

Lost in migration: document quality for batch conversion to PDF/A

Roland Erwin Suri and Mohamed El-Saad
ETH Library, ETH Zurich, Switzerland

Document
quality for batch
conversion to
PDF/A

Received 31 October 2017
Revised 24 April 2018
Accepted 24 April 2018

Abstract

Purpose – Changes in file format specifications challenge long-term preservation of digital documents. Digital archives thus often focus on specific file formats that are well suited for long-term preservation, such as the PDF/A format. Since only few customers submit PDF/A files, digital archives may consider converting submitted files to the PDF/A format. The paper aims to discuss these issues.

Design/methodology/approach – The authors evaluated three software tools for batch conversion of common file formats to PDF/A-1b: LuraTech PDF Compressor, Adobe Acrobat XI Pro and 3-Heights™ Document Converter by PDF Tools. The test set consisted of 80 files, with 10 files each of the eight file types JPEG, MS PowerPoint, PDF, PNG, MS Word, MS Excel, MSG and “web page.”

Findings – Batch processing was sometimes hindered by stops that required manual interference. Depending on the software tool, three to four of these stops occurred during batch processing of the 80 test files. Furthermore, the conversion tools sometimes failed to produce output files even for supported file formats: three (Adobe Pro) up to seven (LuraTech and 3-Heights™) PDF/A-1b files were not produced. Since Adobe Pro does not convert e-mails, a total of 213 PDF/A-1b files were produced. The faithfulness of each conversion was investigated by comparing the visual appearance of the input document with that of the produced PDF/A-1b document on a computer screen. Meticulous visual inspection revealed that the conversion to PDF/A-1b impaired the information content in 24 of the converted 213 files (11 percent). These reproducibility errors included loss of links, loss of other document content (unreadable characters, missing text, document part missing), updated fields (reflecting time and folder of conversion), vector graphics issues and spelling errors.

Originality/value – These results indicate that large-scale batch conversions of heterogeneous files to PDF/A-1b cause complex issues that need to be addressed for each individual file. Even with considerable efforts, some information loss seems unavoidable if large numbers of files from heterogeneous sources are migrated to the PDF/A-1b format.

Keywords Digital documents, Digital libraries, Academic libraries, Archives, Conversion, Digital preservation
Paper type Technical paper

Introduction

If one opens an old digital document, its visual presentation has sometimes slightly changed. Vector graphics, special characters, text formats, links and other features may have deteriorated due to changes in the software that is used to view the file. To avoid such degradation of the file content, digital archives usually favor a format migration strategy. Although it is unclear whether digital content can be preserved over the long term, it seems clear that bridging a time span of 20 years typically requires substantial efforts. In a large migration project, thousands of files that had been produced over the previous 20 years in the formats Lotus 1-2-3, CDF, netCDF and HDF were migrated to current versions of these formats (Frisz, Brown and Waggoner, 2012). Most of these files were converted with standard software, but a fraction of files could not, or not perfectly, be converted, because some of the used features were not supported by the new format version. The authors wrote for this migration project almost 2,000 lines of C-Code. Most of this code was needed to find the unsupported file features, since off-the-shelf conversion software did usually not provide warning messages when features could not be converted.

To preserve digital documents containing text and images, archives often favor migration to the PDF/A format. The PDF/A format differs from the common PDF format by



prohibiting some features that are ill-suited for long-term preservation. The most common subtype of PDF/A is PDF/A-1b, which is especially designed for long-term preservation of the visual appearance of documents. PDF/A-1b is based on the Adobe Acrobat PDF 1.4 format and restricted by the long-term preservation Standard ISO 19005-1 (Sullivan, 2006; Evans and Moore, 2014; Klindt, 2017). In a review about the level of trust placed by archives in document file formats for long-term preservation, the PDF/A format took the fourth place, only superseded by the quite limited formats MARC, TXT and XML (Rimkus *et al.*, 2014). Despite these successes, the suitability of PDF/A as an archival format has been debated (Klindt, 2017). Among other issues, automated text extraction is difficult because the text is interrupted by other content, such as page numbers, headers, and footers. Furthermore, PDF/A-1b requires embedding fonts in the document, such that licensing problems may arise that lead to embedding the most similar font that happens to be available. Nevertheless, even this author does currently not see a viable alternative to PDF formats for preservation of digital documents.

Thus, archives often recommend to their customers converting their Word and PowerPoint documents to PDF/A-1b. Some archives even require submissions in the PDF/A format, such as the archives at the German universities Marburg, Potsdam and Münster (Müller, 2015) and the Swiss Bundesarchiv[1]. Furthermore some governments require the PDF/A format for government documents, including the UK[2]. Unfortunately, conversion to PDF/A-1b can be challenging. To advise customers how to convert their documents to PDF/A, the ETH Library conducted some preliminary tests. A set of five Microsoft Word and five Microsoft PowerPoint files with complex scientific content was selected and converted to PDF/A-1b using four commonly used conversion methods. Unfortunately, conversion errors were quite common and included broken hyperlinks, unacceptable figures, failures to write files and violations of the PDF/A-1b standard[3].

