Detailed analysis of PDF/A-1b with embedded TrueType font
Embedding of a corrupt TrueType font

1 Management Summary

On 20 August 2014, KOST identified a worrying error when printing a PDF document[1] using a PCL printer. The main problem is that in a number of instances, entire passages are not printed and the document's visual appearance is not preserved (of the 2,971 characters on the first page, only 2,340 were reproduced, meaning that more than 21% was not printed – see Fig. 1; the relevant passages are highlighted with pink marker).

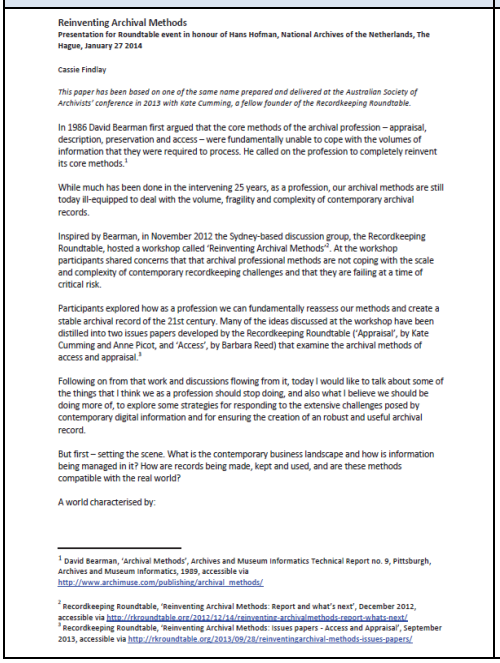
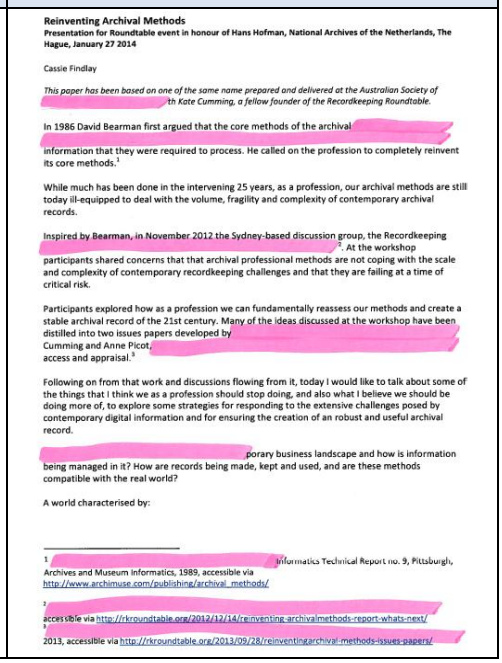
Document as displayed	Printed document
	

Fig. 1: left: screenshot of the first page; right: the document as printed, with the missing passages highlighted in pink

[1] The PDF document (MD5 hash: 05ff9afaf7ded808c3200eb1beff69fc) was downloaded from http://www.nationaalarchief.nl/sites/default/files/docs/nieuws/cassie_findlay_reinventing_archival_methods_the_hague_27jan_2014a.pdf.

Tests conducted by KOST in September 2014 revealed that some characters in the embedded TrueType font “Calibri” appear to be incorrectly defined. However, the leading PDF/A validators identify the document as valid.

An initial analysis was published by KOST in November 2014[2] and sent to those concerned. It defined measures on a number of levels: further clarification of the ISO 19005 standard, correction of the error in the corresponding PDF/A converters, and recognition of the error by PDF/A validators.

Since no satisfactory response was received from manufacturers and ISO within 6 months, KOST felt it necessary to further analyse and document the error. This detailed analysis precisely identifies where the error lies and explains why entire passages, rather than just the characters affected, are not printed. It also demonstrates that the embedded font in the PDF document is the problem, and not Adobe Viewer.[3]

No changes have been made to the measures defined in October 2014, except that a written response is requested from the manufacturers and ISO/TC 171/SC 2/WG 5.

