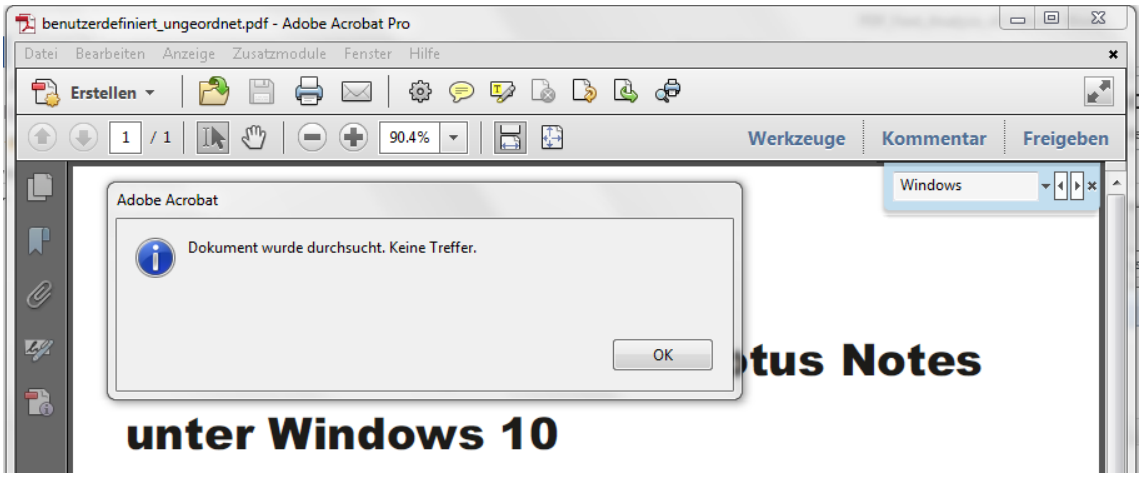
PDF mit mangelhaftem Font: Text ist nicht durchsuch- und extrahierbar

PPEG: Report & Recommendations

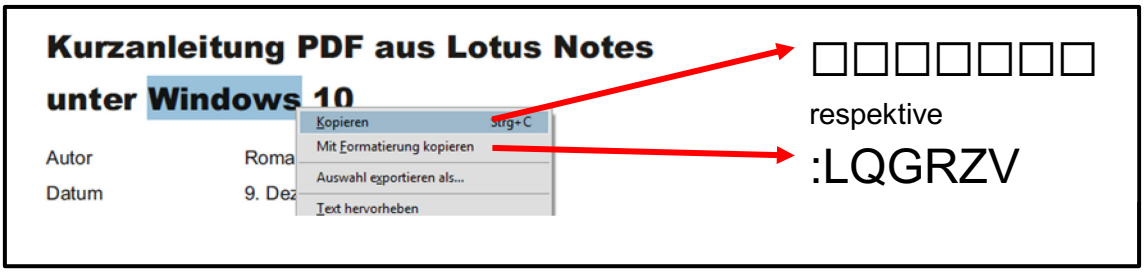
1 Incident

1.1 Beschreibung

Das Staatsarchiv Zürich stellte nach einer Systemumstellung fest, dass das neuerstellte PDF/A-Dokument nicht mehr durchsuchbar ist.



Beim Kopieren von Textpassagen wurden zudem nur noch „Unbekanntes Zeichen“ oder falsche Zeichen ausgegeben.



1.2 Auswirkungen

Die PDF-Dateien können zwar noch gelesen und ausgedruckt, jedoch höchstwahrscheinlich in Zukunft nicht mehr gefunden werden, da die Volltextsuche über diesen kryptischen Text nicht funktioniert. Problematisch ist zudem, dass dieser Fehler nicht auf den ersten Blick ersichtlich ist.