Since customers often struggle with the conversion of their documents to PDF/A, archives may consider converting documents themselves. This requires conversion programs that are able to convert large batches of files with minimal human interference. Evans and Moore (2014) describe a large-scale effort in converting PDF files to PDF/A-1b for digitized and born digital reports from the OASIS system in Scotland and England. They noticed that conversion of PDF to PDF/A-1b with Adobe Acrobat required serious manual intervention. Although PDF/A Manager (by PDFTron) facilitated batch processing, about 20 percent of the produced PDF/A-1b files were considered invalid. In a similar study, Koo and Chou (2013) analyzed the performance of pdfaPilot, 3-HeightsTM and PDF/A Manager using 80 test files. They meticulously tracked down the reported validation errors using the various documents that define the PDF/A-1b standard. This requires advanced expert knowledge and substantial time resources even for this small sample. It turned out that some of the substantial discrepancies between the validators was caused by different interpretations of the standards. About 20 percent of the produced files were truly non-valid PDF/A-1b files. Furthermore, several files converted by 3-HeightsTM and by PDF/A Manager showed visual differences in fonts that hampered the readability of the document. Unfortunately, these substantial errors were not detected by the file validation because they did not violate PDF/A standards. In a more recent product review by KOST (2016), the capabilities of seven software products were investigated for converting to PDF/A-2b. The PDF/A-2b format is very uncommon, but should cause fewer problems with transparent objects than conversion to PDF/A-1b (Debenath *et al.*, 2013). The study reviews costs, license models, operating systems and document quality for these software products. The KOST study used 288 non-corrupt test files of common file types. The software products failed to convert between 4 percent (dmstools) and 28 percent (Foxit) of the files to PDF/A-2b. The faithfulness of the conversions was assessed by visually comparing the first page of the converted

documents with that of the original documents. Between 0 percent (Neevia) and 11 percent (Foxit) of the produced files either contained image compression artefacts or had lost some other information.

For the current study, we were particularly interested in recording the errors pertaining to the visual reproducibility, as they directly affect the usability of the files. Since files migrated to the PDF/A format are often valid but still do not show the original content (Drümmer, 2007; Koo and Chou, 2013), we checked all pages of all our test files for errors in visual reproducibility. Since catching these errors is time consuming, the current study uses a sample of 80 files to investigate the reliability and visual reproducibility of PDF/A-1b conversions.

Characteristics of test files

Some statistics of the test files are summarized in Table I. We selected formats and files to reflect the characteristics of files in our digital archive that may be convertible to PDF/A. We received most test files from a university archive interested in PDF/A conversion. In order to broaden the variability between test files, some further files with scientific content originated from public and personal sources. The file complexity varied from single bitmap images to PhD theses. All but one example file had been created on the Microsoft Windows operating system because this is the most common operating system at our university. Most files shown in the figures and a detailed numerical evaluation on the level of single files are publicly available for download (Suri, 2017).

Software and settings

We tested three software tools with batch processing capabilities on the Microsoft Windows 7 operating system: LuraTech PDF Compressor (LuraTech), Adobe Acrobat XI Pro and 3-Heights™ Document Converter (PDF Tools AG)[4]. The three tested software tools offer a large variety of settings that can be optimized by the user. Unfortunately, the best settings are usually difficult to know and may be specific for the input files. We kept the standard settings unless adaptations were necessary to receive comparable results. In particular, the optical character recognition (OCR) was disabled for all software products due to a large variety in font licenses and language settings. OCR settings may influence computing time but should have no influence on the visual appearance of the documents.

	Total size of example files (MB)	Total number of pages	File dates	Total number of files	Files with bitmaps	Files with vector graphs	Files with special text characters	Files with links
JPG	20.9	10	2000-2016	10	10	0	0	0
PDF	30.4	1346	2004-2016	10	3	2	10	2
PNG	2.6	10	2008-2016	10	10	0	0	0
MS Word (5 files *.doc, 5 *.docx)	6.6	119	2002-2016	10	3	1	2	2
MS Excel (8 *.xls, 2 *.xlsx)	7.6	na	1998-2016	10	0	3	0	1
MS PowerPoint (9 *.ppt, 1 *.pptx)	36.6	173	2003-2016	10	9	6	1	1
E-mail (*.msg, no attachments)	0.5	na	2002-2016	10	1	0	0	10
MS Webseiten (7 *.html, 2 *.mht, 1 *.htm)	10.6	na	1999-2016	10	1	0	0	9

Table I.
Characteristics
of the 80 test files

LuraTech PDF Compressor

We tested PDF Compressor Enterprise version 7.4.02.19 with Born Digital Modul by LuraTech (now with the company Foxit Europe). The resolution maximum was raised from 300 dpi to 1000 dpi to avoid image degradation. We followed the recommendations of the handbook and selected the “MRC Check” document class for complex objects in order to avoid the loss of image resolution.

In order to avoid any stops in batch processing, we selected the preference “Continue job on critical error.” Nevertheless, the program sometimes stopped in the middle of a job without providing an explanation.

The company’s support was helpful. Questions were answered within 24 hours.

Adobe Acrobat XI Pro

For creation and validation of PDF files, Adobe Acrobat XI Pro used the software component Preflight (version 11.0). Preflight is a product of the company Callas Software.

We adopted the recommendations of the “Staatsarchiv des Kantons Zürich” (Staatsarchiv, 2013) for the Adobe Acrobat XI Pro settings, which we briefly summarize here. For each installation of Adobe Acrobat XI Pro, the following settings had to be set only once. In the tab Edit, the menu Settings was selected to choose “Microsoft Office Word” (and all the other file types tested here) from the list “Convert to PDF.” The output format PDF/A-1b:2005 (RGB) was chosen. With the button Edit, one reaches the settings for the PDF printer (except for *.png). In an attempt to avoid any stops during processing of the files, we selected for the compliance standard “PDF/A-1b” on the line “When not compliant” the drop down menu “Continue.” Nevertheless, occasional stops occurred during batch processing.