Archives will not be able to identify and correct the PDF documents affected until the PDF/A validators recognise the error.

2 Analyses

2.1 Application used to create the document

The file properties of the PDF document in question indicate that it was created using Microsoft® Word 2010 and generated using Acrobat Distiller 11.0 (Windows).

Both are current systems and are in widespread use.

2.2 Results of printing

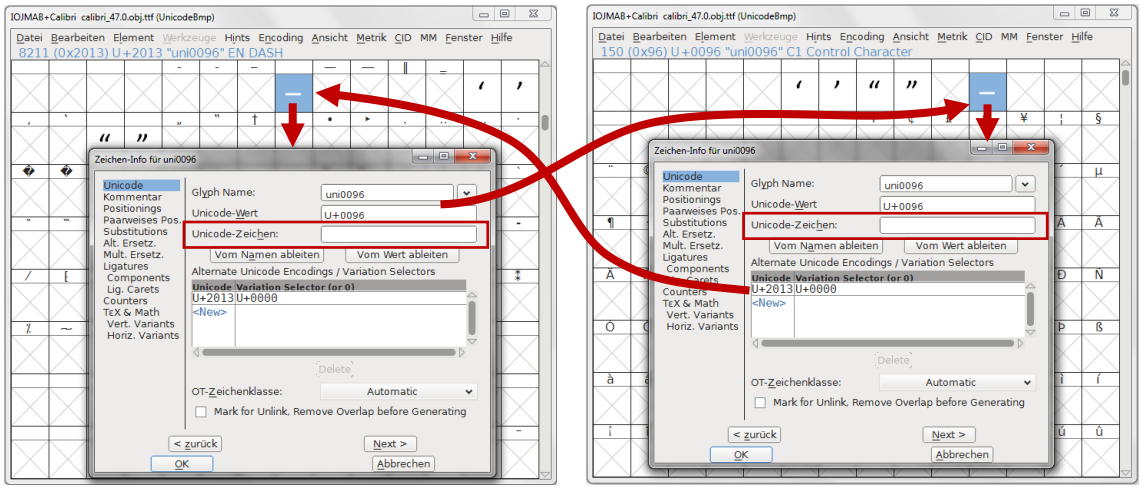
If the PDF document is opened using an Adobe product and printed on a printer without PostScript (and without the advanced option “Print as Image”), significant errors occur in its visual appearance. The type of PCL printer used makes no difference; the error also occurs when converting to XPS. Printing tests carried out by KOST in September 2014 establish the following conclusions:

- All PCL printers are affected
- Adobe uses the embedded font, which is corrupt[4]
- At least the following Calibri TrueType fonts are incorrectly defined:
  – [U+2013] or ‘ [U+2018] and ’ [U+2019][5]

[2] http://kost-ceco.ch/cms/download.php?4a479f8b024ab61dfc53bc2c7c83b45a
[3] It is assumed that only Adobe uses the embedded font for standard fonts.
[4] It is assumed that only Adobe uses the embedded font for standard fonts.
[5] In the first analysis, incorrect Unicode numbers were assigned to the characters in some cases.

2.3 Calibri TrueType analysis

The font affected, Calibri (47 0 obj), was extracted using PDFXplorer[6] and saved as “calibri_47.0.obj.ttf”. The error became apparent when analysing this TrueType font with FontForge.[7]



The characters affected are not defined and are also involved in a circular reference. That circular reference is the reason why entire passages cannot be printed.

The following characters are affected in this document:

- –: U+2013 refers to U+0096 refers back to U+2013
- ‘: U+2018 refers to U+0091 refers back to U+2018
- ’: U+2019 refers to U+0092 refers back to U+2019
- “: U+201C refers to U+0093 refers back to U+201C
- ”: U+201D refers to U+0094 refers back to U+201D

Additionally, the Unicode characters U+0091 to U+0096 are not supported in the Calibri font:[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

The embedded Calibri font is SFNT Revision 5.62, which was supplied with Microsoft® Office 2010 and Windows 7.[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] The corrupt Calibri font has already been replaced by Microsoft® in updates. Windows 7 with Microsoft® Office 2010 and all updates installed contains Revision 5.73.