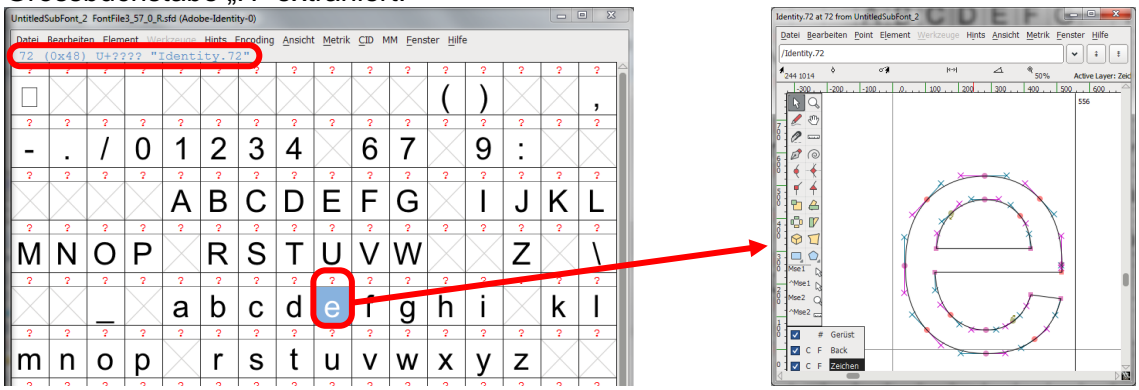
2 Analysis

2.1 Incident Dokument

Beim vorliegenden Dokument handelt es sich um ein PDF/A-2b, welches nicht nur von Adobe-Preflight sondern auch von den Validatoren von PDF Tools, Callas und PDFTron als valide befunden werden. Alle getesteten Viewer konnten den Text nicht korrekt herauskopieren beziehungsweise im Dokument finden.

Im Dokument wurde nur die Schriftart Arial verwendet, welche in Untergruppen eingebettet ist.

Die Analyse hat ergeben, dass die eingebetteten Schriften die Durchsuch- und Extrahierbarkeit nicht gewährleisten. Mit „Adobe Acrobat Pro X 116“ wurde das gesamte PDF nach Word exportiert. Dabei stellte sich heraus, dass jede der eingebetteten Arial-Schriften betroffen ist. Der Buchstabe „e“ wird immer als Grossbuchstabe „H“ extrahiert.



Dies ist auch ersichtlich, wenn die eingebettete Schrift mit FontForge geöffnet wird. Es ist danach sichtbar, dass die Zuordnung zum Unicode Zeichen nicht in der Schrift definiert ist (?). Da die Konturen aber definiert sind, gibt es keine Probleme bei der Darstellung am Bildschirm und beim Drucken. Dies hat zur Folge, dass die visuelle Reproduzierbarkeit erfüllt ist und keinen Fehler in der PDF/A-2b Validierung ergibt. Da auch die Zuordnung zum Unicode Zeichen innerhalb der PDF-Datei mittels einer CMap «ToUnicode» nicht vorhanden ist, wird bei der Kontur «e» der Standard-Wert für Element 72 (U+0048 = H) beim „Kopieren mit Formatierung“ angenommen und ausgegeben. Die Durchsuch- und Extrahierbarkeit ist einzig bei validen PDF/A mit der Konformität A und U gegeben und nicht bei der Konformität B. Die Konformität B garantiert die einwandfreie und systemunabhängige visuelle Wiedergabe, sowie die Unveränderbarkeit der Seiten. Die Extrahierbarkeit von Text ist bei der Konformität B optional.

- Die eingebettete Arial-Schrift erfüllt nicht die Anforderung der Archive.
- Schrift verstösst nicht gegen die PDF/A-Anforderung mit der Konformität B.

2.2 Assessment und KaD

Die KOST empfiehlt im Katalog archivischer Dateiformate KaD[1] PDF/A-1 und PDF/A-2 als archivtaugliche Dateiformate für die Langzeitarchivierung von Textdokumenten. In den Archiven ist das Dateiformat PDF/A eines der meistverwendeten. Ein Assessment der PPEG hatte zum Ziel zu ermitteln, wie häufig diese Fehler in den PDF/A-1b- und PDF/A-2b-Dateien in den verschiedenen Archiven auftaucht.

Da ein solches Assessment mit den bestehenden Tools nicht möglich war, hat die KOST die Firma PDF Tools AG beauftragt für den «3-Heights™ PDF Validator» ein Zusatzmodul zu entwickeln. Dieser Validator ist auch in KOST-Val[2] integriert.