We did not contact Adobe Systems for support.

3-HeightsTM Document Converter

The SME Edition of the 3-HeightsTM Document Converter version 4.8.3.0 from PDF Tools AG was tested. To get detailed error messages and a summary of the errors, we set “PDFA.ERROR = true,” “PDFA.Logsummary = true” and “PDFA-LOGDETAILS = true.” In order to get a warning if embedded files were deleted, PDF.WARNCOLL was set to true.

The support was very helpful with questions that came up during installation.

Software performance

For reasons that usually remained unclear to us, the software tools did not convert all the files to PDF/A-1b. Table II shows the number of files for which the software failed to produce a PDF/A-1b file.

	LuraTech	Adobe Pro	3-Heights TM
JPG	0	0	0
PDF	0	0	0
PNG	1	1	0
MS Word	0	0	1
MS Excel	1	0	2
MS PP	0	0	0
E-mail	0	na ^a	0
Web pages	5	2	4
Total	7	3	7

Table II.
Number of files that
were not converted
to PDF/A-1b

Note: ^aAdobe Acrobat XI Pro does not convert e-mail

Table III shows indicative values of average computing time per file and average change in file size. Please be aware that these numbers are strongly influenced by the selected settings (such as image resolution) and by a few large files (see the section on conversion of Excel files below). Furthermore, not all files were produced by all three software products. The computing time is also influenced by the used computer (Windows 7, Dell Optiflex desktop, CPU Intel i7 with eight processors, Intel HD Graphics) and by processes running in the background. The Excel file in the supplementary materials (Suri, 2017) gives some more details. The table also shows the number of stops in program execution during batch conversion of the files. Since each stop required user interaction, these stops hamper processing of large numbers of files. The time needed for user interaction was not counted as computing time.

Errors in reproducibility

The produced PDF/A-1b files were visually compared with the original files to find reproducibility errors that interfere with the scientific content that is shown in the document. Furthermore, at least one page of text was checked for errors vs the original text by automated text comparison routines. We did not use software to compare graphs with pixel-wise subtraction because this method is quite sensitive to tiny differences in location, color and graph size. We assumed that the PDF/A-1b file was supposed to show the information that the creator most likely viewed when he saved the file in the original application. Each of the 213 created PDF/A-1b files was opened with Adobe Acrobat XI Pro and all pages were visually compared side by side with those of the corresponding original file. The latter was usually opened in the current version of the original application (using the Microsoft Windows 7 operating system with Microsoft Office Professional Plus, 2013).

Conversions from PDF to PDF/A-1b

Original PDF files and produced PDF/A-1b files were visually compared with Adobe Acrobat XI Pro.

In a large and complex PDF 1.5 file (a PhD thesis), we found that the conversion with 3-Heights™ did not faithfully convert a vector graphic: transparent objects in the figure had become in-transparent in the PDF/A-1b file, thereby reducing the information content of the figure (Figure 1). For both other software tools, this error did not occur. The PDF/A-1b format does not allow multiple layers. For successful PDF/A-1b conversion, foreground and background need to be merged properly, which apparently failed in this case.

In another file, Adobe Acrobat XI Pro converted a PDF 1.6 File to PDF/A-1b without any error messages. Nevertheless, the text contained occasional spelling errors (Figure 2). For both other software tools, this error did not occur.

Conversion from Microsoft Word to PDF/A-1b

Microsoft Word files were opened in Microsoft Word 2013. Older files automatically opened in compatibility mode. They were visually compared with the converted PDF/A-1b files (opened with Adobe Acrobat XI Pro).

	LuraTech	Adobe Pro	3-Heights™
Average computing time per file	3 sec	73 sec	8 sec
Total change in file size	+52%	+129%	-39%
Number of stops in batch program	3	4	3

Table III.
Statistics of batch
conversion to PDF/A-
1b for the three
software tools

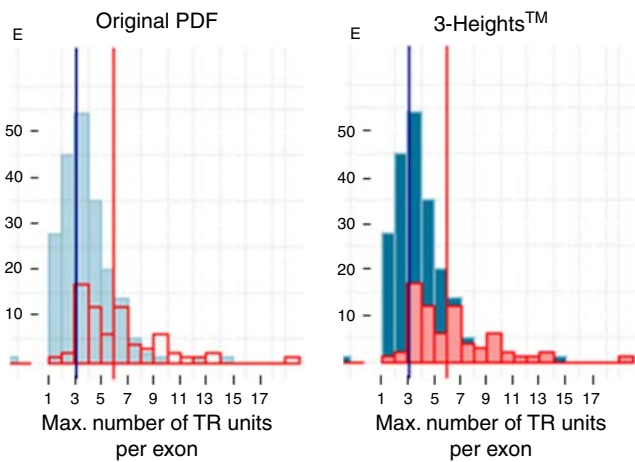


Figure 1.
Red bars lost transparency

Notes: A bar graph in the original PDF (left) is compared with the corresponding graph in the PDF/A-1b file converted by 3-Height™ (right). Unfortunately, the red bars in the foreground had lost their transparency and hid some of the blue bars in the background

