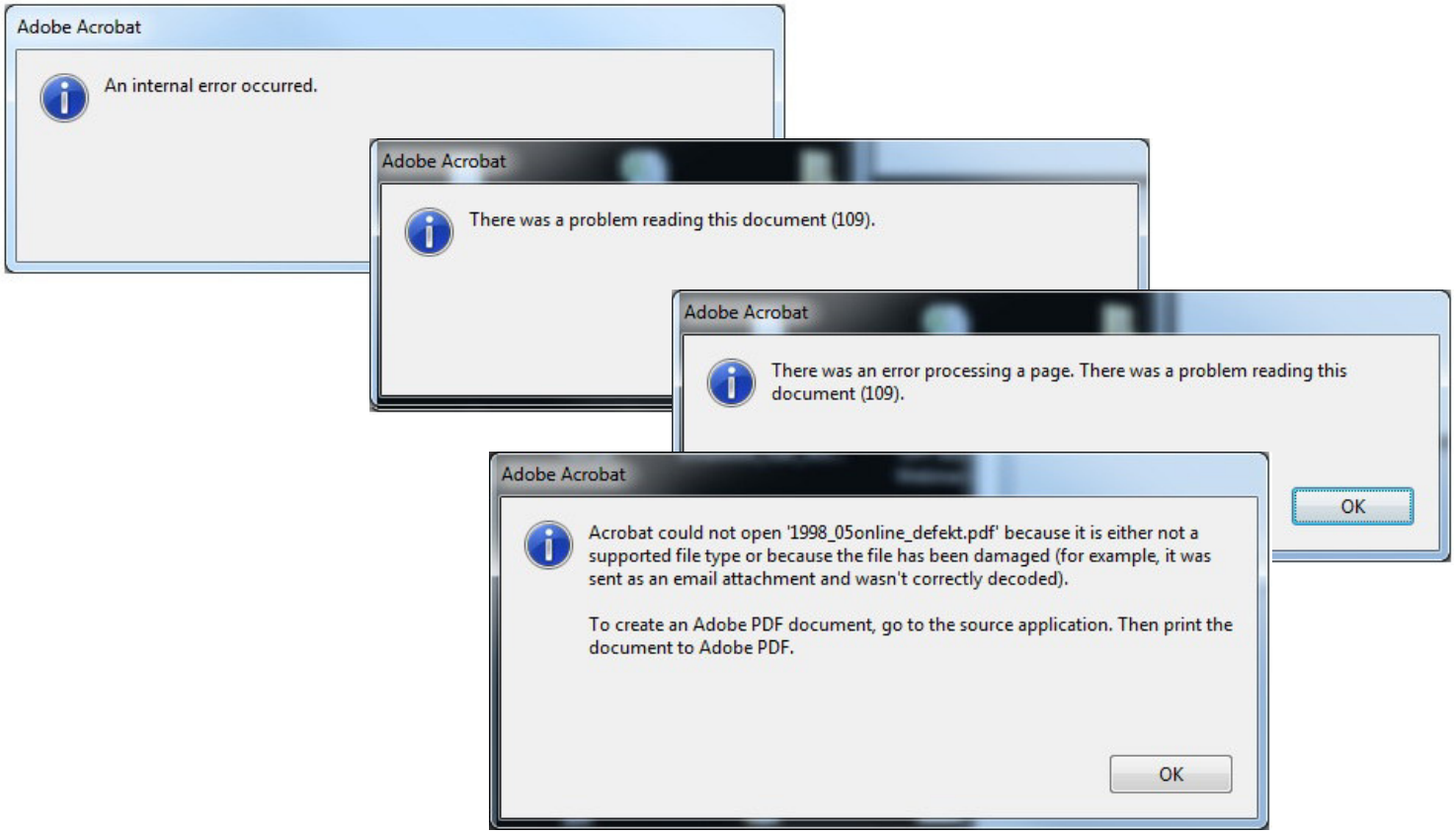
Too broken to be archived?
Practical experience with archiving PDF files

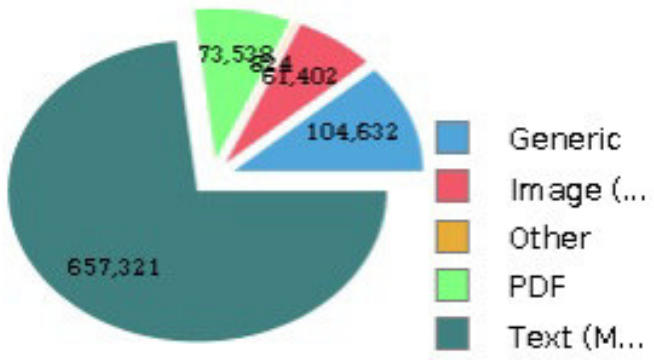
Yvonne Friese
Goportis/ZBW
20th july 2015



