

PDF mit mangelhaftem Font: Text ist nicht durchsuch- und extrahierbar

Arbeitskreises «Archivierung von
Unterlagen aus digitalen Systemen»

23. März 2021

Disposition

- Einleitung
- Problem und Auswirkungen
- Assessment mit KOST-Val
- PPEG Empfehlung
(vermeiden, kennzeichnen und beheben)
- [Behebung]

Die KOST

Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen

- Gemeinschaftsunternehmen diverser Archive der Schweiz sowie des Landesarchivs des Fürstentums Liechtensteins.
- Sie hat den Auftrag, ihre Träger bei der Archivierung elektronischer Unterlagen zu unterstützen
 - Standards und Richtlinien
 - Tools
 - Dienstleistungen zur Lösung konkreter Probleme
 - Studien und Veranstaltungen

Die KOST

Geschäftsstelle

- Georg Büchler, Martin Kaiser, Claire Röthlisberger – Jourdan

PPEG (Preservation Planning Expert Group)

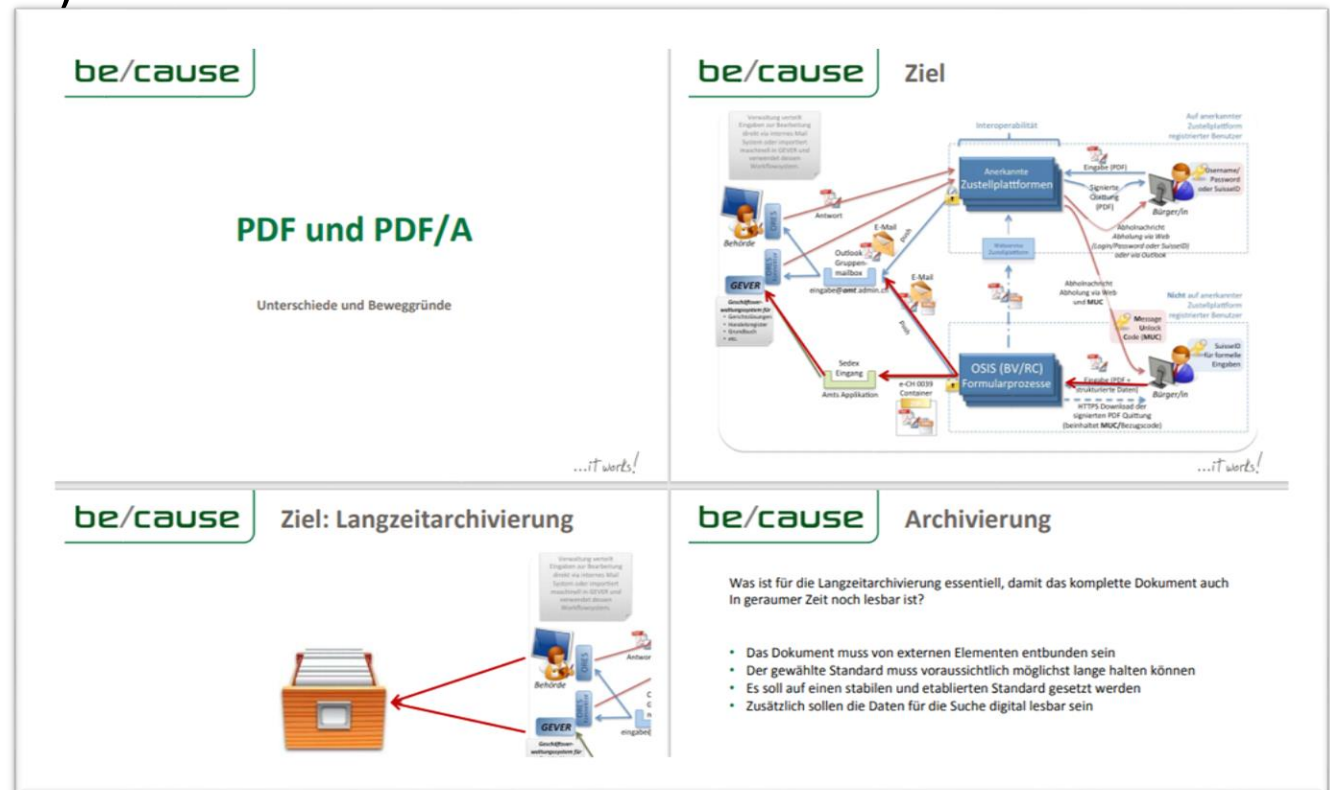
- Lambert Kansy, Staatsarchiv Basel Stadt
- Martin Lüthi, Staatsarchiv St. Gallen
- Kaya Demiroglu, Staatsarchiv Uri
- Romano Padeste, Staatsarchiv Zürich
- Pascal Schneiter, Staatsarchiv Neuenburg
- Christa Ackermann, Bundesarchiv
- Geschäftsstelle der KOST

Problem und Auswirkungen

Was sind unsere Erwartungen und Anforderungen:

[PDF&PDF/A](#) (Seite 4)

Öffnen



Problem und Auswirkungen

Was sind unsere Erwartungen und Anforderungen:

[PDF&PDF/A](#) (Seite 4)

Markieren,
kopieren &
einfügen

be/cause

Archivierung

Was ist für die Langzeitarchivierung essentiell, damit das komplette Dokument auch in geraumer Zeit noch lesbar ist?

- Das Dokument muss von externen Elementen entbunden sein
- Der gewählte Standard muss voraussichtlich möglichst lange halten können
- Es soll auf einen stabilen und etablierten Standard gesetzt werden
- Zusätzlich sollen die Daten für die Suche digital lesbar sein

...it works!

Problem und Auswirkungen

Was sind unsere Erwartungen und Anforderungen:

[PDF&PDF/A](#) (Seite 4)

Was ist für die Langzeitarchivierung essentiell, damit das komplette Dokument auch in geraumer Zeit noch lesbar ist?

- Das Dokument muss von externen Elementen entbunden sein
- Der gewählte Standard muss voraussichtlich möglichst lange halten können
- Es soll auf einen stabilen und etablierten Standard gesetzt werden
- Zusätzlich sollen die Daten für die Suche digital lesbar sein

Problem und Auswirkungen

Die Daten sollen für die Suche digital lesbar sein



- Die Daten sollen die Volltextsuche ermöglichen
- Textpassagen sollen gefunden werden

Die Daten sollen weiterverarbeitet werden können

- Die Daten sollen digital extrahierbar sein.



PDF/A-1b und PDF/A-2b

- garantieren die visuelle Reproduzierbarkeit
- garantieren nicht die Durchsuch- und Extrahierbarkeit
- ermöglichen die Durchsuch- und Extrahierbarkeit meist einwandfrei

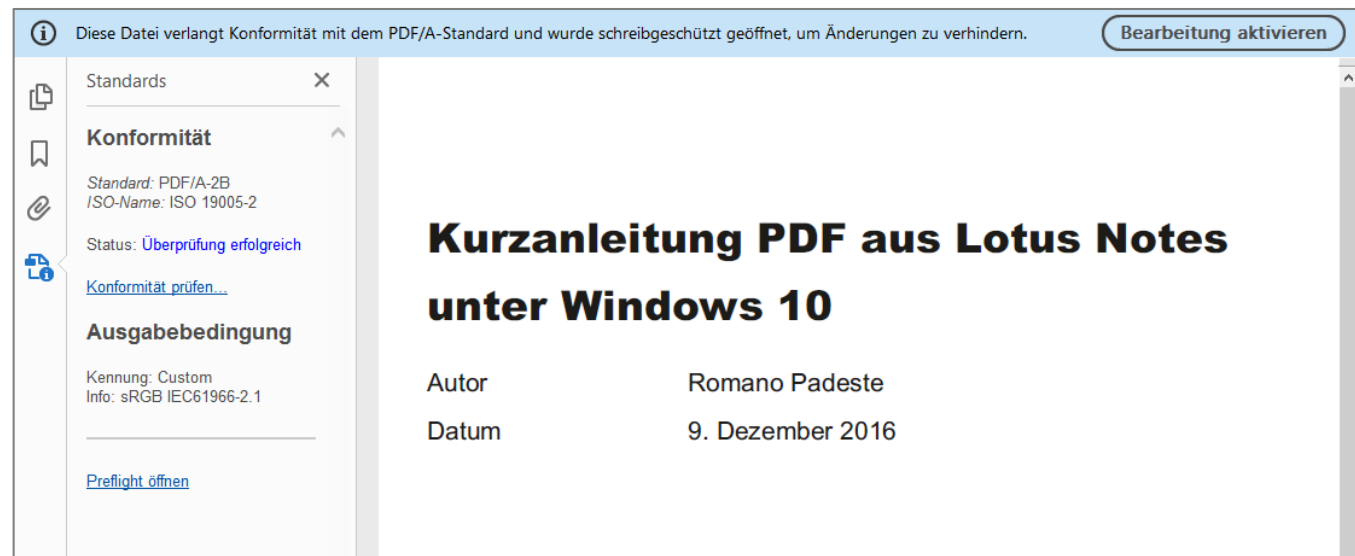


Problem und Auswirkungen

In einigen Fällen werden unsere Erwartungen und Anforderungen nicht erfüllt:

[PDF/A-2b](#) (interner Link)

