Detailanalyse PDF/A-1b mit eingebetteter TrueType Schrift
Einbettung einer korrupten TrueType Schrift

1 Management Summary

Am 20. August 2014 hat die KOST einen beunruhigenden Fehler beim Drucken eines PDF-Dokuments[1] mit einem PCL-Drucker festgestellt. Der Fehler besteht hauptsächlich darin, dass mehrfach ganze Passagen nicht gedruckt werden und die visuelle Reproduzierbarkeit des Dokuments somit nicht gegeben ist. (Von den 2971 Zeichen der ersten Seite wurden lediglich 2340 Zeichen wiedergegeben, was bedeutet, dass über 21 Prozent nicht gedruckt wurde, siehe Abbildung 1, entsprechende Passagen mit rosa Marker hervorgehoben).

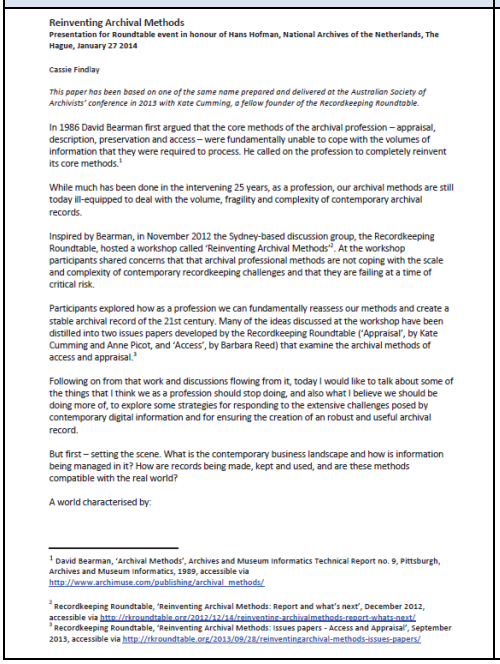
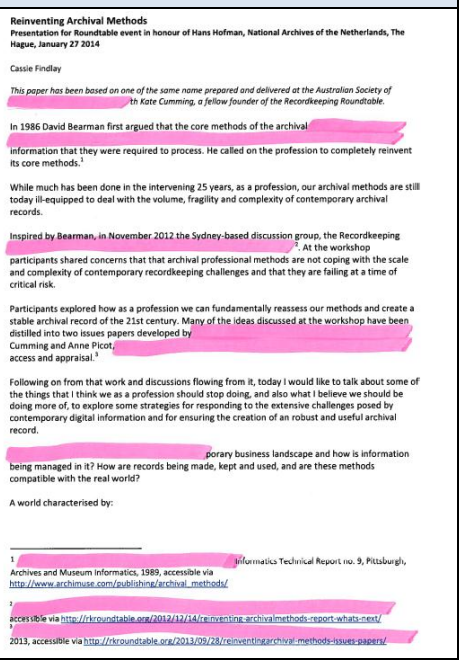
Anzeige	Druckergebnis
	

Abbildung 1: links PrintScreen der ersten Seite, rechts das Druckergebnis mit den fehlenden Passagen in rosa markiert

[1] Das PDF-Dokument (MD5-Summe: 05ff9afaf7ded808c3200eb1beff69fc) wurde von http://www.nationaalarchief.nl/sites/default/files/docs/nieuws/cassie_findlay_reinventing_archival_methods_the_hague_27jan_2014a.pdf heruntergeladen.

Die Tests der KOST im September 2014 haben ergeben, dass einige Zeichen der eingebetteten TrueType Schrift "Calibri" offensichtlich fehlerhaft definiert sind. Dennoch wird das Dokument von den führenden PDF/A-Validatoren als valid identifiziert.

Eine erste Analyse der KOST wurde im November 2014 veröffentlicht[2] und den involvierten Akteuren gesendet. Sie definierte Massnahmen auf verschiedenen Ebenen: Präzisierung der ISO 19005-Standards, Korrektur des Fehlers in den entsprechenden PDF/A-Konvertern sowie Erkennung des Fehlers durch PDF/A-Validatoren.