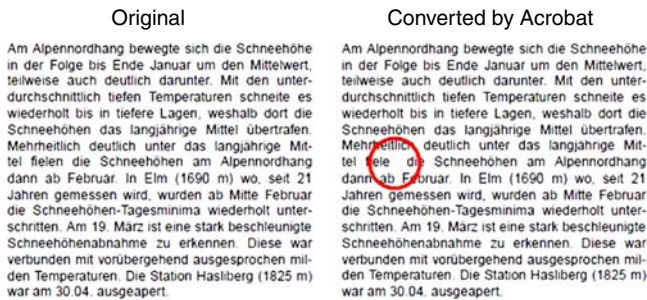


Figure 2.
Spelling error by
acrobat conversion

Notes: The file contained occasional spelling errors if it was converted with Adobe Acrobat XI Pro. The word “fielen” in the original file (left) became “fielen” in the PDF/A-1b file (right)

A Microsoft Word file created in 2014 (extension docx, ISO 29500:2008-2012) showed a semi-transparent object. Unfortunately, the background text became invisible when the file was converted with 3-Heights™ (Figure 3). For both other software tools, this error did not occur.

A Microsoft Word 97-2003 file, saved in 2008, contained some change requests by its reviewers and also included the reviewers’ comments. When the file was opened in Word 2013, the tracked changes and the reviewers’ comments were shown. However, once this file had been converted to PDF/A-1b by LuraTech PDF Compressor, the recommended changes were shown as if they had been accepted. They were shown without mark-ups and without the reviewers’ comments (Figure 4). For both other software tools, the PDF/A-1b document looked just like the original Microsoft Word document.

Another Word file lead to reproducibility errors that were common to all three conversion tools. In a Microsoft Word document that was last saved in 2004 by Word

97-2003 the lines in a table were partially interrupted when the PDF/A-1b files were inspected by Adobe Acrobat XI Pro (Figure 5). These lines had not really disappeared but were only too fine to be shown. For all three PDF/A-1b files, the lines were only shown correctly if the view in Adobe Acrobat was zoomed to a magnification of at least 400 percent.

In the same Word document, the footer information (containing path and date when the file was stored) was updated for all three software tools (Figure 6). As the original Word document was last saved in the year 2004, a user would expect that the date in the footer reflects this year. However, the PDF/A-1b document shows the date when the file was converted to PDF/A-1b. Note that when viewing the original Word document (in MS Word 2013), the date was automatically updated to the current date too. The original view shown in the figure on the left was reproduced by opening the Word document in protected view. In contrast to the date, the path to the folder where it had originally been stored is still correctly shown when viewing the Word document. Unfortunately, in the PDF/A-1b files the path had been updated to the directory in which the conversion to PDF/A-1b had taken place.

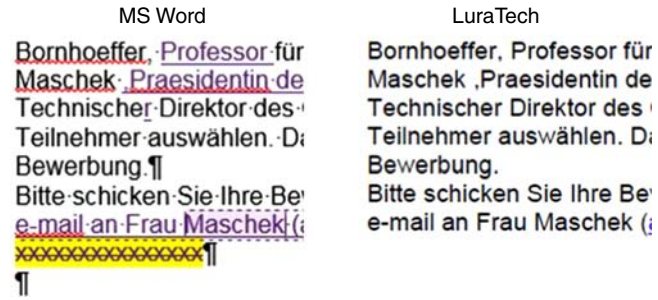
Conversion from MS Excel to PDF/A-1b

Excel files often contain lines, rows or columns with special formatting, such as background colors. This formatting usually extends until the end of the Excel sheet. Our sample of ten

Document
quality for batch
conversion to
PDF/A



Figure 3.
Smiley lost
transparency
when converted
by 3-Heights™



Notes: When a Word document (left) with corrections by the reviewers was converted by LuraTech's PDF Compressor (right), all mark-ups were removed and the reviewers' comments were accepted

Figure 4.
Mark-ups removed
by LuraTech

Word				PDF/A-1b			
Kopie Nr.	Signatur	Bezeichnung des Originals	Anzahl Kopien	Kopie Nr.	Signatur	Bezeichnung des Originals	Anzahl Kopien
1				1			

Figure 5.
Some horizontal lines
became invisible with
all three PDF/A
converters

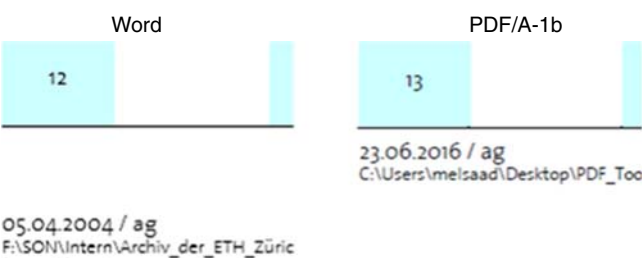


Figure 6.
Footer became
updated

Notes: Word document (left) was converted to PDF/A-1b (right).
Unfortunately, file date and file storage location in the footer had
both been updated

Excel files contained four files with such formatting extending until the end of the sheet. This lead to long computing times for Adobe Acrobat XI Pro (18 minutes for ten files) but not for LuraTech (0.6 minutes for nine files) or 3-Heights™ (0.6 minutes for eight files). The file size most dramatically increased for Adobe Acrobat. It increased from 7.6 MB of the original ten files to 134 MB for the produced PFD/A-1b files. This was the main reason for the large increase in total file size for Adobe Acrobat shown in Table III. For the PDF Compressor by LuraTech a formatted row in an Excel sheet led to 100,000 empty pages in the PDF/A-1b File. Apparently, in an attempt to reduce the file size, the PDF Compressor dramatically reduced the sheet sizes. Unfortunately, for three Excel files this lead to unreadable characters (Figure 7).