Öffnen



Problem und Auswirkungen

In einigen Fällen werden unsere Erwartungen und Anforderungen nicht erfüllt:

[PDF/A-2b](#) (interner Link)

Markieren,
kopieren &
einfügen

Kurzanleitung PDF aus Lotus Notes unter Windows 10

Autor

Romano Padeste

Datum

9. Dezember 2016

Problem und Auswirkungen

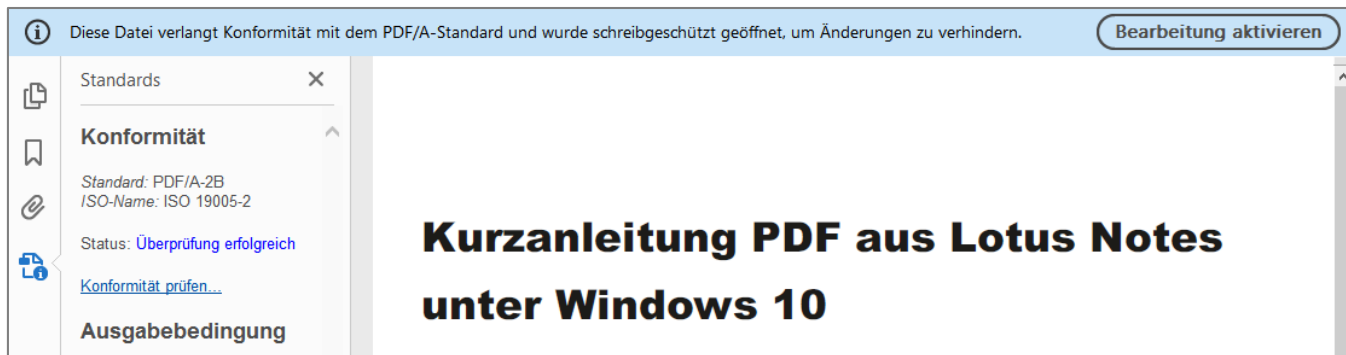
In einigen Fällen werden unsere Erwartungen und Anforderungen nicht erfüllt:

[PDF/A-2b](#) (interner Link)

?
? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
?
? ?

Problem und Auswirkungen

Dieses Dokument ist ein valides PDF/A-2b



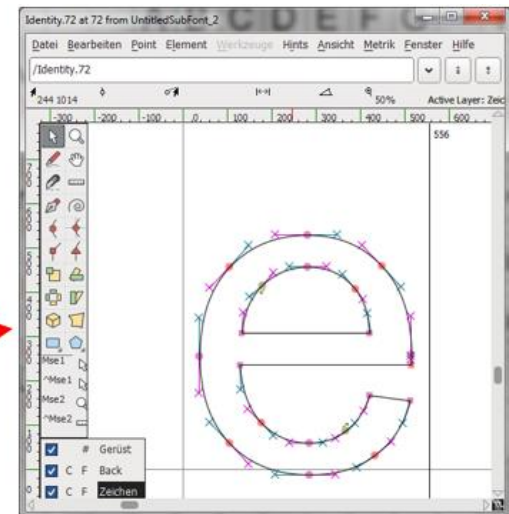
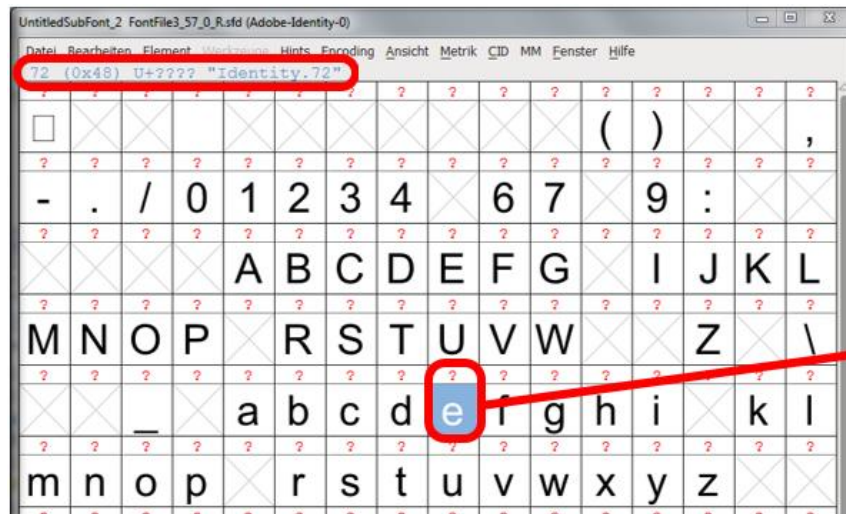
Alle getesteten Viewer konnten den Text nicht korrekt herauskopieren respektive im Dokument finden

- eingebettete Schrift erfüllt nicht die Anforderung der Archive
- Schrift verstösst nicht gegen die PDF/A-2b Anforderung

Problem und Auswirkungen

Font-Problem

- nur die Kontur der Schrift eingebettet
- digitales Abbild in der Schrift nicht bekannt / gegeben (?)
- Zuordnung zum Unicode Zeichen innerhalb der PDF-Datei mittels einer CMap «ToUnicode» nicht vorhanden



Problem und Auswirkungen

Alle PDF-Dateien mit eingebetteten Schriften können diesen Mangel enthalten.

Archive müssen damit rechnen, dass ohne Massnahmen diese PDF-Dateien bei einer Inbetriebnahme einer Volltextsuche nicht gefunden würden.

Assessment: Entwicklung

Ein Assessment mit den bestehenden Tools war nicht möglich. Die KOST beauftragt die Firma PDF Tools AG für die Entwicklung.

- neues Zusatzmodul für den «3-Heights™ PDF Validator»
- «3-Heights™ PDF Validator» ist in KOST-Val integriert
- Ausgabe von diversen Informationen:
 - Metadaten zur Datei; z.B. Hersteller und PDF-Version
 - Metadaten zur Schrift; z.B. Name und Fonttype
 - Informationen zum Zeichen (z.B. Kontur als Bild und Unicodezeichen)

Assessment: Ergebnis «strict»

Mit der Konfiguration «strict», welches dem Ergebnis des «3-Heights™ PDF Validators» entspricht, werden alle unbekannten und undefinierten Zeichen bemängelt.

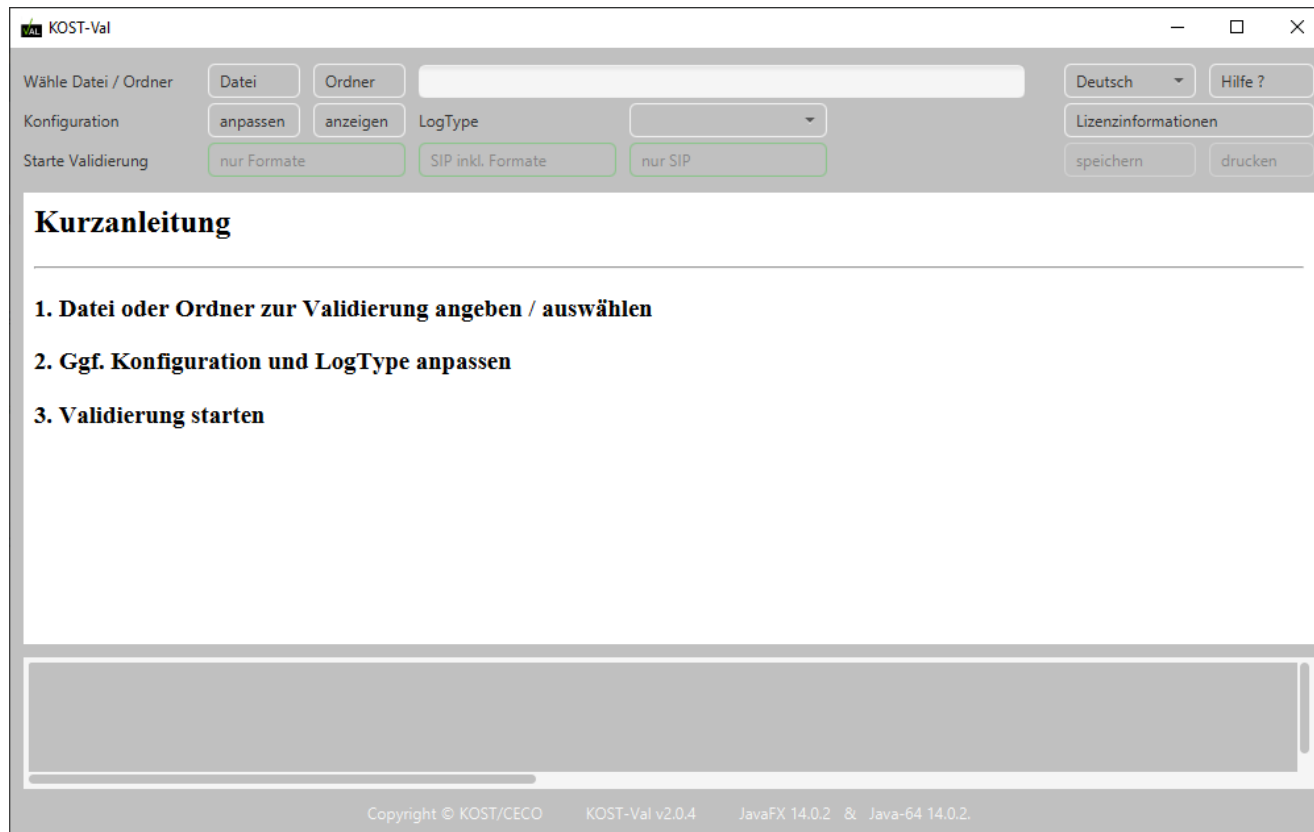
In den KOST-Archiven sind im Schnitt rund 20% aller PDF/A-1b und PDF/A-2b Dateien betroffen.