Da innerhalb 6 Monate keine zufriedenstellende Rückmeldung der Hersteller und von ISO zurückkam, sah sich die KOST gezwungen, den Fehler weiter zu analysieren und zu dokumentieren. Diese hier vorliegende Detailanalyse zeigt sehr genau auf, wo der Fehler liegt und warum nicht nur die betroffenen Zeichen im Ausdruck fehlen, sondern ganze Passagen nicht gedruckt werden. Zudem zeigt sie auch auf, dass die eingebettete Schrift im PDF-Dokument das Problem ist und nicht der Adobe Viewer[3].

An den im Oktober 2014 definierten Massnahmen hat sich nichts verändert, ausser dass zusätzlich von den betroffenen Hersteller und der ISO/TC 171/SC 2/WG 5 eine schriftliche Stellungnahme verlangt wird.

Erst wenn die PDF/A-Validatoren den Fehler erkennen, können die Archive die betroffenen PDF-Dokumente ermitteln und korrigieren.

2 Analysen

2.1 Erstellende Applikation

Die Dateieigenschaften des in Frage stehenden PDF-Dokuments zeigen auf, dass es mit "Microsoft® Word 2010" erstellt und mit "Acrobat Distiller 11.0 (Windows)" erzeugt wurde.

Beides sind aktuelle Systeme und entsprechend verbreitet im Einsatz.

2.2 Druckergebnisse

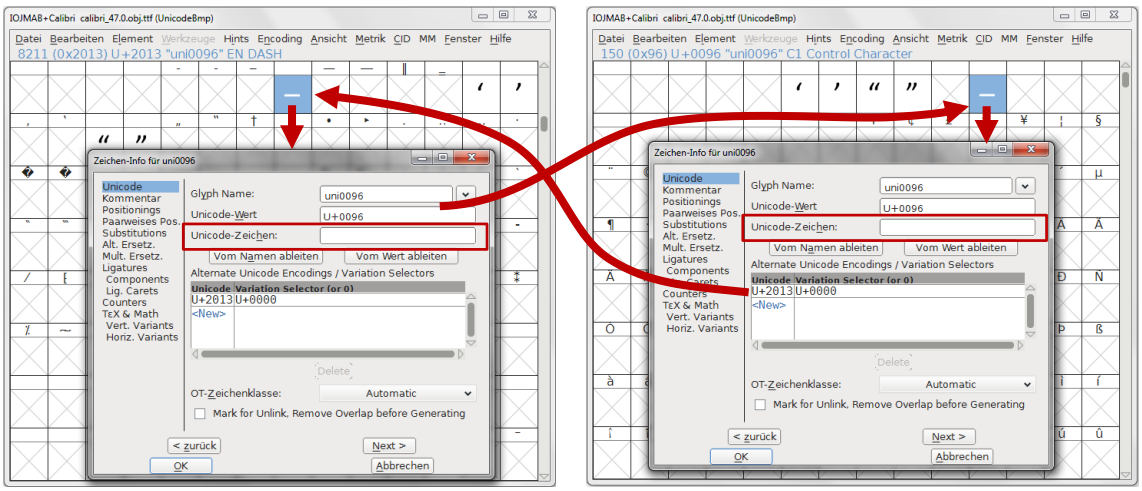
Wenn das PDF-Dokument mit einem Adobe-Produkt geöffnet und auf einem Drucker ohne PostScript (und ohne die erweiterte Druckoption "als Bild drucken") gedruckt wird, treten bedeutende Fehler im Erscheinungsbild auf. Dabei spielt es keine Rolle, welcher PCL-Drucker verwendet wird; der Fehler erscheint zudem bei der Konvertierung in XPS. Die Drucktests der KOST im September 2014 lassen folgende Rückschlüsse zu:

- Es sind alle PCL-Drucker betroffen
- Adobe verwendet die eingebettete Schrift, welche korrupt ist[4]
- Mindestens folgende Calibri-TrueType-Zeichen sind fehlerhaft definiert:
  – [U+2013] oder ‘ [U+2018] respektive ’ [U+2019][5]

[2] http://kost-ceco.ch/cms/download.php?4a479f8b024ab61dfc53bc2c7c83b45a.
[3] Es wird angenommen, dass nur Adobe bei Standard-Schriften die eingebettete Schrift verwendet.
[4] Es wird angenommen, dass nur Adobe bei Standard-Schriften die eingebettete Schrift verwendet.
[5] In der ersten Analyse wurden den Zeichen teilweise falsche Unicode-Nummern zugeordnet.