Conversion from e-mail to PDF/A-1b

Original e-mails were opened with Microsoft Outlook Professional Plus 2013 and visually compared with the PDF/A-1b file opened in Adobe Acrobat XI Pro.

A Microsoft Outlook document, created in the year 2016, contained several active hyperlinks. Unfortunately, some (but not all) of the hyperlinks were lost if the PDF/A-1b document was created by 3-Heights™ (Figure 8). All hyperlinks were still available when the e-mail was converted with the LuraTech tool.

Another Microsoft Outlook document, created in 2003, contained a table that was not faithfully reproduced when the file was converted by LuraTech to PDF/A-1b (Figure 9). The black lines in the table had disappeared and the indentations of the table elements had changed. For both other software tools, this issue did not occur. Since these minor differences hardly impair the information content, they were not counted as errors.

MS Excel		LuraTech	
Bestellungen Nicht Excel		Bestellungen Nicht Excel	
9783642297663	Qin Q. Advanced Mechar	9783642297663	Qin Q. Advanced Mechar
9783642327612	Rapp D. Use of Extraterres	9783642327612	Rapp D. ** ** *
9789400754065	Rovida E. Machines and Sigr	9789400754065	Rovida E. ** ** *
9783642324888	Takayama K. Professor I. I. Glas	9783642324888	Takayama K. Professor I. I. Glas
9781447145639	Williams J.J. Introduction to Ar	9781447145639	Williams J.J. ** ** *

Figure 7.
Some text became
unreadable

Notes: When a Microsoft Excel table (left) was converted to PDF/A-1b by LuraTech PDF Compressor (right), some text became unreadable (right, zoomed by 1,000 percent from actual size). The text was not completely lost, as it could be recovered by selecting the text and pasting it to a text editor



Notes: An e-mail written in Microsoft Office contained several active hyperlinks shown in blue font color. The hyperlink underlying the word “Privacy Policy” works properly in the original Microsoft Outlook hyperlink (left, mouse pointer was placed over hyperlink). Unfortunately, this link had been lost in the PDF/A-1b document created by 3-Heights™ (right)

Document
quality for batch
conversion to
PDF/A

Figure 8.
Underlying
hyperlink was lost

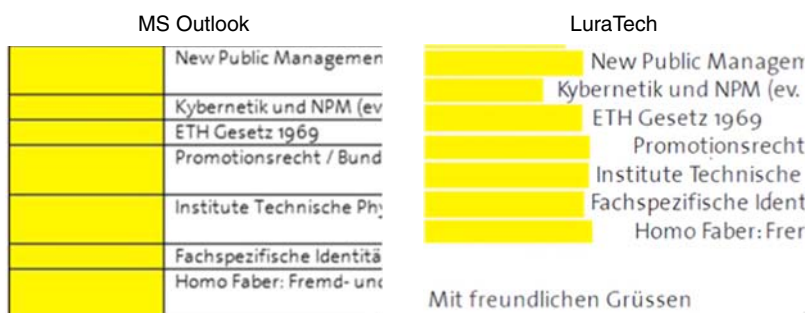


Figure 9.
Black lines
disappeared when
converted by
LuraTech

Conversion from web pages to PDF/A-1b

For the previous file formats we opened the original file with a “standard” software. Since there is no standard software for opening web pages, we compared PDF/A-1b files with their visual appearance in three different browsers. Web pages were opened in internet Explorer 11, Mozilla Firefox 49 and Google Chrome 51 and were compared with the converted PDF/A-1b document. We ignored differences if at least one of the three browsers showed the same shortcomings as the PDF/A-1b file.

An HTML webpage saved in the year 2010 and written in Hypertext Markup Language version 5 looked identical in Internet Explorer, Mozilla Firefox or Google Chrome. Also the PDF/A-1b created by Adobe Acrobat XI Pro looked exactly like the original. Unfortunately, when a PDF/A-1b was created by LuraTech or 3-Heights™, font colors and line breaks changed (Figure 10). These minor differences were not counted as errors.

Overview of reproducibility errors

Both following tables summarize the errors found by visual comparison between the original file and the converted PDF/A-1b file. We classified conversion errors according to the error types shown in Table IV. If the same type of error occurred several times in one file, it was only counted once. The last item in Table IV gives the number of files with one or more errors. Since some files contained more than one type of error, these numbers in the bottom line are often smaller than the sum of the numbers above.

We counted these changes in the visual appearance as errors because they hamper in our opinion the usability of the document. However, changes that did not impair the information content were not counted as errors. These neglected issues include:

- additional spaces in text;



Figure 10.
HTML webpage
looked different

Notes: When an HTML webpage (top) was converted to PDF/A-1b with LuraTech (middle) or 3-Heights™ (bottom), the coloured characters became black, the grey background became white, and the text “In [35]:” got misplaced

Table IV.
Reproducibility errors
found in the files
converted to PDF/A-
1b for each of the
three software tools

	LuraTech	Adobe Pro	3-Heights™
Lost links	0	1	1
Loss of other document content (unreadable characters, missing text, part of document missing)	7	3	3
Updated fields (reflection date or storage location for conversion)	3	4	4
Vector graphics error (including transparency)	0	0	2
Spelling errors	0	2	0
Files with one or more of above errors	9	8	7

- changes in color;
- loss of black lines in tables (such as in Figure 9);
- changes in the location of page breaks or in the arrangement of the objects (such as in Figure 10);
- huge increases in file size and page numbers that sometimes happen when Excel files get converted; and
- changes in the structural information of the PDF File (thumbnails, bookmarks) or metadata.