Im nachfolgenden Testset mit PDF/A-1a, -1b, -2a, -2b, -2u und normalen PDF sind rund 11% betroffen.

Assessment: Ergebnis «strict»

Screenshots KOST-Val Nr. 1:

KOST-Val 2.x:
NEUES GUI

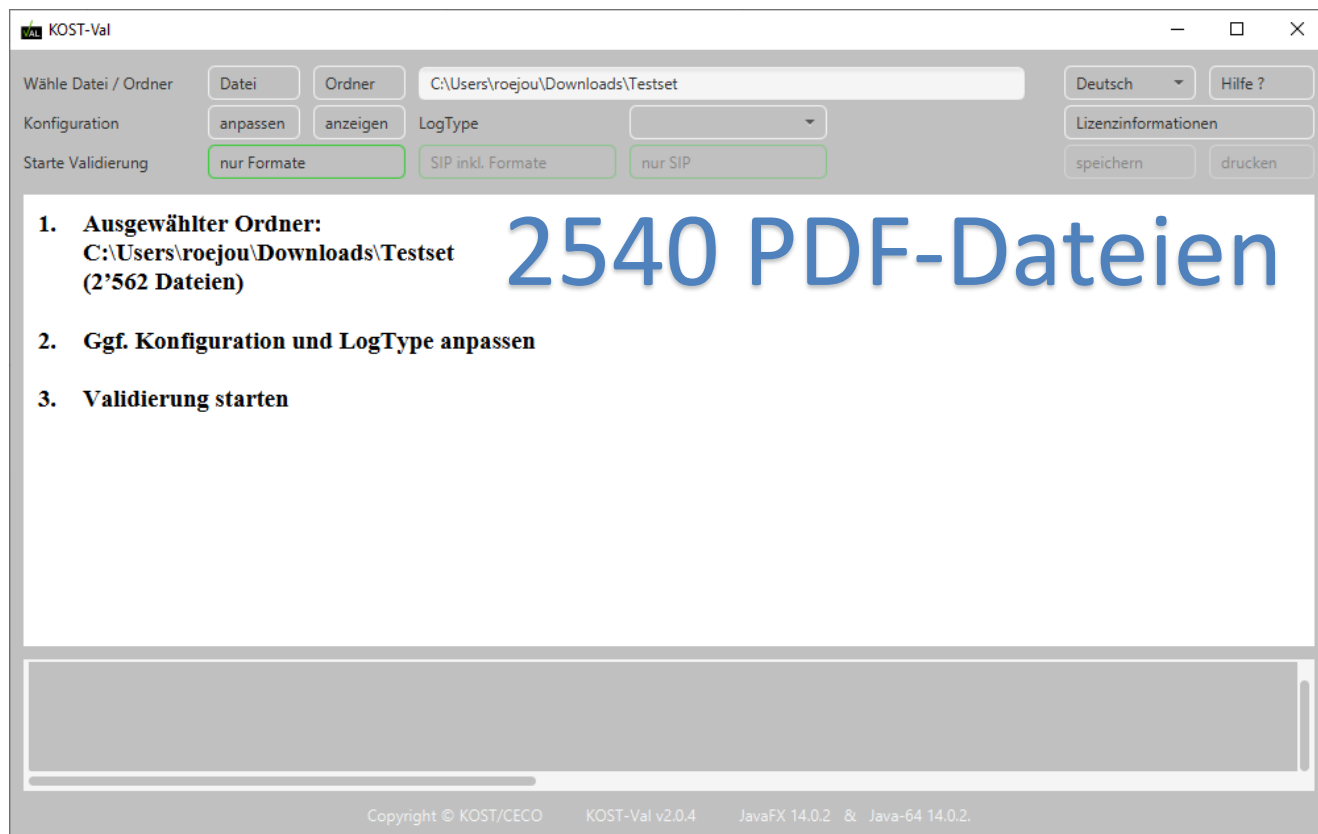


Screenshots KOST-Val Nr. 2:



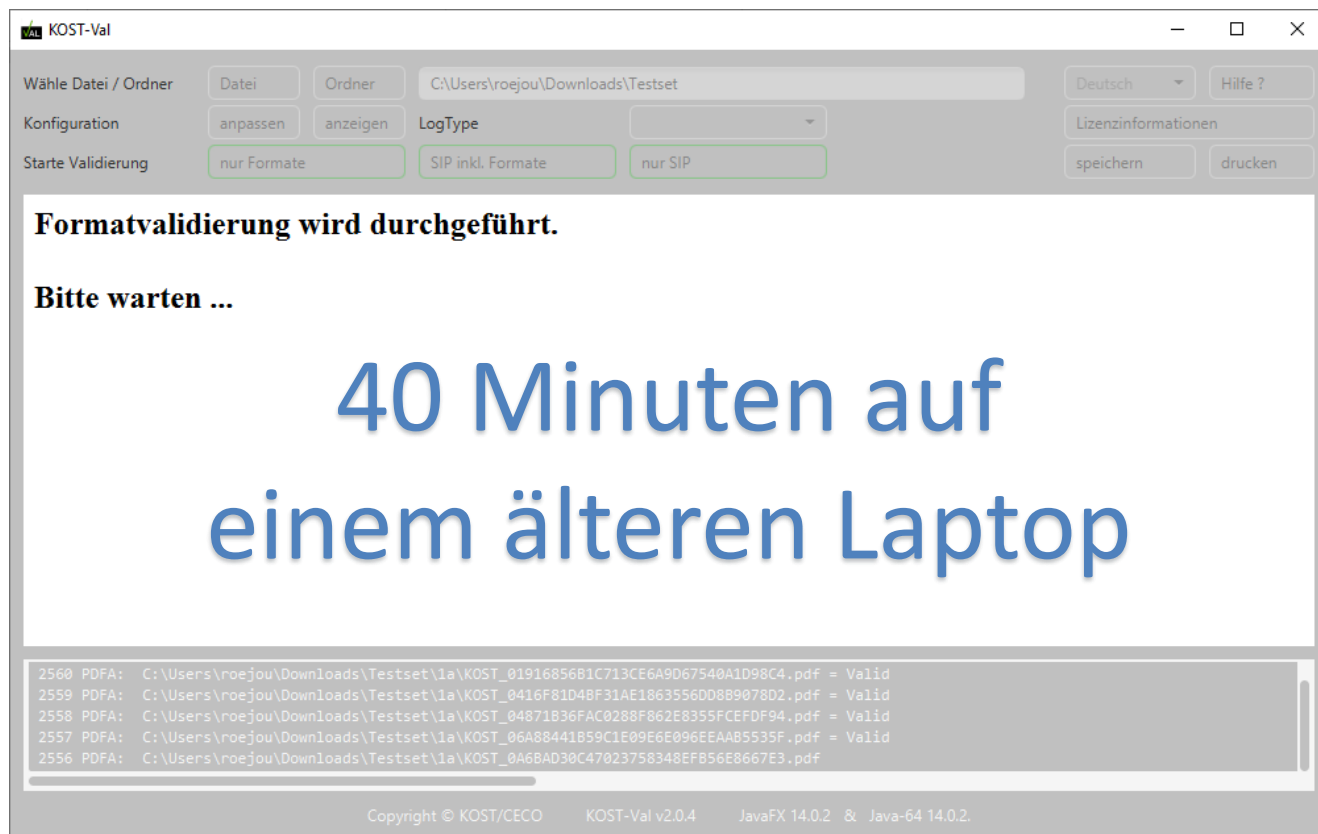
Assessment: Ergebnis «strict»

Screenshots KOST-Val Nr. 3:



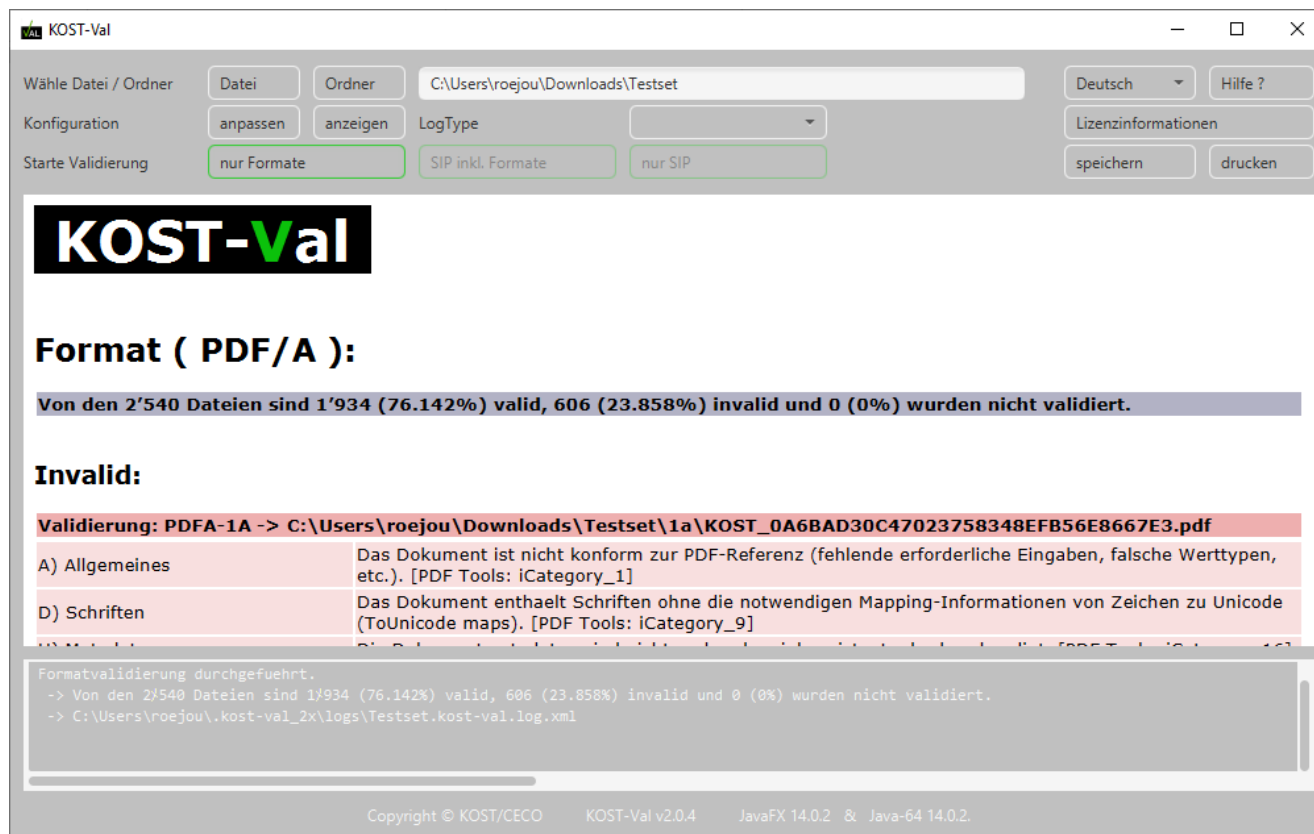
Assessment: Ergebnis «strict»

Screenshots KOST-Val Nr. 4:



Assessment: Ergebnis «strict»

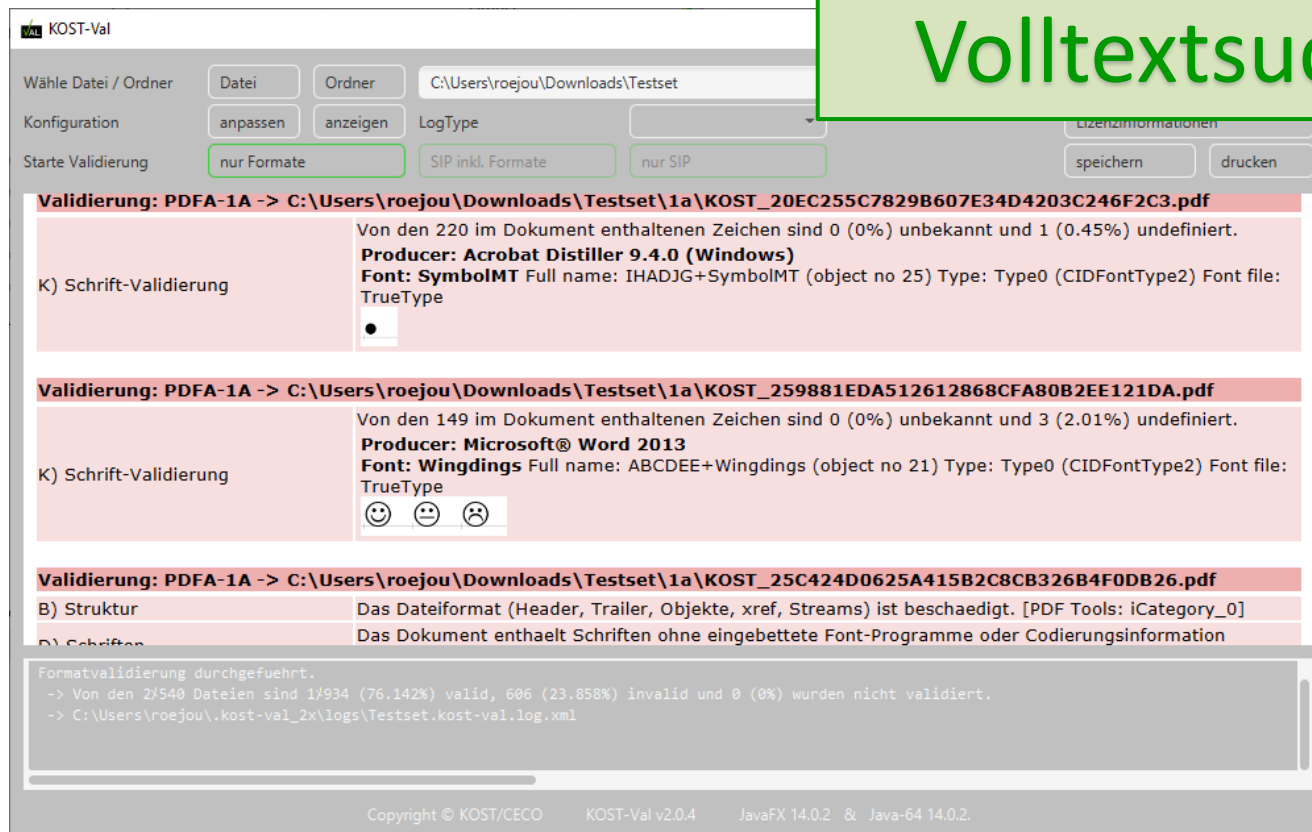
Screenshots KOST-Val Nr. 5:



Assessment: Ergebnis «strict»

Screenshots KOST-Val Nr. 6:

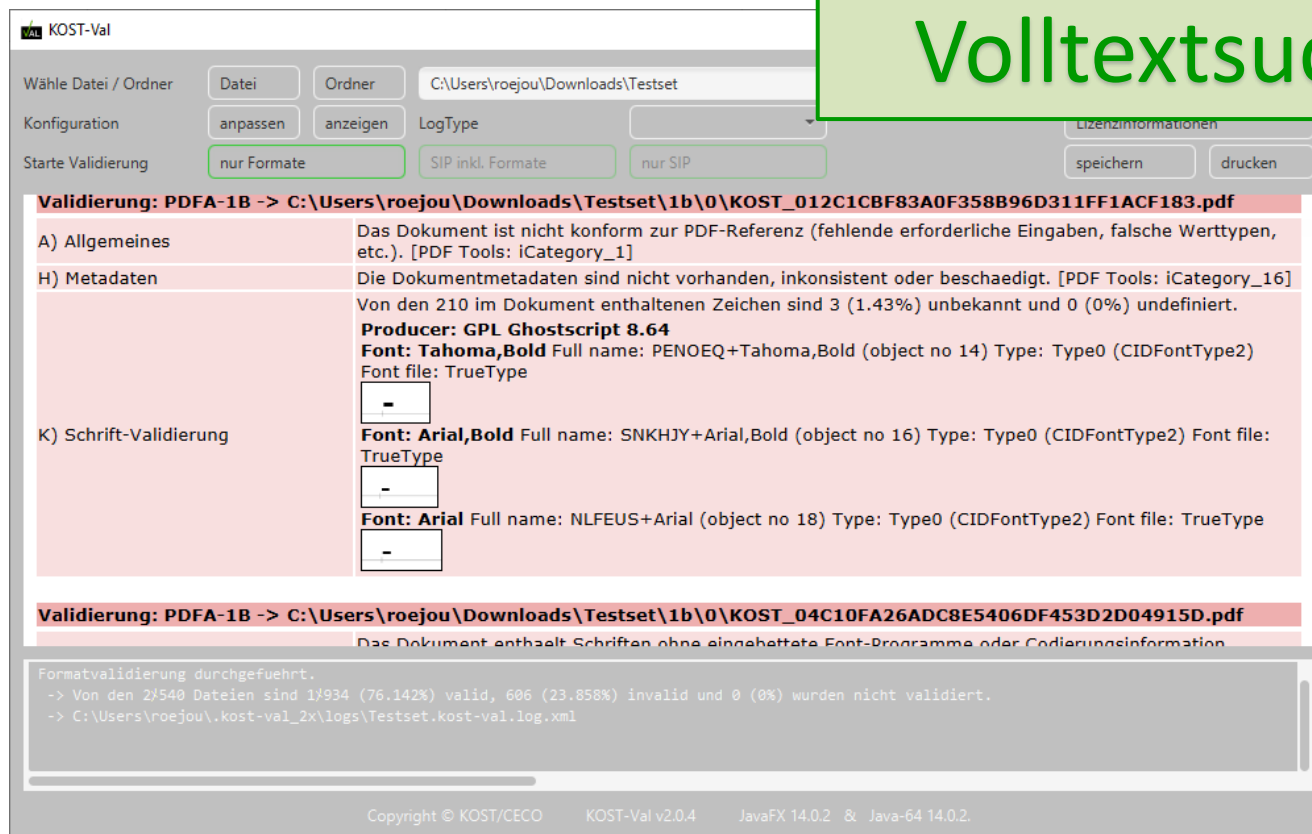
Unkritisch für die
Volltextsuche



Assessment: Ergebnis «strict»