2.2.1 Assessment-Ergebnis

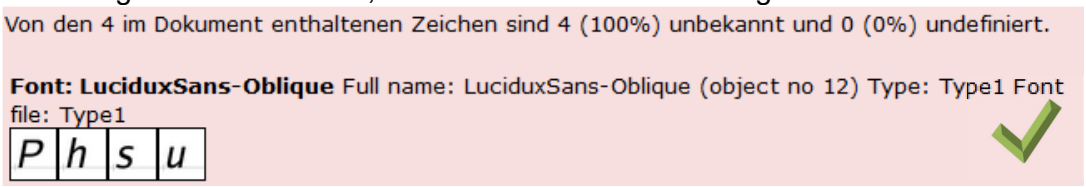
Das Assessment mit Hilfe von KOST-Val hat ergeben, dass solche PDF/A-Dokumente in allen Archiven[3] vorhanden sind, und dass der Fehler nicht auf einzelne Schriftarten oder Herstellungsjahre eingeschränkt werden kann. PDF/A-Dokumente von Herstellern qualitativ hochwertiger Software sind im Schnitt weniger betroffen als z.B. von kostenlose Produkte. Der Fehler wurde zudem in allen PDF-Arten gefunden: in invaliden PDF/A-Dateien mit der Konformität A oder U, in PDF-Dateien mit eingebetteten Schriften sowie hauptsächlich in PDF/A-Dateien mit der Konformität B.

Mit der Konfiguration «strict» bemängelt der «3-Heights™ PDF Validators»[4] alle unbekannten und undefinierten Zeichen, welche verwendet werden[5]. Undefinierte Zeichen sind z.B. jene Zeichen die den Unicode-Wert U+FFFD (= unbekannt) enthalten. In den am Assessment beteiligten KOST-Archiven sind im Schnitt rund 20% aller PDF/A-1b- und PDF/A-2b-Dateien betroffen.

Diese Erkenntnis hatte zur Folge, dass die KOST in KOST-Val eine Nachauswertung programmierte, um diejenigen Dokumente zu ignorieren, welche für die Volltextsuche unkritische Mängel beinhalten, und die wirklich kritischen Dokumente schneller zu finden.

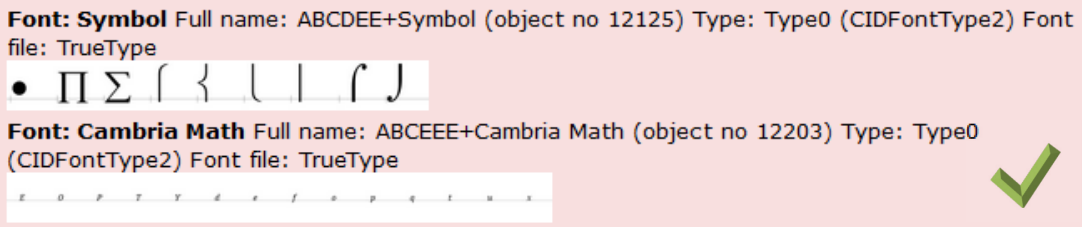
Mit der Konfiguration «tolerant» erfolgt die Nachauswertung durch KOST-Val.
Dabei werden nachfolgende Mängel in den Schriften ignoriert:

1. Alle Mängel in den Schriften, wenn maximal 5 Zeichen eingebettet wurden

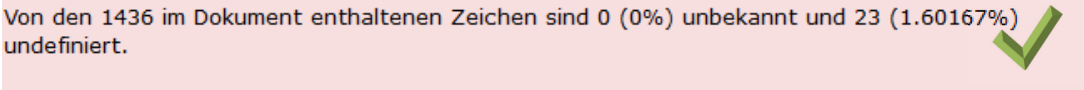


[1] Katalog archivischer Dateiformate: https://kost-ceco.ch/cms/kad_main_de.html
[2] KOST-Val: https://kost-ceco.ch/cms/kost-val.html
[3] Für das Assessment wurden nebst den Dateien der KOST-Geschäftsstelle (Sammlung aus unterschiedlichen Quellen und Archiven) auch archivierte PDF/A-Dateien von den Staatsarchiven SG, BS und UR sowie dem Bundesarchiv verwendet.
[4] 3-Heights™ PDF Validators: https://www.pdf-tools.com/pdf20/de/produkte/pdf-converter-validation/pdf-validator/
[5] Es kann allerdings sein, dass ein Zeichen zwar verwendet wird, auf der Seite aber nicht sichtbar ist, falls dieses z.B. überdeckt wird oder auf Grund von Transparenz-Effekten kein Kontrast erzeugt.