Archives may wish to estimate the number of conversion errors for their own file population. Such an estimate should be based on the number of converted document features. We thus summarize the reproducibility errors (as defined in Table IV) with respect to the document contents:

- For the 37 files with bitmaps, all bitmaps were converted faithfully.
- For 12 files with vector graphs, two conversions contained reproducibility errors. Both had been converted by 3-Heights™.

- For the 13 files with special characters, all the special characters were faithfully converted. However, two converted files contained spelling errors. Both had been converted by Acrobat.
- Links occurred in 25 input files and were lost in two conversions, one by Adobe and one by 3-HeightsTM.
- Between three and four files per software contained some updated fields such that a new date or folder (related to the PDF/A conversion) was written into the document.

Document
quality for batch
conversion to
PDF/A

Table V shows for each file type the number of files with reproducibility errors. The ten JPG and the ten PNG files in our test set usually contained single pictures that all got converted without any reproducibility issues. We also noticed hardly any reduction in terms of image quality presumably due to settings that were sufficiently generous regarding the size of the produced PDF/A-1b files. Reproducibility errors were not very frequent for PDF (two errors for 30 converted files) and PowerPoint files (three errors for 30 converted files). However, reproducibility errors occurred quite frequently in Microsoft Word (five errors for 29 converted files) or Excel files (11 errors in 27 converted files), as in these file types many objects and complex featured had to be converted to PDF/A-1b.

Disagreements between PDF/A-1b validators

The current study investigates errors in visual reproducibility for migration to PDF/A-1b. As explained in the introduction, these errors are usually not related to the validity of the PDF/A-1b file. Nevertheless, we used three validator software tools[5] to check whether the produced PDF/A-1b files were consistent with the PDF/A-1b format:

- Preflight Module of Adobe Acrobat XI Pro;
- PDF Tools 3-HeightsTM Validator (included in KOST-Val version 1.7.4[6]); and
- PDFTron Validator (included in KOST-Val version 1.7.4).

Table VI shows the number of non-compliant PDF/A-1b files as reported by these three PDF/A validators. Although we did not further analyze this result, we noticed that

	LuraTech	Adobe Pro	3-Heights TM
JPG	0	0	0
PDF	0	1	1
PNG	0	0	0
MS Word	2	1	2
MS Excel	5	4	2
MS PP	1	1	1
E-mail	1	na	1
Web pages	0	1	0

Table V.
Number of files with
reproducibility errors
for each of the eight
file types

	Number of produced PDF/A-1b files	Preflight Module of Adobe Acrobat XI Pro	PDF Tools 3-Heights TM Validator	PDFTron Validator
LuraTech	73 files	2	2	11
Adobe Pro	67 files	0	8	7
3-Heights TM	73 files	9	2	25

Table VI.
Numbers of non-
compliant PDF/A-1b
files according to
three validator
software tools

validation errors are quite common and that validators often disagree. For a thorough investigation of PDF/A validators, we refer to previous studies (Koo and Chou, 2013; Evans and Moore, 2014; KOST, 2017).

Discussion

The current study gave for the first time a qualitative and quantitative overview of the various errors in the visual content that are typically caused by migration of documents to PDF/A. Meticulous inspection revealed that reproducibility errors reduced the information content for 24 files out of 213 files produced PDF/A-1b files. Spotting these reproducibility errors is only possible by time-consuming visual inspection. Large-scale migration of heterogeneous files to PDF/A-1b thus faces a trade-off between consumption of substantial human resources and degeneration in the document content. Migrating of documents that were produced over the last 20 years to the more sustainable PDF/A-1b format leads to small but noticeable degradation of visual content.

For 230 attempted conversions to PDF/A-1b, the software tools failed to produce PDF/A-1b files in 17 cases. Furthermore, the batch processing stopped 10 times requiring manual interference. Large-scale batch conversions of heterogeneous files to PDF/A-1b thus cause a large variety of complex issues that need to be addressed for each individual file, presumably by collaborating with the software vendor. The decision for a converter software will thus not only depend on the performance of the software tool, but also on the reliability of the software vendor support. Unfortunately, the results of the KOST (2016) study suggest that no software product has solved the reproducibility problems studied here. Among seven converter tools tested by KOST, the 3-HeightsTM converter was among the best three tools in terms of reproducibility errors (2 percent of files with errors). Nevertheless, in the current study, the 3-HeightsTM converter had similar reproducibility errors as the other tools (10 percent of files with errors). Although Neevia's Document Converter Pro performed in the KOST (2016) study substantially better in terms of reproducibility errors (0 percent of files with errors), it only converted 78 percent of the files.