Screenshots KOST-Val Nr. 7:

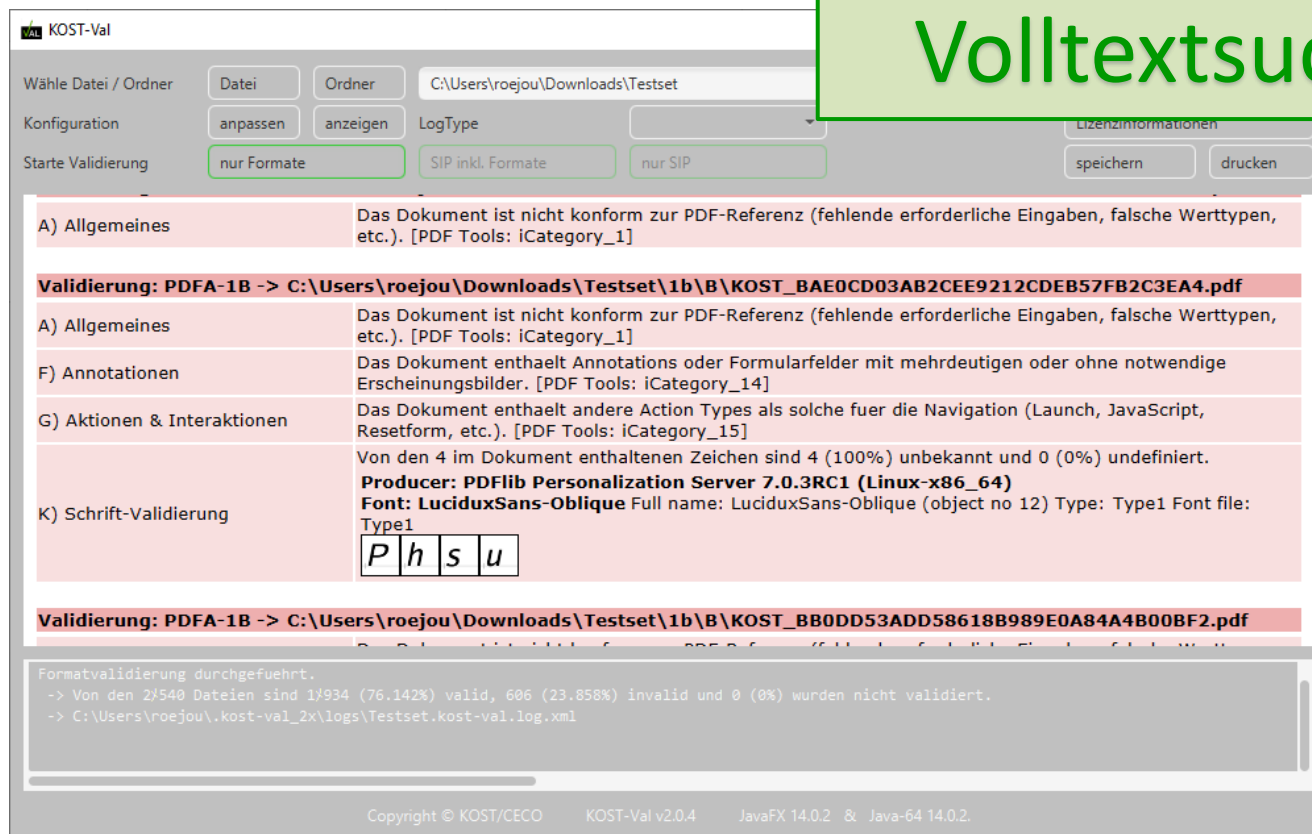
Unkritisch für die
Volltextsuche



Assessment: Ergebnis «strict»

Screenshots KOST-Val Nr. 8:

Unkritisch für die
Volltextsuche



Assessment: Ergebnis «strict»

Screenshots KOST-Val Nr. 9:

Kritisch für die
Volltextsuche

The screenshot displays the KOST-Val application window. The top section contains configuration options: 'Wähle Datei / Ordner' with a file path 'C:\Users\roejou\Downloads\Testset', 'Konfiguration' with buttons 'anpassen' and 'anzeigen', and 'LogType' with a dropdown menu. Below this are 'Starte Validierung' buttons: 'nur Formate', 'SIP inkl. Formate', and 'nur SIP'. The main area shows validation results for a PDF document. It includes sections for 'E) Transparenz', 'H) Metadaten', and 'K) Schrift-Validierung'. The 'K) Schrift-Validierung' section displays two font validation tables. The first table shows the font 'font00000000196ed131' (Mac OS X 10.9.4 Quartz PDFContext) and lists characters from ' ' to '9'. The second table shows the font 'font00000000196ed131' (GWXWPR+font00000000196ed131) and lists characters from ' ' to 'l'. At the bottom, a status bar indicates 'Formatvalidierung durchgeführt.' and provides statistics: 'Von den 2/540 Dateien sind 1/934 (76.142%) valid, 606 (23.858%) invalid und 0 (0%) wurden nicht validiert.' The footer contains copyright information: 'Copyright © KOST/CECO KOST-Val v2.0.4 JavaFX 14.0.2 & Java-64 14.0.2.'

Wähle Datei / Ordner: Datei Ordner C:\Users\roejou\Downloads\Testset

Konfiguration: anpassen anzeigen LogType

Starte Validierung: nur Formate SIP inkl. Formate nur SIP

Lizenzinformationen: speichern drucken

E) Transparenz: Das Dokument enthält Transparenz. [PDF Tools: iCategory_10]

H) Metadaten: Die Dokumentmetadaten sind nicht vorhanden, inkonsistent oder beschädigt. [PDF Tools: iCategory_16]

Von den 301 im Dokument enthaltenen Zeichen sind 90 (29.9%) unbekannt und 0 (0%) undefiniert.

Producer: Mac OS X 10.9.4 Quartz PDFContext

Font: font00000000196ed131 Full name: GYHNOU+font00000000196ed131 (object no 935) Type: Type0 (CIDFontType2) Font file: TrueType

'	(	)	,	-	.	/	0	1	2	3	4	5	6	7	8	9	
:	A	B	C	D	E	F	G	H	I	L	N	O	P	S	a	b	c
d	e	f	g	h	i	l	m	n	o	p	q	r	s	t	u	v	y
à	ç	é	è														

K) Schrift-Validierung

Font: font00000000196ed131 Full name: GWXWPR+font00000000196ed131 (object no 337) Type: Type0 (CIDFontType2) Font file: TrueType

'	,	?	E	I	L	P	a	b	c	d	e	f	g	h	i	l	
m	n	o	q	r	s	t	u	v	x	y	à	é	è				

Formatvalidierung durchgeführt.

-> Von den 2/540 Dateien sind 1/934 (76.142%) valid, 606 (23.858%) invalid und 0 (0%) wurden nicht validiert.

-> C:\Users\roejou\.kost-val_2x\logs\Testset.kost-val.log.xml

Copyright © KOST/CECO KOST-Val v2.0.4 JavaFX 14.0.2 & Java-64 14.0.2.

Assessment: Ergebnis «strict»

Screenshots KOST-Val Nr. 10:

Kritisch für die
Volltextsuche

The screenshot displays the KOST-Val application window. The top section contains controls for file selection (Datei, Ordner), configuration (anpassen, anzeigen, LogType), and validation options (Starte Validierung: nur Formate, SIP inkl. Formate, nur SIP). The main area shows the validation results for the file **C:\Users\roejou\Downloads\Testset\2b\KOST_0D2CD11DA306E1077E18D963BE36D42B.pdf**.