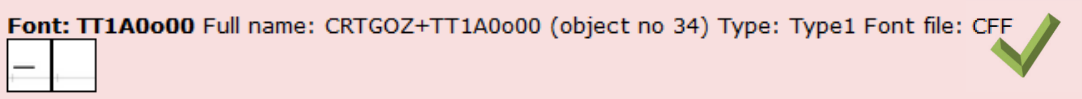
2. Mängel in den Schriften, welche im Fontnamen «Webdings», «Wingdings», «Symbol» oder «Math» enthalten



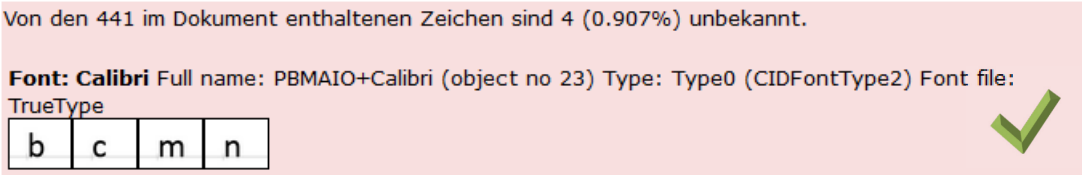
3. Alle undefinierten Zeichen, sofern diese nicht die Schwelle von 20% aller Zeichen überschreiten



4. Der KOST bekannte Zeichen, welche Abstände, Aufzählungszeichen und Auslassungspunkte (...) darstellen[6]



5. Wenn am Schluss nur noch maximal 4 Zeichen bemängelt würden[7]



Mit der Konfiguration «tolerant» von KOST-Val werden in den am Assessment beteiligten KOST-Archiven im Schnitt noch rund 1% aller PDF/A-1b- und PDF/A-2b-Dateien bemängelt. Dies ist massiv weniger als mit der Konfiguration «strict» und identifiziert so nur die kritischen PDF/A-Dateien.

Fazit:

- Alle PDF-Dateien mit eingebetteten Schriften können diesen Mangel enthalten.
- Archive müssen damit rechnen, dass ohne Massnahmen 1% aller PDF/A-Dateien mit der Konformität B mit einer Volltextsuche nicht gefunden würden.

[6] Bei der Durchführung des Assessments wurde festgestellt, dass in sehr vielen Dokumenten nur diese Zeichen-Kategorien bemängelt werden. Auf weitere Zeichen wurde verzichtet, da diese manuell respektive visuell definiert werden müssen. Dies ist mit einem erheblichen Aufwand verbunden.

[7] Nach der Durchführung von Schritt 1-4 des Assessments wurde festgestellt, dass ein Viertel aller restlichen Mängel nur vier oder weniger Zeichen bemängelt werden. Im schlechtesten Fall sind dies Buchstaben oder Zahlen. Dennoch ist es sehr unwahrscheinlich, dass aufgrund dieser vier Zeichen eine Volltextsuche scheitern würde. Meistens handelt es sich um Zeichen die in die Kategorie 2 und 4 zugeordnet werden könnten.

3 Recommendations

3.1 Mängelvermeidung

Möchte ein Archiv in Zukunft nur noch durchsuch- und extrahierbar PDF/A-Dateien haben, muss es zur Vermeidung dieses Mangels mindestens die Konformität U bei neu erstellten PDF/A-Dateien verlangen. Im KaD wird im Fazit bereits darauf hingewiesen.

- Vermeidung: Konvertierungen mit qualitativ hochwertige Software erzeugen und mindestens PDF/A-2u verlangen.

3.2 Mängelkennzeichnung

Nach dem Assessment kann jedes Archiv diejenigen bereits existierenden PDF- und PDF/A-Dateien im Archiv identifizieren, die nicht vollständig durchsuch- und extrahierbar sind.

Als minimale Preservation Action sollen diese Dateien im Findmittel mit einer Notiz gekennzeichnet[8] werden.

- Kennzeichnung: Im Findmittel alle Dateien kennzeichnen, die die Validierung im Modul K mit der Konfiguration «strict» nicht bestanden haben.

3.3 Mängelbehebung

Bereits existierende PDF- und PDF/A-Dateien im Archiv, welche die Validierung im Modul K mit der Konfiguration «tolerant» nicht bestanden haben, riskieren für die Volltextsuche nicht auffindbar zu sein.

Diese Dateien sollten nebst der Kennzeichnung auch eine Mängelbehebung durchlaufen, bei der die einzelnen fehlenden Zeichen mit den entsprechenden Unicodezeichen ergänzt werden.

- Behebung: Preservation Action auf alle Dateien durchführen, die die Validierung im Modul K mit der Konfiguration «tolerant» nicht bestanden haben.

[8] Das BAR verzichtet auf eine Kennzeichnung und die vorgängige Behebung bestehender PDF-Dateien. Im Gegenzug werden bei der Auslieferung konsequent alle bestellten PDF-Dateien in ein neues PDF/A-2u mit OCR konvertiert. Damit sind die ausgelieferten PDF-Dateien alle durchsuch und extrahierbar. Sollte das BAR zu einem späteren Zeitpunkt die Volltextsuche über das digitale Magazin einführen, riskiert es ohne weitere Massnahmen, dass viele Dateien für die Volltextsuche nicht auffindbar sein werden.

