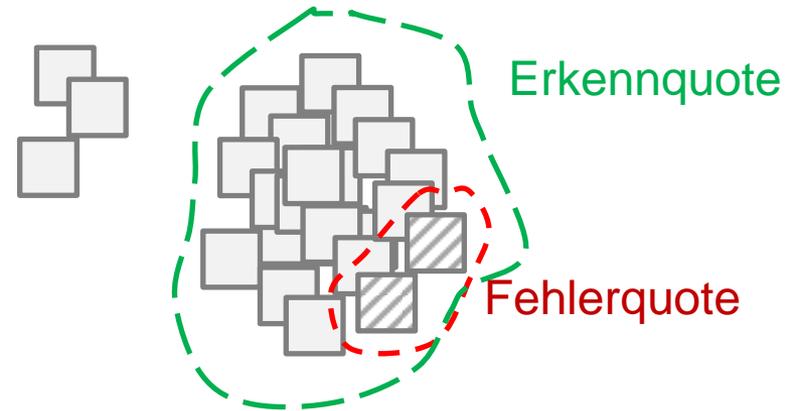
Fazit

- trotz Weiterentwicklung



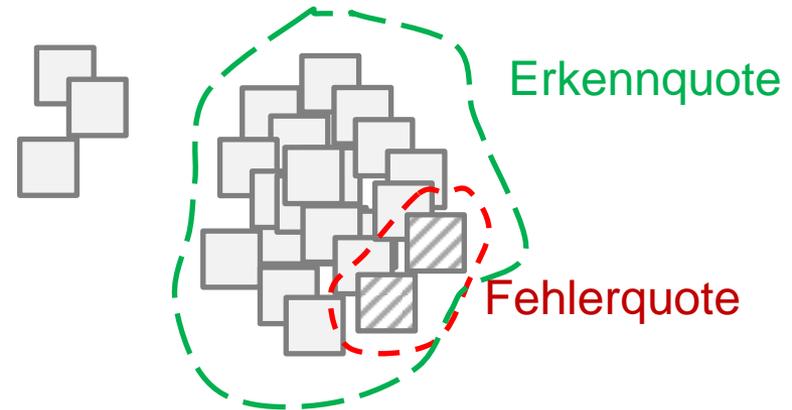
Fazit

- 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?

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