Validierung: PDF/A-2B -> C:\Users\roejou\Downloads\Testset\2b\KOST_0D2CD11DA306E1077E18D963BE36D42B.pdf

A) Allgemeines
Das Dokument ist nicht konform zur PDF-Referenz (fehlende erforderliche Eingaben, falsche Werttypen, etc.). [PDF Tools: iCategory_1]
Von den 289 im Dokument enthaltenen Zeichen sind 104 (35.99%) unbekannt und 0 (0%) undefiniert.
Producer: PDF Writer - bioPDF / <http://www.biopdf.com> / User GS EFD
Font: PalatinoLinotype-Roman Full name: PalatinoLinotype-Roman (object no 35) Type: Type1 Font file: CFF
tt
Font: FrutigerLTStd-ExtraBlackCn Full name: LEQGX+FrutigerLTStd-ExtraBlackCn (object no 36) Type: Type0 (CIDFontType0) Font file: CFF
- . : A B C D E F G H I J L M N O P R S T U V X Z
' A E I O
Font: FrutigerLTStd-BlackCn Full name: RKZMWZ+FrutigerLTStd-BlackCn (object no 37) Type: Type0 (CIDFontType0) Font file: CFF
K) Schrift-Validierung
() , - . 0 1 2 3 4 5 7 8 9 : A B C D E F G H I J
L M N O P R S T U V W X Z ' a b c d e a h i k l m n

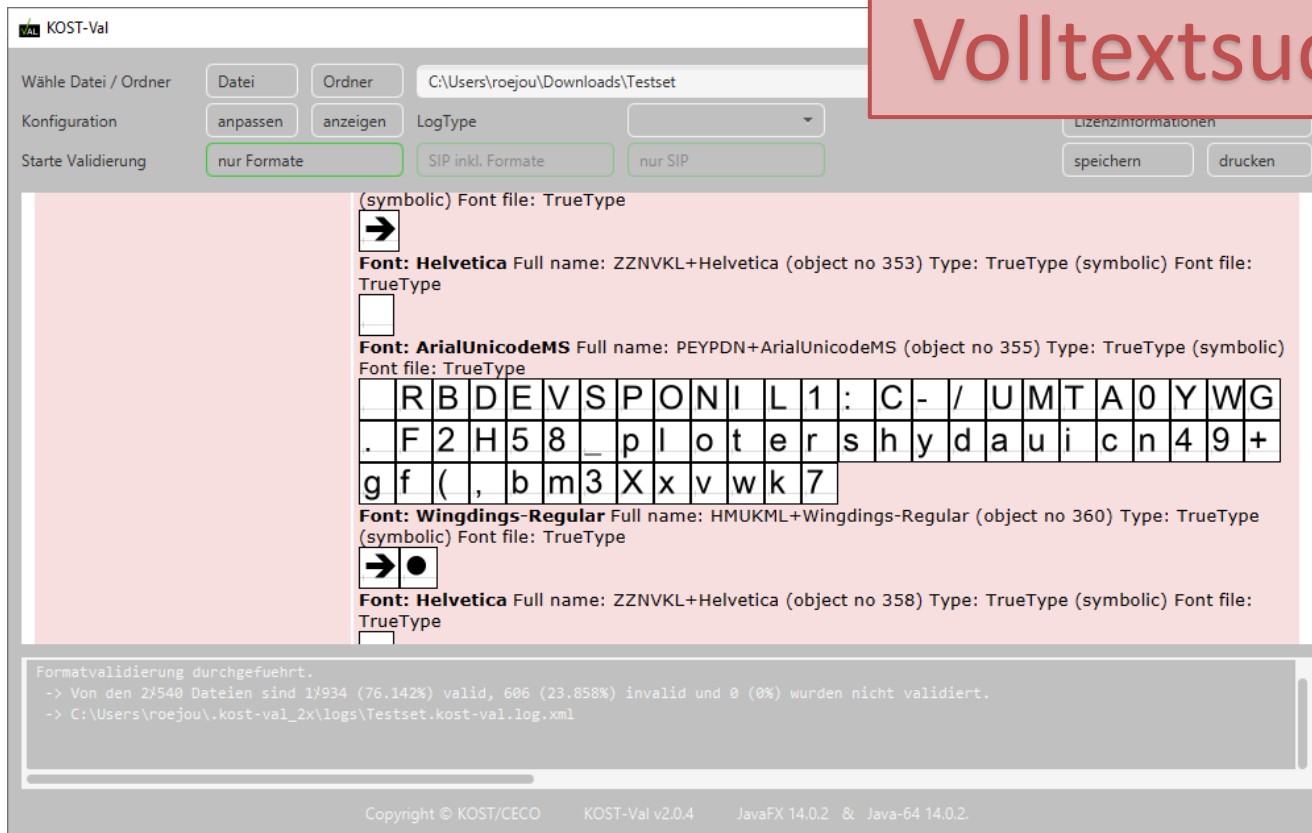
Formatvalidierung durchgeführt.
-> Von den 24540 Dateien sind 17934 (76.142%) valid, 606 (23.858%) invalid und 0 (0%) wurden nicht validiert.
-> C:\Users\roejou\kост-val_2x\logs\Testset.kost-val.log.xml

Copyright © KOST/CECO KOST-Val v2.0.4 JavaFX 14.0.2 & Java-64 14.0.2

Assessment: Ergebnis «strict»

Screenshots KOST-Val Nr. 11:

Kritisch für die Volltextsuche



Assessment: Nachauswertung

Bedarf an einer Nachauswertung «tolerant», bei der unkritische Mängel ignoriert werden

- maximal 5 Zeichen eingebettet
- Webdings-, Wingdings-, Symbol- oder Math-Schriften
- alle undefinierten Zeichen, sofern $\leq 20\%$
- bekannte Zeichen (Abstände, Aufzählungen und Auslassungspunkte)
- wenn nur noch maximal 4 Zeichen bemängelt würden

Nachauswertung erfolgt in KOST-Val

Durch die KOST definiert und entwickelt

Assessment: Ergebnis «tolerant»

Mit der Konfiguration «tolerant» wird die Nachauswertung durchgeführt.

In den KOST-Archiven sind im Schnitt rund 1% aller PDF/A-1b und PDF/A-2b Dateien betroffen.

Diese können bei einer Volltextsuche kritisch sein.

Im gezeigten Testset werden mit tolerant rund 1% weiterhin bemängelt. → 12% der Strikten

Assessment: Ergebnis «tolerant»

Screenshots KOST-Val Nr. 12:

KOST-Val - Configuration

☐ JPEG

☐ JPEG2000

☐ PNG

☐ SIARD ☒ SIARD-1.0 (eCH-0165 v1) ☒ SIARD-2.1

☒ **PDF/A**

PDF Tools: ☒ PDF Tools, ☐ details, ☒ Font, ☒ Tolerant

Callas: ☐ Callas, ☐ N-Entry

Versions: ☒ PDF/A-1a, ☒ PDF/A-1b, ☒ PDF/A-2a, ☒ PDF/A-2b, ☒ PDF/A-2u

Sonstiges: ☒ JBIG2

☐ **TIFF**

Komprimierungsalgorithmus: ☒ Uncompressed, ☒ CCITT 1D, ☒ T4/Group 3 Fax, ☒ T6/Group 4 Fax, ☒ LZW, ☐ JPEG, ☐ Deflate, ☒ PackBits

Farbraum: ☒ WhiteIsZero, ☒ BlackIsZero, ☒ RGB, ☒ RGB Palette, ☐ transparency, ☐ CMYK, ☐ YCbCr, ☐ CIE L*a*b*

Bits per Sample (pro Kanal): ☒ Bps 1, ☐ Bps 2, ☒ Bps 4, ☒ Bps 8, ☒ Bps 16, ☐ Bps 32

Sonstiges: ☒ Multipage, ☐ Tiles, ☐ Size

☐ **eCH-0160**

Pfadlänge: 179

SIP Name: SIP_[1-2][0-9]{3}[0-1][0-9]{3}[0-9]{3}

PUID: TXT PDFa1 PDFa2 TIFF JP2 JPEG PNG WAVE FLAC MP3 MP4 MJ2 MPEG2 CSV SIARD WARC

Arbeitsverzeichnis:

Inputverzeichnis:

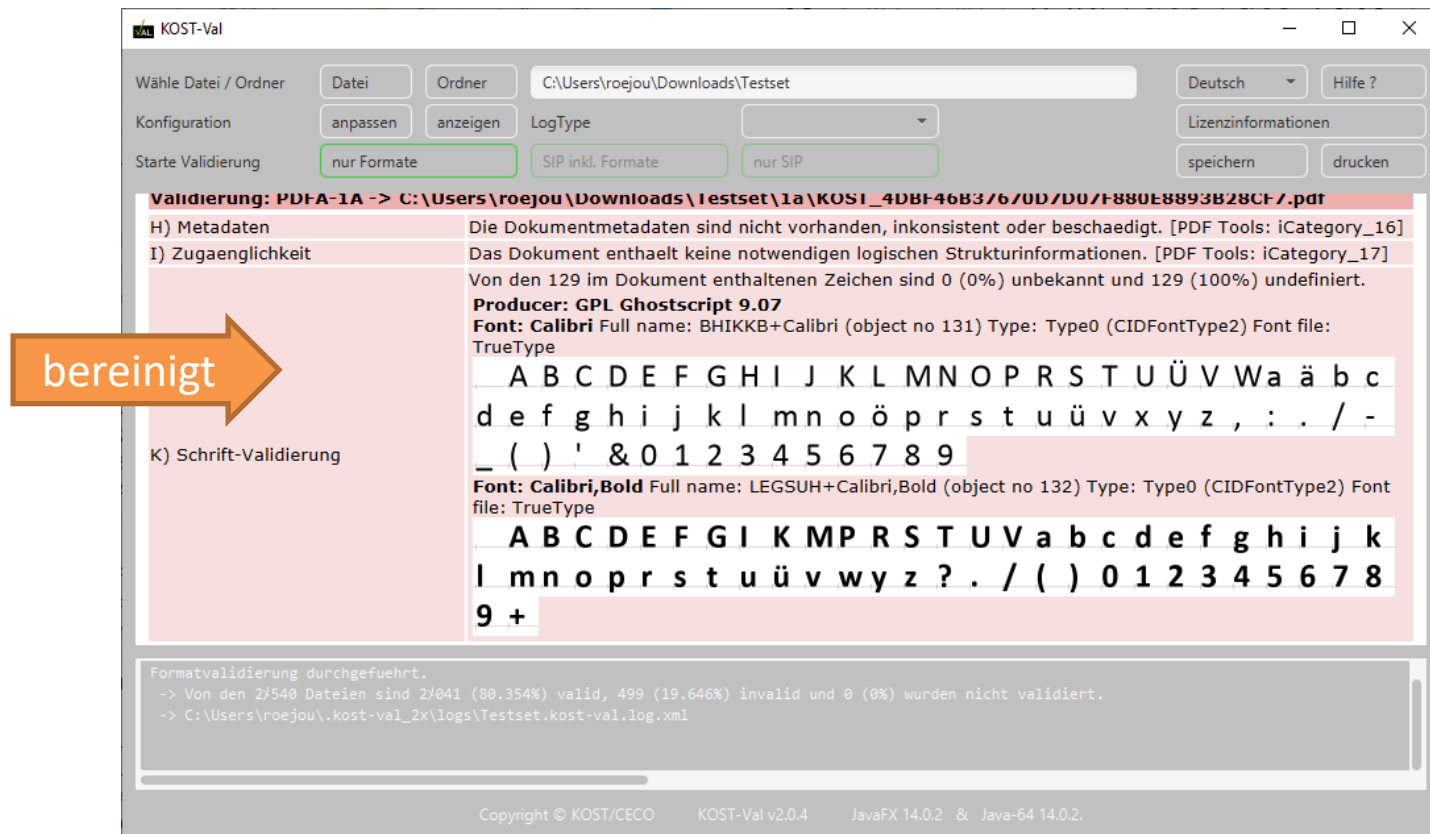
Copyright © KOST/CECO KOST-Val v2.0.4 JavaFX 14.0.2 & Java-64 14.0.2