Bei dem Mangel muss zwischen zwei verschiedenen Arten von Fehlern unterschieden werden:

a) Modul D und K werden bemängelt

D) Schriften	Das Dokument enthaelt Schriften ohne eingebettete Font-Programme oder Codierungsinformation (CMAPs). [PDF Tools: iCategory_8]
K) Schrift-Validierung	Von den 27 im Dokument enthaltenen Zeichen sind 23 (85.1852%) unbekannt und 0 (0%) undefiniert. **Font: Calibri** Full name: Calibri (object no 65) Type: Type0 (CIDFontType2) T a b c d e g h i l m n o p r s t u w . “ ”

b) Modul K wird bemängelt aber nicht Modul D

K) Schrift-Validierung	Von den 971 im Dokument enthaltenen Zeichen sind 14 (1.44181%) unbekannt und 0 (0%) undefiniert. **Font: StoneSans-Semibold** Full name: JLPQRJ+StoneSans-Semibold (object no 397) Type: Type0 (CIDFontType0) Font file: CFF R T a b d e l n o r t u x

3.3.1 Mängelbehebung: Modul D und K werden bemängelt

Im Fall a) geschieht die Mängelbehebung meist automatisch, wenn die PDF-Datei mit einem professionellen Tool nach PDF/A-2u konvertiert wird. Dabei werden die nicht valid eingebetteten Schriften, wenn möglich korrekt eingebettet und mit Unicodewerten ergänzt. Dies funktioniert sehr zuverlässig, wenn die Schrift auch auf dem Rechner installiert respektive eine Standardschrift ist. Der Fall a) tritt in weniger als einem Drittel aller Fälle auf.

3.3.2 Mängelbehebung: Modul K wird bemängelt aber nicht Modul D

Im Fall b) sind die Schriften zwar eingebettet, aber nur die Kontur, wie es beim eingangs beschriebenen Incident der Fall ist. Das Assessment hat aufgezeigt, dass dies der häufigste Fall ist. Für diese Mängelbehebung ist es am hilfreichsten, die Kontur auszuwerten.

Die Firma PDF Tools AG hat parallel zur Erweiterung ihres 3-Heights™ PDF Validators[4] auch 3-Heights™ PDF OCR[9] angepasst. Dabei werden die bemängelten Zeichen in der Schrifttabelle mittels OCR Engine (ABBYY FineReader ausgelesen und das Ergebnis als Unicodezeichen in der Schrifttabelle beim entsprechenden Zeichen hinterlegt. Natürlich bleibt die PDF/A-Konformität erhalten. Die OCR-Erkennung ist hauptsächlich auf Zahlen und Buchstaben spezialisiert; die Erkennung von anderen Zeichen ist weniger genau.

Die OCR-Erkennung kann deutlich verbessert werden, indem die Optionen respektive Einstellungen jeweils an das Archiv und an das Dokument angepasst werden.

Die KOST hat folgenden Aufruf für das Testset verwendet:

```
pdfocr -v -ocr "service" -ocp "Profile=C:\Dateipfad\abbyy_text.txt"
-ocl "English,German,French,Mathematical" -otm update -ots
"knownSymbolic" -otu "installedFont,fallbackAllPua" input.pdf
output.pdf
```

[9] 3-Heights™ PDF OCR: https://www.pdf-tools.com/pdf20/de/produkte/pdf-converter-validation/pdf-ocr/

Das Profile «abbyy_text.txt», enthält nachfolgenden Inhalt, der es erlaubt Texte in einer anderen Orientierung respektive Ausrichtung zu erkennen. Die Erkennung des Textes in Bilder ist auch möglich.:

```
[PagePreprocessingParams]
CorrectOrientation=TRUE

[PageAnalysisParams]
DetectVerticalEuropeanText=TRUE

[ObjectsExtractionParams]
DetectTextOnPictures = TRUE
```

Als Hauptsprache respektive Typ (-ocl) wurden Englisch, Deutsch, Französisch sowie mathematische Zeichen definiert. Dies ist hilfreich, da das OCR-Programm mit Wörterlisten der jeweiligen Sprache arbeitet und damit eine höhere Qualität und bessere Performanz erzielt.