Introduce myself

NOT: a PDF expert

BUT: a practitioner having to archive A LOT of PDF files



Formats Breakdown Of Classification Group: PDF

Format ID	Version	Description	Number Of IEs
fmt/157		Portable Document Format - Exchange 1a:2001	1
fmt/95	1	Portable Document Format - Archival	8
fmt/158		Portable Document Format - Exchange 3:2002	156
fmt/354	1b	Acrobat PDF/A - Portable Document Format	309
fmt/276	1.7	Acrobat PDF 1.7 - Portable Document Format	1092
fmt/15	1.1	Portable Document Format	7355
fmt/19	1.5	Portable Document Format	8315
fmt/17	1.3	Portable Document Format	8869
fmt/18	1.4	Portable Document Format	14178
fmt/16	1.2	Portable Document Format	14833
fmt/20	1.6	Portable Document Format	18324

Topics

- Format identification
- Format validation
- Fixing PDFs
- Quality check
- PDF/A-2**a**
- Planned activities

Format Identification:
I'll give you three days to guess my name …

Format ID	Format Mime Type	Format Name
fmt/14	application/pdf	fmt/14
fmt/144	application/pdf	fmt/144
fmt/145	application/pdf	fmt/145
fmt/146	application/pdf	fmt/146
fmt/147	application/pdf	fmt/147
fmt/148	application/pdf	fmt/148
fmt/15	application/pdf	fmt/15
fmt/157	application/pdf	fmt/157
fmt/158	application/pdf	fmt/158
fmt/16	application/pdf	fmt/16
fmt/17	application/pdf	fmt/17
fmt/18	application/pdf	fmt/18
fmt/19	application/pdf	fmt/19
fmt/20	application/pdf	fmt/20
fmt/276	application/pdf	fmt/276
fmt/354	application/pdf	fmt/354

Format identification uncertainty: reasons

The tools come to different results, so the ingest stopps and somebody has to decide.

```
<fileinfo>
  <size toolname="Jhove" toolversion="1.5">174080</size>
  <creatingApplicationName toolname="Jhove" toolversion="1.5" status="CONFLICT">Acrobat PDFWriter 5.0 per Windows NT/WP-Lem Giarratana Et al.doc
  - Microsoft Word</creatingApplicationName>
  <creatingApplicationName toolname="Exiftool" toolversion="9.13" status="CONFLICT">Acrobat PDFWriter 5.0 per Windows NT/Labeipi8old
  </creatingApplicationName>
  <lastmodified toolname="Exiftool" toolversion="9.13" status="CONFLICT">2015:07:02 11:33:31+02:00</lastmodified>
  <lastmodified toolname="Tika" toolversion="1.3" status="CONFLICT">2003-10-27T13:56:50Z</lastmodified>
  <created toolname="Exiftool" toolversion="9.13" status="SINGLE_RESULT">2003:10:27 12:04:48Z</created>
  <filepath toolname="OIS File Information" toolversion="0.2" status="SINGLE_RESULT">
  C:\FileSample\text\PDFs\TechnicalMD_EconStor\fmt_multiple\391324195.pdf</filepath>
  <filename toolname="OIS File Information" toolversion="0.2" status="SINGLE_RESULT">391324195.pdf</filename>
  <md5checksum toolname="OIS File Information" toolversion="0.2" status="SINGLE_RESULT">a790167996166f58e6673fa2bce7a5bb</md5checksum>
  <fslastmodified toolname="OIS File Information" toolversion="0.2" status="SINGLE_RESULT">1435829611151</fslastmodified>
</fileinfo>
```

Inconsistencies within the PDF file: possible solution

Possible solution: e. g. PDF Tools search for inconsistencies and then decide how to convert via a table like this:

Info Dictionary	XMP	Conversion
Value present	Value not present	⇒ Add to XMP
Value not present	Value present	⇒ Add to doc info
Value 2	Value 1	⇒ Use XMP

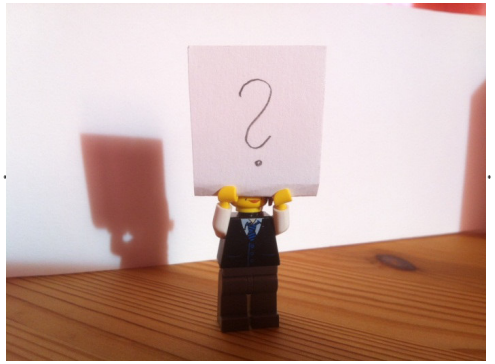
(Sebastian Ryffel „Challenges of Conversion from PDF to PDF/A, PDF Tools, during the PDF Days 2015 in Cologne)

Format Validation

So far, I have run into 60 JHOVE error messages.