KOST-Val Konfiguration

Angabe, ob eine JPEG-Validierung stattfinden soll [yes]:	no
Angabe, ob eine JP2-Validierung stattfinden soll [yes]:	no
Angabe, ob eine PNG-Validierung stattfinden soll [yes]:	no
Angabe, ob eine SIARD-Validierung stattfinden soll [yes]:	no
Angabe, welche SIARD-Versionen erlaubt sind [1.0, 2.1]:	1.0 2.1
Angaben zur PDF/A-Validierung [yes]:	yes
Angabe, ob eine PDF/A-Validierung mit PDF Tools stattfinden soll [yes]:	yes
Angabe, ob seitens PDF Tools auch detaillierte Fehler in Englisch ausgegeben werden sollen [yes]:	no
Angabe, ob eine Schrift-Validierung (Durchsuch- und Extrahierbarkeit) stattfinden soll [tolerant]:	tolerant
Angabe, ob eine PDF/A-Validierung mit callas stattfinden soll [yes]:	no
Angabe, ob seitens callas ein Fehler (E) oder eine Warnung (W) ausgegeben werden soll, wenn der N-Eintrag nicht übereinstimmt [W]:	W
Angabe, welche PDF/A Versionen erlaubt sind [1A, 1B, 2A, 2B, 2U]:	1A 1B 2A 2B 2U
Angabe, ob die JBIG2-Komprimierung erlaubt ist [yes]:	yes
Angabe, ob eine TIFF-Validierung stattfinden soll [yes]:	no
Angabe, der erlaubten Komprimierungsalgorithmen [Uncompressed, CCITT 1D, T4/Group 3 Fax, T6/Group 4 Fax, LZW, PackBits]:	Uncompressed CCITT 1D T4/Group 3 Fax T6/Group 4 Fax LZW PackBits
Angabe, der erlaubten Farbräume [WhiteIsZero, BlackIsZero, RGB, RGB Palette]:	WhiteIsZero BlackIsZero RGB RGB Palette
Angabe, der erlaubten Bits per Sample [1, 4, 8, 16]:	1 4 8 16
Angabe, ob Multipage-TIFFs erlaubt sind oder nicht [yes]:	yes
Angabe, ob der Aufbau in Kacheln erlaubt ist oder nicht [no]:	no
Angabe, ob Dateigrößen von 1000MB (~1GB) und grösser erlaubt sind oder nicht [no]:	no
Angaben zur SIP-Validierung [yes]:	no
Erlaubte maximale Anzahl Zeichen in Pfadlängen [179]:	179
Vorgaben zum Aufbau des SIP-Namens [SIP_[1-2][0-9]{3}[0-1][0-9]{3}[0-9]{3}]:	SIP_[1-2][0-9]{3}[0-1][0-9]{3}[0-9]{3}
Auflistung der erlaubten Dateiformate [TXT, PDFa1, PDFa2, TIFF, JP2, JPEG, PNG, WAVE, MP3, MP4, MJ2, CSV, SIARD, WARC]:	TXT PDFa1 PDFa2 TIFF JP2 JPEG PNG WAVE FLAC MP3 MP4 MJ2 MPEG2 CSV SIARD WARC

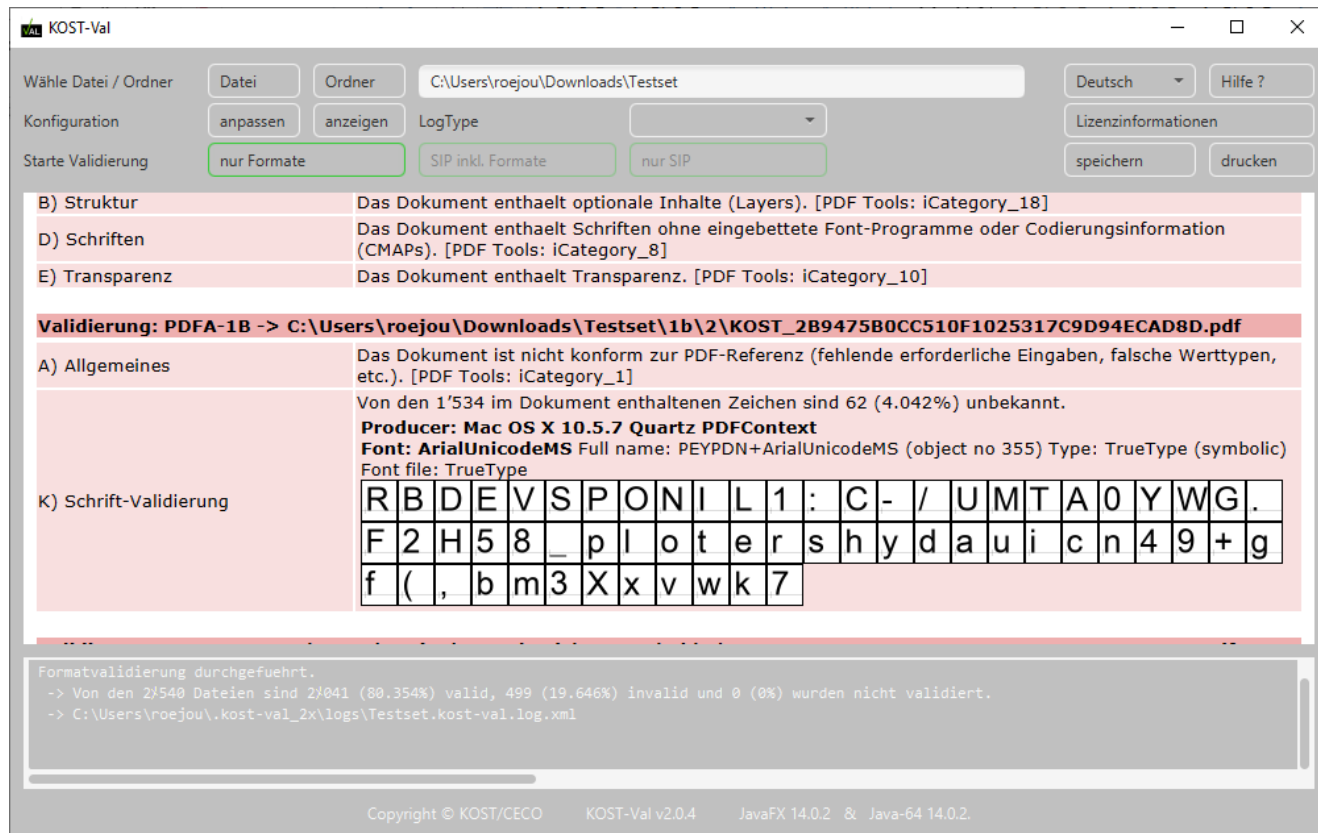
Assessment: Ergebnis «tolerant»

Screenshots KOST-Val Nr. 13:



Assessment: Ergebnis «tolerant»

Screenshots KOST-Val Nr. 14:



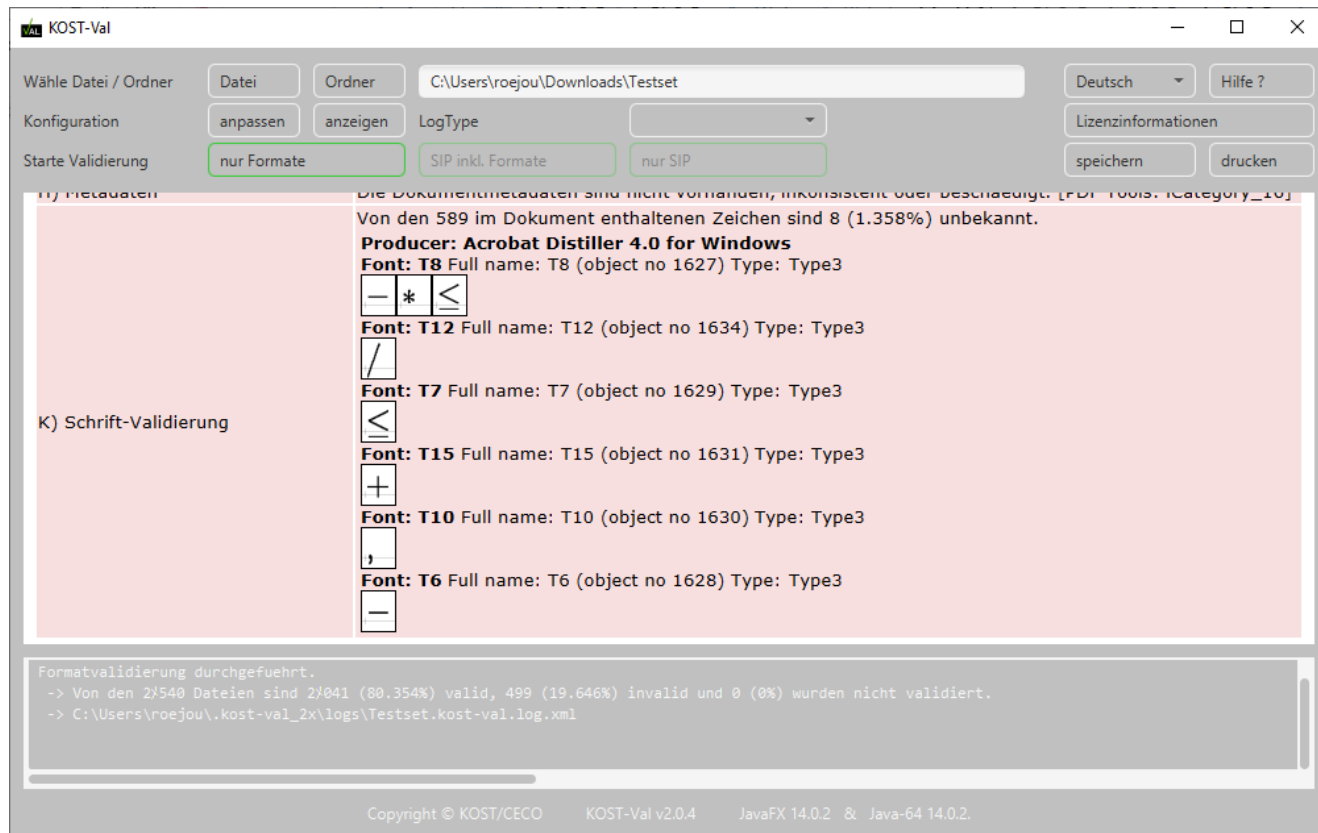
Assessment: Ergebnis «tolerant»

Screenshots KOST-Val Nr. 15:

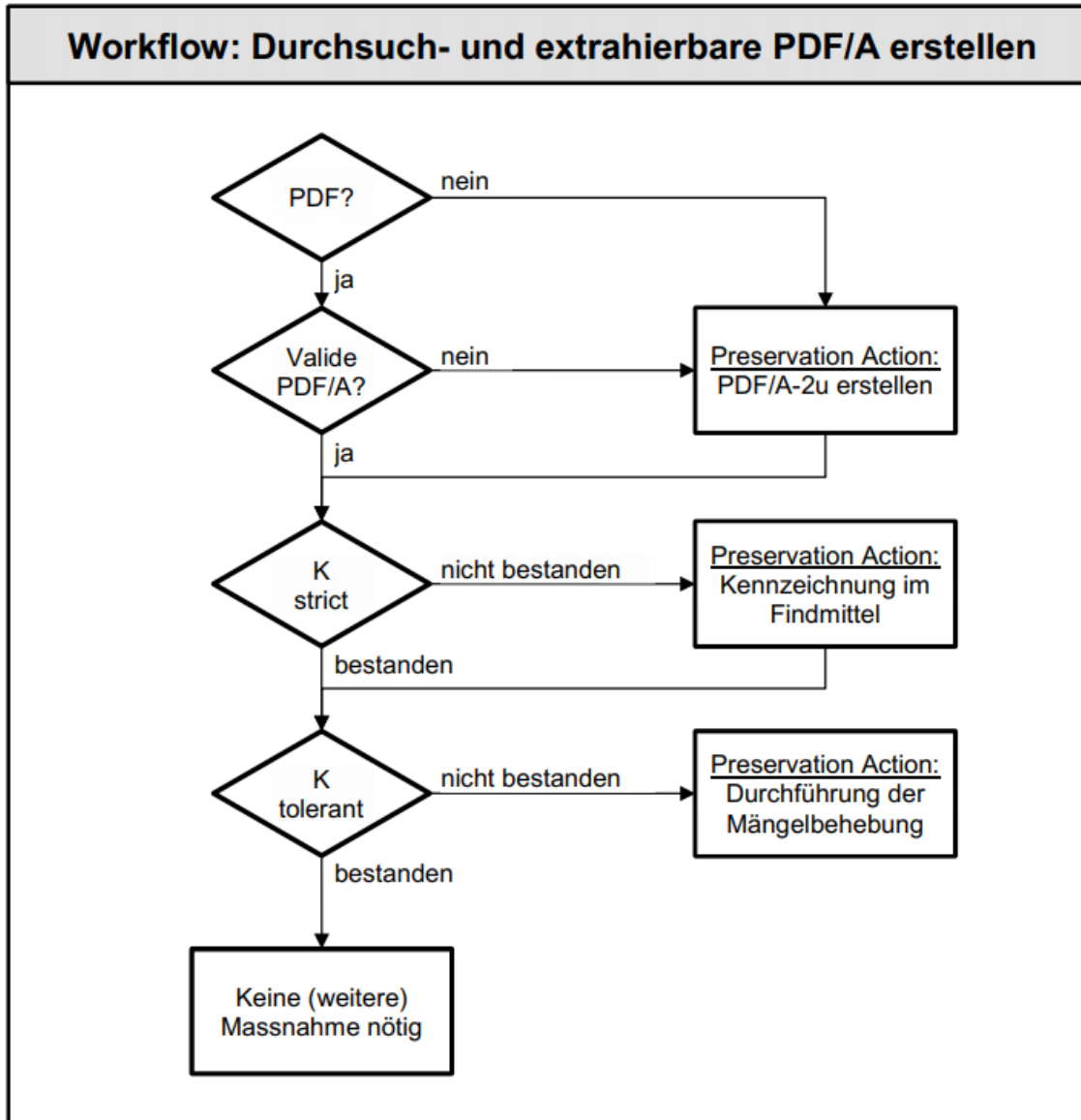


Assessment: Ergebnis «tolerant»

Screenshots KOST-Val Nr. 16:



PPEG Empfehlung



Vermeiden

Kennzeichnen

Beheben

PPEG Empfehlung

vermeiden

- Konvertierungen mit qualitativ hochwertige Software erzeugen und mindestens PDF/A-2u verlangen
- Per sofort

im Findmittel kennzeichnen

- strict-Assessment nicht bestanden
- «Datei nicht vollständig durchsuch- und extrahierbar»
- Schnellst möglich kennzeichnen

PPEG Empfehlung

beheben in einer neuen Representation

- tolerant-Assessment nicht bestanden
- vor der Einführung der Volltextsuche
- Mögliches Behebungstool (OCR):
 - «3-Heights™ PDF OCR» von PDF Tools AG mit «FineReader» Lizenz
 - «FineReader» von ABBYY

Behebung

Benötigte Installationen und Lizenzen:

- 3-Heights™ PDF OCR Service → v. 6.3.1.5
- 3-Heights™ PDF OCR Shell → v. 6.3.1.5
- FineReaderEngine → v. 12

Behebung



Behebung

Behebung mit 3-Heights™ PDF OCR:

```
pdfocr -v -ocr "service"
```

```
-ocp "Profile=C:\TEMP\ocr_profile.txt"
```

→ Text Ausrichtung

```
-ocl "English,German,French,Mathematical"
```

→ Sprache Wörterlisten

```
-otm update
```

→ alle Zeichen ohne sinnvollen Unicode-Wert beheben

```
-ots "knownSymbolic"
```

→ Symbolschriftarten werden übersprungen

```
-otu "installedFont,
```

→ installierte Schriften verwenden

```
fallbackAllPua"
```

→ bei Problemen PUA-Unicode verwenden

```
C:\TEMP\Behebung_Demo_todo.pdf
```

→ Input PDF-Datei

```
C:\TEMP\Behebung_Demo_done.pdf
```

→ Output PDF-Datei

Behebung



FAZIT

Preservation Action wird ausgelöst durch die veränderte Anforderung der Archive.

➤ Kein grundsätzliches Problem von PDF

Besten Dank

Weitere Infos inkl. Videos sind hier ersichtlich:

<https://kost-ceco.ch/cms/pdf-nicht-durchsuchbar.html>

KOST Koordinationsstelle für die dauerhafte Archivierung
elektronischer Unterlagen

Ein Gemeinschaftsunternehmen von Schweizer Archiven

Claire Röthlisberger-Jourdan

claire.roethlisberger@kost.admin.ch