2.4 PDF/A-1B

2.4.1 PDF/A-1B – aims and purpose

PDF/A claims to ensure that the visual appearance of documents is correctly reproduced. This claim is also explicitly stated in the third paragraph of the *Introduction*:

> 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.

While the embedding of corrupt fonts in the present PDF/A-1b document does not breach a specific requirement,[10] it is almost certainly incompatible with this general statement. It is therefore reasonable to expect that a PDF/A validator would react to it.

2.4.2 PDF/A-1B validation

The PDF/A-1b document in question was tested using the following PDF/A validators:

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

All the validators identified the document as a valid PDF/A-1b document.

3 Conclusion from analyses

If a PDF/A document generated using current tools and tested as valid cannot be printed correctly, archives have an as yet unknown, but potentially major problem that requires an urgent solution.

It is unacceptable for archives that entire passages of text in a valid PDF/A are not printed when using Adobe Acrobat Pro and Adobe Reader. This does not even come close to the claimed facility of preserving visual appearance over time, independent of the systems used.

The defective Unicode characters – [U+2013], ‘ [U+2018], ’ [U+2019], “ [U+201C] and ” [U+201D] are in everyday use.

The applications used to create the document (Microsoft® Word 2010 with Acrobat Distiller 11.0) are up to date, and it is therefore reasonable to assume that further PDF documents with corrupt fonts are in existence and continue to be produced, as not everyone necessarily installs updates from Microsoft®.

[10] ISO 19005-1 does not explicitly state that these fonts must be embedded correctly (6.3.2 Font types: All fonts used in a conforming file shall conform to the font specifications defined in PDF Reference 5.5.), but only that all the characters used must be embedded (6.3.3 ff). It is implicitly assumed that these must be correct, but this is not expressly stated. ISO 19005-2 expands this sentence to include the following statement: “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.”

We therefore reiterate the following preservation planning measures set out in October 2014:

4 Preservation planning measures

The problem described must be addressed on four levels simultaneously.[11] The following measures were therefore set in train in October 2014:

A. The creators of the document analysed[12] are informed of the problem and requested to investigate the reproducibility of the error.
B. The manufacturer of the PDF/A converter is contacted and requested to embed only valid TrueType fonts.
C. The manufacturers of the validators tested are contacted and requested to expand testing of the embedded content so that PDF/A documents with corrupt content are not identified as valid.
D. The secretariat of ISO/TC 171/SC 2/WG 5, which is responsible for ISO 19005, is notified and requested to incorporate the addition to the sentence in ISO 19005-2 6.2.11.2 (“as well as to the font specifications referenced by these provisions”) into ISO 19005-1 6.3.2, either by means of a “corrigendum” or otherwise.[13]

Measure C is necessary so that PDF/A documents of this type can be identified and corrected. Since the application used to create the document is a current one, Measure B can reduce the incidence of such PDF documents. Measure D is designed to ensure that the manufacturers of PDF/A software either do not produce this error or update their validators accordingly.

Archives will not be able to identify and correct the PDF documents affected until the PDF/A validators recognise the error.

5 Position statement

Since the first enquiry made in November 2014 unfortunately did not yield the desired results, KOST now requests software manufacturers and ISO/TC 171/SC 2/WG 5 to respond in writing by 31 August 2015. The manufacturers of the validators tested are also expected to supply a precise description of which embedded content is not validated, and the reasons for this.

[11] All four levels must be addressed simultaneously because in our experience the levels involved support each other and it is not expedient to address only one of them.
[12] Both the author Cassie Findlay and the publishing institution, the Dutch National Archives.
[13] The further clarification in ISO 19005-2 6.2.11.2 should have been incorporated into Version 1 by means of a corrigendum. It is not sufficient to address the problem only in the new versions, as the previous versions explicitly remain valid and are intended to do so.