Mit «-otm update» werden alle Zeichen ohne sinnvollen Unicode-Wert bearbeitet. Wird gleichzeitig auch «-ots "knownSymbolic"» mitgegeben, werden aber klassische Symbolschriftarten wie z.B. «ZapfDingbats» und «Wingdings» bei der Text-OCR-Verarbeitung übersprungen.

Durch die Verwendung von «-otu "installedFont,fallbackAllPua"» werden auf dem System installierte Schriften und deren Unicode-Zuordnung mitverwendet, und mit fallbackAllPua werden die Zeichen, für welche kein geeigneter Unicode-Wert bestimmt werden konnte, als undefiniert erklärt. Damit können anschliessend mehr Dokumente die tolerant-Schriftvalidierung bestehen.

Beim Test in der KOST-Geschäftsstelle konnte mit dem 3-Heights™ PDF OCR 66% der fehlerhaften Dokumente korrigiert werden. Zudem wurde die Anzahl der bemängelten Zeichen um 87% reduziert. Die restlichen 13% der bemängelten Zeichen konnten nicht erkannt werden und wurden durch die Option «fallbackAllPua» als undefiniert erklärt. Dadurch konnte aus der Sicht der tolerant-Schriftvalidierung alle Dokumente behoben werden. Diese Zahlen hängen jedoch stark von den Ausgangsdokumenten ab.

Original Screenshot	Original Extraktion	Extraktion nach Behebung
© Photo-montage	□□□□□□□□□□□□□□□	© Photo-montage
$ax^2 + bx + c = 0$	__ + __ + _ = 0	ax2 + bx + c = 0
newsletter	newsle□er	newsletter

Mit dem produktiven Einsatz von 3-Heights™ PDF OCR wird erwartet, dass sich das Tool weiter verbessern wird – dies unter der Voraussetzung, dass der Hersteller laufend Dokumente zur Verfügung gestellt bekommt, welche noch nicht zufriedenstellend behoben werden konnten.

3.3.3 Kontrolle der Mängelbehebung

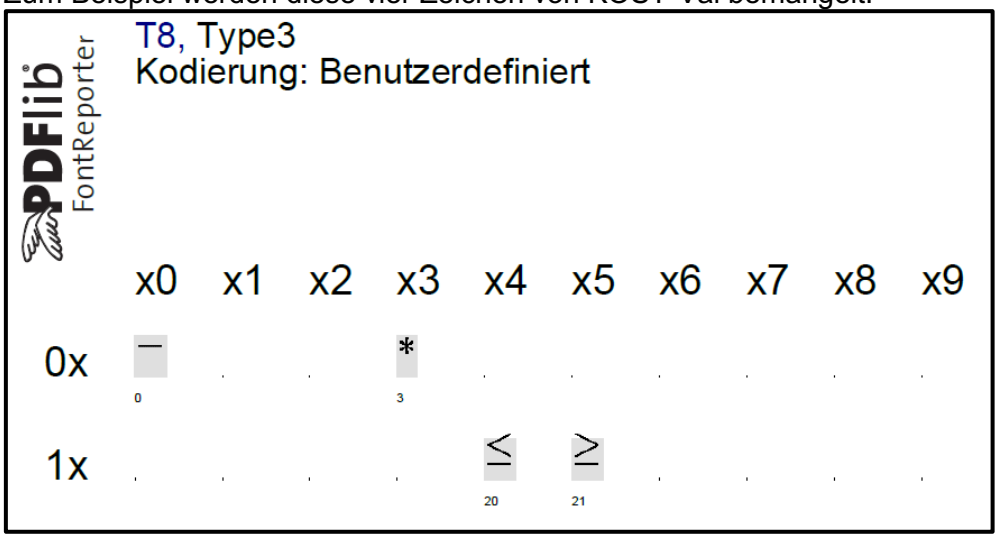
Nach der Mängelbehebung sollte eine neue Validierung mit KOST-Val durchgeführt werden. Idealerweise müsste dann das Problem behoben sein. Eine Kontrolle der Qualität der korrigierten respektive erkannten Zeichen ist nur visuell möglich. Da dies sehr aufwändig ist, können nicht viele Dokumente entsprechend kontrolliert werden. Die KOST Geschäftsstelle hat zur Kontrolle neben KOST-Val auch PDFlib FontReporter[10] verwendet. PDFlib FontReporter ist ein kostenloses Plug-In für Adobe Acrobat zur Analyse von Fonts in PDF-Dokumenten. Per einfachem Mausklick erstellt PDFlib FontReporter einen detaillierten Bericht über eine in Acrobat geöffnete PDF-Datei. PDFlib FontReporter erfasst allgemeine Daten, Font-Informationen sowie Glyphen-Tabellen zu allen in einem PDF vorhandenen Fonts und erstellt ein separates PDF mit einem Report.