2.3 Calibri TrueType Analyse

Mit Hilfe von PDFXplorer[6] wurde die betroffene Schrift Calibri (47 0 obj) extrahiert und als „calibri_47.0.obj.ttf“ abgespeichert. Bei der Analyse dieser TrueType-Font mit FontForge[7] wurde der Fehler offensichtlich.



Die betroffenen Zeichen sind nicht definiert und zusätzlich in einem zirkularen Verweis eingebunden. Der zirkulare Verweis ist der Auslöser dafür, dass ganze Passagen nicht gedruckt werden können.

Folgende Zeichen sind in diesem Dokument betroffen:

- –: U+2013 verweist auf U+0096 verweist zurück auf U+2013
- ‘: U+2018 verweist auf U+0091 verweist zurück auf U+2018
- ’: U+2019 verweist auf U+0092 verweist zurück auf U+2019
- “: U+201C verweist auf U+0093 verweist zurück auf U+201C
- ”: U+201D verweist auf U+0094 verweist zurück auf U+201D

Die Unicode Zeichen U+0091 bis U+0096 sind zudem in der Calibri Schrift nicht unterstützt[8]:

- U+0091 private use one
- U+0092 private use two
- U+0093 set transmit state
- U+0094 cancel character
- U+0096 start of guarded area

Bei der eingebetteten Calibri-Schrift handelt es sich um die SFNT Revision 5.62, welche mit Microsoft® Office 2010 respektive Windows 7 ausgeliefert wurde[9].

[6] http://www.o2sol.com/pdfxplorer/overview.htm
[7] http://fontforge.github.io/en-US/
[8] http://www.fileformat.info/info/unicode/font/calibri/missing.htm
[9] Die korrupte Calibri-Schrift wurde durch Microsoft® bereits in Updates ersetzt. Windows 7 mit Microsoft® Office 2010 und allen Updates installiert enthalten die Revision 5.73.

2.4 PDF/A-1B

2.4.1 PDF/A-1B – Ziel und Zweck

Der grundsätzliche Anspruch von PDF/A ist es, die visuelle Reproduzierbarkeit korrekt zu gewährleisten. Dieser Anspruch ist auch im dritten Absatz der *Introduction* explizit festgehalten:

> The primary purpose of this part of ISO 19005 is to define a file format based on PDF, known as PDF/A, which provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems used for creating, storing or rendering the files.

Das vorliegende PDF/A-1b-Dokument verletzt mit der Einbettung von korrupten Schriften zwar nicht ein konkretes Requirement[10], aber sehr wohl dieses generelle Statement. Es wäre also zu erwarten, dass ein PDF/A-Validator darauf reagiert.

2.4.2 PDF/A-1B-Validierung

Das in Frage stehende PDF/A-1b-Dokument wurde mit folgenden PDF/A-Validatoren getestet:

- Preflight in Adobe Acrobat Pro Version 10.1.10 & 10.1.13
- PDF/A-Manager Version V6.1121853 & V6.500 von PDFTron
- 3-Heights ™ PDF Validator Version 4.3 & 4.5.6 von PDF Tools AG
- pdfaPilot Version 5.1.211 & 5.5.232 von Callas

Alle Validatoren identifizierten das Dokument als valides PDF/A-1b-Dokument.

3 Fazit der Analysen

Wenn ein mit den aktuellen Tools erzeugtes und als valid geprüftes PDF/A-Dokument nicht korrekt druckbar ist, haben die Archive ein (noch unbekanntes, aber potentiell grosses) Problem, für das eine Lösung dringend nottut.

Für die Archive ist es inakzeptabel, dass in einem validen PDF/A ganze Textpassagen mit Adobe Acrobat Pro und Adobe Reader nicht gedruckt werden. Von einem Erhalten der visuellen Reproduzierbarkeit unabhängig der verwendeten Systeme kann nicht annähernd die Rede sein.

Die fehlerhaften Unicode-Zeichen – [U+2013], ‘ [U+2018], ’ [U+2019], “ [U+201C] und ” [U+201D] sind geläufige Zeichen.