Following, some of my favourites.

No PDF Header



Issue: saved as a PDF, but is none.

In this case, the reason: User got a 404 without realising and saved the 404-page as a PDF instead of the PDF he wanted to save.

Too broken to archive: Yes. Too abscent to archive.

Invalid PDF Trailer

Issue:The PDF ends before the EOF (end of File) – tag.

In this case, the reason: The upload has stopped before is was successfully uploaded.

Too broken to archive: Yes. The PDF is not complete.

URL: wiki.opf-labs.org/download/attachments/101613571/567147525.pdf?version=1&modificationDate=1436357226677

Those are easy to understand, but impossible to fix

But most JHOVE errors are easy to fix and impossible to understand

Jhove Validation Findings

FileName	Year of Creation	Creation Software	Jhove Module	Status	JhoveMessages	Message1	Message2	Message3
609865870	2009	MiKTeX pdfTeX-1.40.4	PDF-hul	Not well-formed	1	Improperly constructed page tree		
722289138	2011	Acrobat Distiller 9.4.5 (Windows)	PDF-hul	Not well-formed	1	Expected dictionary for font entry in page resource		
730732967	2012	Acrobat Distiller 9.5.2 (Windows)	PDF-hul	Not well-formed	1	Expected dictionary for font entry in page resource		
733248489	2012	Acrobat Distiller 9.5.2 (Windows)	PDF-hul	Not well-formed	1	Expected dictionary for font entry in page resource		
803538693	2014	Adobe Acrobat Pro 11.0.9	PDF-hul	Not well-formed	3	Expected dictionary for font entry in page resource	Invalid object number or object stream	Outlines contain recursive references.
819634859	2015	Acrobat Distiller 11.0 (Windows)	PDF-hul	Not well-formed	2	Improperly constructed page tree	Outlines contain recursive references.	
546281818	2007	Acrobat Distiller 7.0.5 (Windows)	PDF-hul	Not well-formed	1	Improperly nested array delimiters		
605029652	2009	Adobe PDF Library 9.0	PDF-hul	Not well-formed	1	Invalid object number in cross-reference stream		
605059357	2009	pdfTeX-1.40.9	PDF-hul	Not well-formed	1	Invalid object number in cross-reference stream		
571271863	2008	Acrobat Distiller 7.0.5 (Windows)	PDF-hul	Not well-formed	1	Lexical error		

Efforts to understand JHOVE error messages

http://wiki.opf-labs.org/display/Documents/JHOVE+issues+and+error+messages

Contains:

- Explanation: Well-Formedness & Validity
- Some basic PDF knowledge
- A list of JHOVE errors
    - Source code
    - Explanation
    - Impact
    - Cure
    - PDF example
    - (if available, yet to be continued)

Other validators: PDFBox for PDF/A

675 invalid PDF/A files contain 995 different error messages, all in all 219,265 Errors:

- Graphics: 117,457
- Fonts: 85,993
- Syntax: 11,196
- Annotations: 2,960
- MetaData: 1,520
- Transparency: 60
- Action: 79

In comparison, JHOVE is relatively silent.

Fixing JHOVE issues: the shallow way

Copy a PDF site-per-site into a new PDF (via iText) → Copy the XMP metadata into the new PDF(via iText) → Done

```
int n = reader.getNumberOfPages();
for (int i = 0; i < n;) {
copy.addPage(copy.getImportedPage(reader,
++i));
}
```