Für die visuelle Qualitätskontrolle wurden bei den von KOST-Val bemängelten Zeichen die Unicode-Werte aus dem Report gelesen und kontrolliert. Dabei kann der Unicode-Wert in einer Tabelle[11] nachgelesen werden oder in Word eingegeben und mittels der Tastenkombination [Alt] [C] umgewandelt werden.

3.3.4 Probleme bei der Mängelbehebung

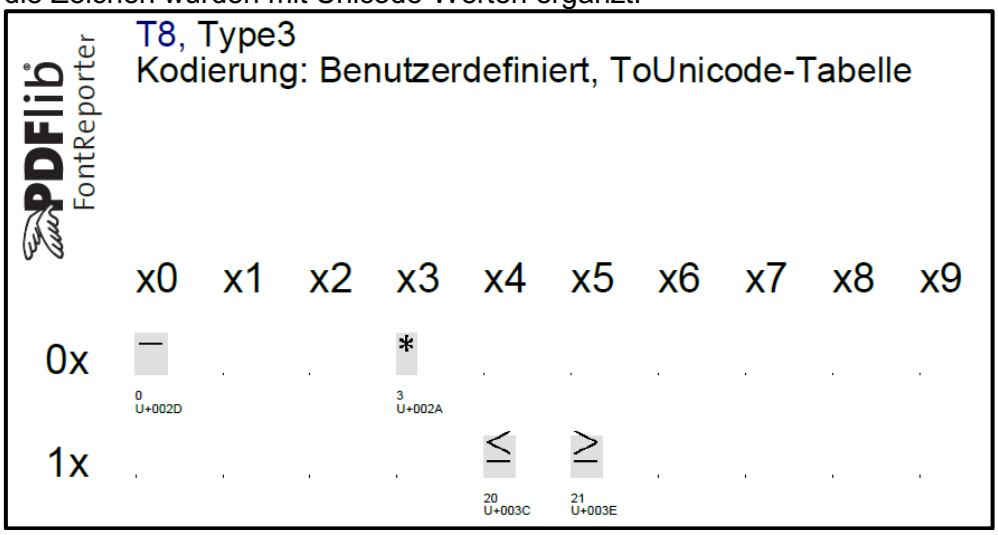
Bei der Qualitätskontrolle wurden insbesondere Probleme bei benutzerdefinierten Schriften und bei Zeichen, welche keine lateinischen Buchstaben und Zahlen darstellen, festgestellt.

Zum Beispiel werden diese vier Zeichen von KOST-Val bemängelt:



[10] PDFlib FontReporter: https://www.pdflib.com/download/free-software/fontreporter/
[11] z.B. https://www.compart.com/de/unicode/

Nach der Behebung ist die Kodierung neu «Benutzerdefiniert, ToUnicode-Tabelle» und die Zeichen wurden mit Unicode-Werten ergänzt:

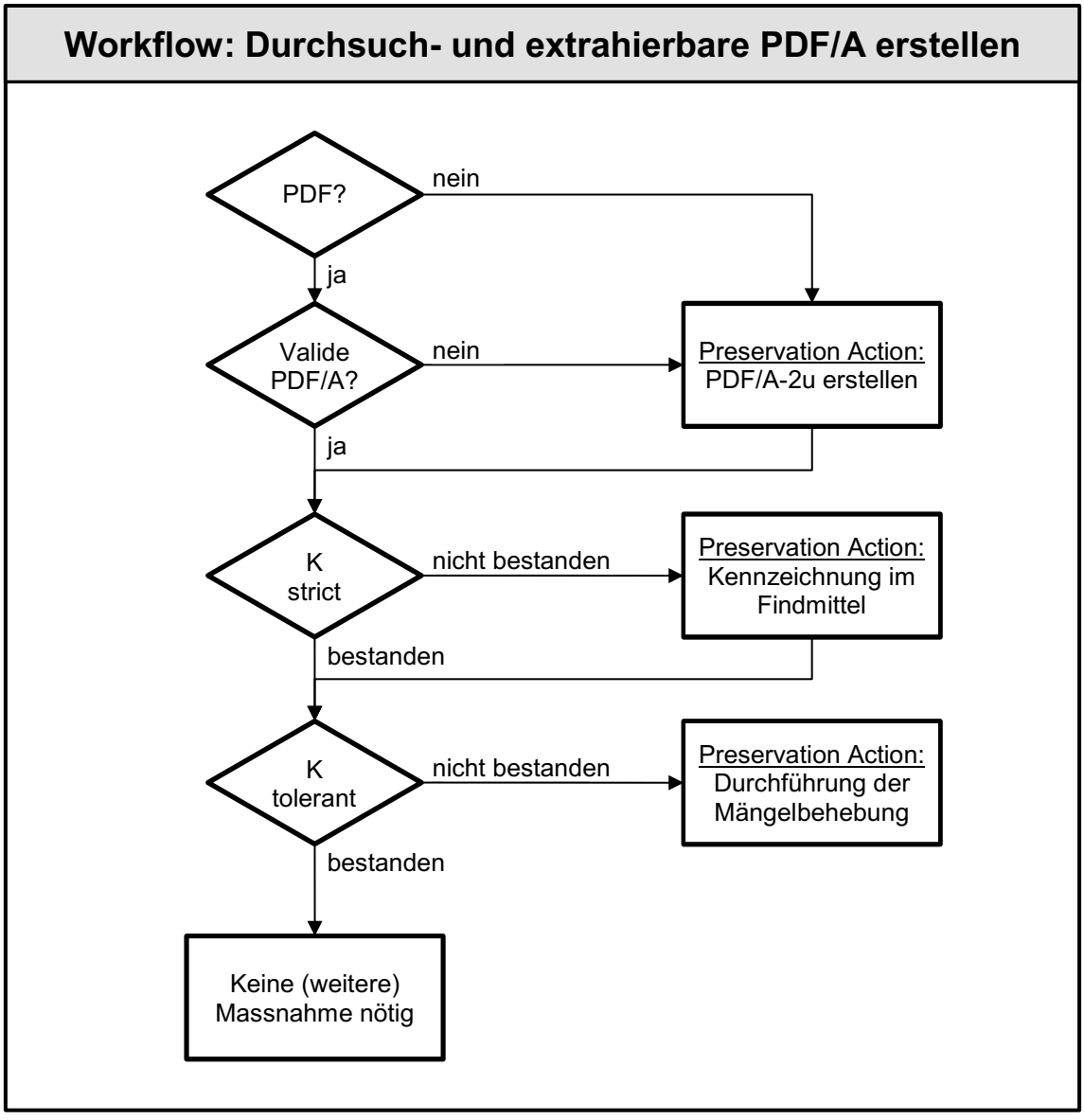


Wenn jetzt die einzelnen Unicode-Werte kontrolliert werden, stellt sich heraus, dass die Erkennung nicht 100% stimmt.

	Probleme bei der Mängelbehebung			
Darstellung KOST-Val	—	*	≤	≥
Darstellung FontReporter	— 0 U+002D	* 3 U+002A	≤ 20 U+003C	≥ 21 U+003E
Textpassage	range −12	$V * 10^{-N}$	$\leq p \leq n.$	If $N \geq 0,$
Erkannter Unicode-Wert	U+002D	U+002A	U+003C	U+003E
Erkanntes Unicode-Zeichen	-	*	<	>
Problem bei der Erkennung	keines	Erkannt wurde ein 5-strahliger Stern, welcher hochgestellt ist	Der untere Strich wurde von der OCR Engine (ABBYY FineReader) nicht erkannt, da er in dieser Schriftart unterhalb der Schriftgrundlinie liegt, was für diese Zeichen unüblich ist	
Korrekter Unicode-Wert		U+2217 ∗	U+2264 ≤	U+2265 ≥

3.4 Recommendations: Zusammenfassung

Die Archive sollten den beschriebenen Mangel in Zukunft vermeiden, indem sie mindestens PDF/A-2u verlangen, bestehende Dateien, welche «K strict» nicht bestehen, im Findmittel kennzeichnen und zusätzlich jene Dateien, die «K tolerant» nicht bestanden haben, korrigieren.



- Vermeidung: Konvertierungen mit qualitativ hochwertige Software erzeugen und mindestens PDF/A-2u verlangen.
- Kennzeichnung: Im Findmittel alle Dateien kennzeichnen, die die Validierung im Modul K mit der Konfiguration «strict» nicht bestanden haben.
- Behebung: Preservation Action auf alle Dateien durchführen, die die Validierung im Modul K mit der Konfiguration «tolerant» nicht bestanden haben.