Further results can be compared with those of the KOST (2016) study. That study found that 22 and 13 percent of the files were not converted to PDF/A by Adobe Acrobat and PDF Tools (3-HeightsTM), respectively. These numbers were lower for the current study with 3.7 percent and 8.8 percent of files that were not converted. For the files converted by Adobe Acrobat and 3-HeightsTM, the KOST study reported that the visual reproducibility of the documents was impaired in 5 and 2 percent of the files, respectively. The percentages we found in the current study were much higher (12 and 10 percent). These differences may be explained by different software versions, different test files, or the more rigorous visual inspection that was used here. Whereas we visually checked all file contents, the 2016 KOST study only checked the first page of the documents. They checked whether the visual appearance of a file was sufficient, whether there were no compression artefacts in images, and whether all content was still available. In addition to these differences, in the KOST study files were mostly converted to PDF/A-2b, whereas they were converted to PDF/A-1b here. The format PDF/A-2b does not prohibit transparency, such that there should be no errors due to semi-transparent objects (KOST, 2013). We found in the current study two errors due to semi-transparent objects (Table IV, Figures 1 and 3), both with the product by PDF Tools (3-HeightsTM).

Archives and their customers often wonder who should convert the documents to PDF/A. Customers are in a somewhat better position than archives to check the faithfulness of PDF/A conversions, as they know how their document should look like. Although archives can acquire professional software that allows for batch processing, these software tools do not solve the main issues with the conversion: the reproducibility

errors of the other tools were similar to those of the popular Adobe Acrobat XI Pro. In addition, there are also further widely-available options with similar performance not using Adobe Acrobat XI Pro (see footnote 3). Taken together, we recommend customers to perform the conversion of single files. Nonetheless, if a customer provides a multitude of files of the same origin, the setup and experience of the archive may favor conversion by the archive.

In the current study, three PDF/A validators were used to check whether the produced files really adhered to the criteria of the PDF/A-1b standard. The three PDF/A validators often claimed that the produced PDF/A-1b files were not valid PDF/A-1b documents. Unfortunately, there was substantial disagreement between these assessments. Also previous studies reported a large percentage of supposedly invalid PDF/A files (Müller, 2015; KOST, 2017). To reduce the number of invalid files, the KOST study required that two validators (PDF Tools 3-HeightsTM Validator and PDFTron Validator) agreed in order to consider a file to be invalid. Even with this stringent criterion, invalid PDF/A files were still common (13 percent for conversion by Adobe Acrobat and 9 percent for conversion by 3-HeightsTM). Since error messages are usually rather cryptic, it would be difficult and time consuming to figure out the reasons for alleged non-conformances with PDF/A standards. The effort of figuring out and fixing all validation errors is often not feasible for archives. Lists of validation errors reported by the PDF validator (JSTOR/Harvard Object Validation Environment (JHOVE)) indicate that for more than half of the PDF validation errors it is neither clear whether they impair long-term preservation nor how the files could be corrected[7]. Although many archives check all incoming files with validators, they often resort to setting up rules such that some of the non-specific validation errors are automatically ignored (Geisser, 2017). Nevertheless, validation errors that are common to many files from a single file producer need to be investigated. Institutional file producers are often able to correct such errors by changing the settings of programs that produce the files.

The German Federal Archives (Bundesarchiv) automatically convert heterogeneous files of various common file formats to PDF/A[8]. The Federal Archives state that factors such as missing fonts, macros, hyphenation, transparent objects, embedded objects and reduced image resolution sometimes impair the quality of their document conversions. This is consistent with our results suggesting that reproducibility errors can hardly be avoided when converting to PDF/A. Customers are often advised, or even required (Müller, 2015), by archives to convert their documents to PDF/A-1b even though they face similar problems. Archives may thus be forced to adopt a more tolerant policy with regards to the acceptance of various file formats even though this only postpones difficulties with the long-term preservation of documents.

Notes

1. Archivtaugliche Dateiformate, Standards für die Archivierung digitaler Unterlagen, Schweizerisches Bundesarchiv, January 2014, version 2014/01, available at: www.bar.admin.ch/dam/bar/de/dokumente/konzepte_und_weisungen/archivtaugliche_dateiformate.1.pdf.download.pdf/archivtaugliche_dateiformate.pdf (accessed February 20, 2018).
2. Viewing government documents, GOV.UK Cabinet Office, August 29, 2017, available at: www.gov.uk/government/publications/open-standards-for-government/viewing-government-documents (accessed May 22, 2018).
3. Conversion of Microsoft Word or PowerPoint Files to PDF/A-1b “ETH-Bibliothek, Digital Curation Office”, available at: <https://documentation.library.ethz.ch/pages/viewpage.action?pageId=6685488>. (accessed May 22, 2018).