Die erstellenden Applikationen (Microsoft® Word 2010 mit Acrobat Distiller 11.0) sind aktuell, und entsprechend muss davon ausgegangen werden, dass weitere PDF-

[10] In ISO 19005-1 steht nicht explizit, dass diese Schriften korrekt eingebettet werden müssen (6.3.2 Font types: All fonts used in a conforming file shall conform to the font specifications defined in PDF Reference 5.5.), sondern nur, dass alle verwendeten Zeichen eingebettet sein müssen (6.3.3 ff). Dass sie korrekt sein müssen, wird implizit angenommen, aber nicht ausdrücklich gesagt. In ISO 19005-2 wurde der Satz noch mit der Aussage ergänzt, dass die zitierten Spezifikationen zur PDF-Referenz konform sein müssen (6.2.11.2 Font types: All fonts and font programs used in a conforming file, regardless of rendering mode usage, shall conform to the provisions in ISO 32000-1:2008, 9.6 and 9.7, as well as to the font specifications referenced by these provisions.).

Dokumente mit korrupten Schriften existieren und weiterhin produziert werden, da nicht alle zwingend die Updates von Microsoft® durchführen.

Entsprechend wird an den folgenden Preservation-Planning-Massnahmen vom Oktober 2014 festgehalten.

4 Preservation-Planning-Massnahmen

Das beschriebene Problem muss auf vier Ebenen gleichzeitig[11] angegangen werden. Deshalb werden im Oktober 2014 folgende Massnahmen in die Wege geleitet:

A. Die Hersteller des analysierten Dokumentes[12] werden über das Problem informiert und gebeten, die Reproduzierbarkeit des Fehlers abzuklären.
B. Die Herstellerin des PDF/A-Konverters wird angeschrieben und gebeten, nur valide TrueType-Schriften einzubetten.
C. Die Hersteller der getesteten Validatoren werden angeschrieben und gebeten, die Prüfung der eingebetteten Inhalte zu erweitern, damit PDF/A-Dokumente mit korrupten Inhalten nicht als valid gelten.
D. Das Sekretariat des ISO/TC 171/SC 2/WG 5, welches für den ISO 19005 zuständig ist, wird informiert und gebeten, mit einem 'Corrigenda' oder einem anderen Mechanismus die Satzergänzung von ISO 19005-2 6.2.11.2 "as well as to the font specifications referenced by these provisions" auch in ISO 19005-1 6.3.2 nachzutragen[13].

Die Massnahme C ist notwendig, damit solche PDF/A-Dokumente erkannt und korrigiert werden können. Da die herstellende Applikation eine aktuelle ist, kann mit der Massnahme B die Entstehung solcher PDF-Dokumente reduziert werden. Die Massnahme D soll dazu führen, dass die Hersteller von PDF/A-Software diesen Fehler entweder nicht produzieren oder ihre Validatoren entsprechend aufdatieren.

Erst wenn die PDF/A-Validatoren den Fehler erkennen, können die Archive die betroffenen PDF-Dokumente ermitteln und korrigieren.

5 Stellungnahme

Da die erste Anfrage vom November 2014 leider nicht zielführend war, bittet die KOST jetzt die Software-Hersteller und die ISO/TC 171/SC 2/WG 5 eine schriftliche Stellungnahme bis zum 31. August 2015. Bei den Hersteller der getesteten Validatoren wird in dieser Stellungnahme zudem erwartet, dass genau beschreiben wird, welche eingebetteten Inhalte nicht mitvalidiert werden und aus welchen Gründen.

[11] Alle vier Ebenen werden gleichzeitig angegangen, da unsere Erfahrung darin besteht, dass die beteiligten Ebenen sich gegenseitig abstützten, und es nicht zielführend ist, wenn nur eine Ebene angegangen wird.
[12] Sowohl die Autorin Cassie Findlay als auch die publizierende Institution, das Niederländische Nationalarchiv.
[13] Die Präzisierung in ISO 19005-2 6.2.11.2 hätte gleichzeitig mit einem Corrigenda in die Version 1 zurückfliessen sollen. Das Problem nur in den neuen Versionen anzugehen, greift zu kurz, da die bisherigen Versionen explizit gültig bleiben und bleiben sollen.