4. Although we intended to test Callas pdfToolbox, iText, and dp4Convert, our IT service could not install them on a Windows 7 desktop computer. Installation of a test version of Callas pdfToolbox (www.callassoftware.com) turned out to be cumbersome: even after many weeks of frequent e-mail exchanges with their customer support, we had not managed to get a properly running trial version of the software. We did not receive much support from iText (<https://itextpdf.com>) and an installation apparently required the installation of a Java or “.NET” development environment. A full version of dp4Convert from SEAL Systems AG (www.sealsystems.de) can only be installed on a server.
5. VeraPDF (verapdf.org) was not tested because only a prototype was available in summer 2016. JHOVE is commonly used for PDF validation but was not used here, as it does not test all requirements of PDF/A (<http://jhove.openpreservation.org/modules/pdf>, accessed May 22, 2018). FITS uses JHOVE for PDF validation (<https://projects.iq.harvard.edu/fits/tools>, accessed May 22, 2018).
6. Validator für TIFF-, SIARD-, PDF/A-, JP2- und JPEG-Dateien sowie SIPs (KOST-Val), KOST-Val v 1.7.4, available at: http://kost-ceco.ch/cms/index.php?kost_val_de (accessed May 22, 2018) and <http://github.com/KOST-CECO/KOST-Val> (accessed May 22, 2018).
7. JHOVE errors list, Tab PDF-hul, from Open Preservation Foundation, available at: <https://docs.google.com/spreadsheets/d/1zyg4eqH6akoehI10fNhzroaW3CgcwUjkNlvvVZvFDyo/edit#gid=0> (accessed May 22, 2018); and a similar list by Yvonne Friese and Becky McGuinness in OPF knowledge base wiki, available at: <http://wiki.opf-labs.org/display/Documents/JHOVE+issues+and+error+messages> (accessed May 22, 2018).
8. Aussonderung digitaler Unterlagen und deren Archivierung im Bundesarchiv. Ein Leitfaden, Bundesarchiv 2010, version 1.2, available at: www.bundesarchiv.de/imperia/md/content/abteilungen/abtb/bbea/behoerdenleitfaden-v1.2-2010-08-27-internet.pdf (accessed February 20, 2018).

References

- Debenath, O., Merzaghi, M. and Röthlisberger, C. (2013), “PDF/A-2 und PDF/A-3: was ist neu?”, KOST Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen, April 10, 15pp., available at: <https://kost-ceco.ch/cms/download.php?000c45fdeb8ba0cc368d4c286e048c82> (accessed October 24, 2017).
- Drümmer, O (2007), “PDF/A – a look at the technical side”, PDF/A Competence Center, September 22, 9pp., available at: www.pdfa.org/wp-content/uploads/2011/08/pdf-a_a_look_at_the_technical_side-2b.pdf (accessed May 17, 2018).
- Evans, T.N.L. and Moore, R.H. (2014), “The Use of PDF/A in digital archives: a case study from archaeology”, *International Journal of Digital Curation*, Vol. 9, No. 2, pp. 123-138, available at: <http://dx.doi.org/10.2218/ijdc.v9i2.267> (accessed May 22, 2018).
- Frisz, C., Brown, G. and Waggoner, S. (2012), “Assessing migration risk for scientific data formats”, *The International Journal of Digital Curation*, Vol. 7 No. 1, pp. 27-38, available at: www.ijdc.net/index.php/ijdc/article/view/202/271; <https://doi.org/10.2218/ijdc.v7i1.212>
- Geisser, F. (2017), “Handling file format issues in Rosetta – and beyond”, Rosetta Advisory Group 8th Annual Meeting, Sheffield, available at: www.youtube.com/watch?v=nyNaNJDeB14 (accessed May 22, 2018).
- Klindt, M. (2017), “PDF/A considered harmful for digital preservation”, iPRES, available at: <https://ipres2017.jp/wp-content/uploads/15.pdf> (accessed May 17, 2018).
- Koo, J. and Chou, C.C.H. (2013), “PDF to PDF/A: evaluation of converter software for implementation in digital repository workflow”, *New Review of Information Networking*, Vol. 18 No. 1, pp. 1-15, doi: 10.1080/13614576.2013.771989 (accessed May 22, 2018).
- KOST (2013), “PDF/A-2 und PDF/A-3: was ist neu?”, Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST), available at: https://kost-ceco.ch/cms/index.php?pdf-a-2_3_study_de (accessed May 17, 2018).

- KOST (2016), "PDF/A: produktreview batchtauglicher PDF/A-Konverter", Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST), in German or French, available at: http://kost-ceco.ch/cms/index.php?pdfa_konverter_de (accessed June 14, 2016).
- KOST (2017), "PDF/A: produktreview batchtauglicher PDF/A-Validatoren by kost-ceco", in German and French, available at: <https://kost-ceco.ch/cms/index.php?id=139,217,0,0,1,0> (accessed October 24, 2017).
- Müller, D. (2015), "Art und Verteilung von PDF-Formaten auf wissenschaftlichen deutschen Repositorien mit Hinblick auf Langzeitarchivierung", available at: https://ub-madoc.bib.uni-mannheim.de/38775/1/Mueller_Thesis_PDF_Formate.pdf (accessed October 24, 2017).
- Rimkus, K., Padilla, T., Popp, T. and Martin, G. (2014), "Digital preservation file format policies of ARL member libraries: an analysis", *D-Lib Magazine*, Vol. 20 Nos 3/4, available at: www.dlib.org/dlib/march14/rimkus/03rimkus.html
- Staatsarchiv (2013), "Anleitung zum Konvertieren von Office- und PDF-Formaten nach PDF/A", Staatsarchiv des Kantons Zürich, Version: 1.0, January 28.
- Sullivan, S.J. (2006), "An archival/records management perspective on PDF/A", *Records Management Journal*, Vol. 16 No. 1, pp. 51-56, available at: <https://doi.org/10.1108/09565690610654783>
- Suri, R.E. (2017), "Lost in migration: document quality for batch conversion to PDF/A", ETH Research Collection, data collection, ETH Zurich, available at: <http://doi.org/10.3929/ethz-b-000200243>

Corresponding author

Roland Erwin Suri can be contacted at: roland.suri@library.ethz.ch