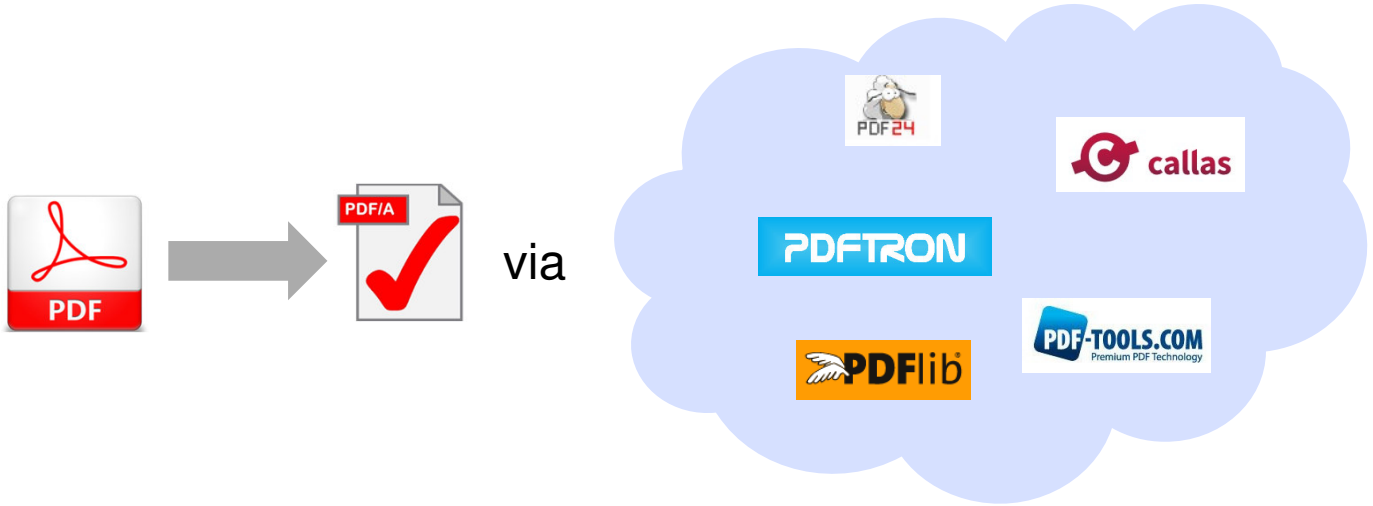
Conclusion: always works (as long as the original PDF is readable and not password-protected)

Findings: The output-PDF looks like the original, JHOVE considers it to be well-formed and valid

Suspicion: JHOVE just cannot find the issues any more..

https://github.com/friesey/preservation-tools/blob/workingfriese/pdf-tools/src/main/java/filetools/pdf/iTextRepairPdf.java

Fixing JHOVE issues: thoroughly



Issues:

- original PDF often too broken to be migrated
- migrated PDF does not look and feel as the original PDF

Quality Check: the twin tester tool

character-checking only

original PDF

This is Kitten

Saturday and Fun

Rosettacode and Programing

altered PDF

This is Sitting

Sunday and Fun

Raisethysword and Programing

Compared PDF Files

Original File	Number of Lines Original	Migrated File	Number of Lines Original Migration	PDF Twins	Different Lines	Levenshtein Distance Full PDF	Line Irregulaties (lines shifting etc.)
KittenSentences	3	KittenSentencesMig	3	false	3	15	not likely

Found differences in PDF Files

Difference in Line Number	Original Line	Migrated Line	Different Word (Original)	Different Word (Migration)	Levenshtein Distance
1	This is Kitten	This is Sitting			4
2	Saturday and Fun	Sunday and Fun			3
3	Rosettacode and Programing	Raisethysword and Programing			8

The twin tester tool: how it works

1. Extracts the text from the original PDF and the migrated PDF line-by-line
2. Compares each line
3. Puts out different lines, and – Levenshtein-distance is only 1– the different word in it
4. Calculates Levenshtein-distance to test if a line has shifted
5. Puts out all the Levenshtein-distances for each different line
6. Puts out the Levenshtein distance for all the lines in the PDF (very useful, if just a line or a few lines have shifted)

Source code: https://github.com/friesey/preservation-tools/blob/workingfriese/pdf-tools/src/main/java/filetools/pdf/PdfTwinTest.java

The twin tester in the real world: example 1

Compared PDF Files

Original File	Number of Lines Original	Migrated File	Number of Lines Original Migration	PDF Twins	Different Lines	Levenshtein Distance Full PDF	Line Irregulaties (lines shifting etc.)
Beispiel_5	1593	Beispiel_5_PDF24	1593	false	19	20	not likely

Found differences in PDF Files

Difference in Line Number	Original Line	Migrated Line	Different Word (Original)	Different Word (Migration)	Levenshtein Distance
143	"The rulers are not necessarily given any advance notice about (…)	" The rulers are not necessarily given any advance notice about (…)	"The	" The	1
199	satisfaction variables: 1) "satisfaction with the living conditions of one's family (henceforth	satisfaction variables: 1) "satisfaction with t he living conditions of one's family (henceforth	the	t he	1
200	"private satisfaction"); and 2) expectations concerning the living conditions of one's family in	"private satisfaction"); and 2) expectations conc erning the living conditions of one's family in	concerning	conc erning	1
713	Answers from 1 "very bad" to 5 "very good" (Country satisfaction); Do you think that in a year your life	Answers from 1 "very bad" to 5 "very good" (Country sa tisfaction); Do you think that in a year your life	satisfaction);	sa tisfaction);	1
714	and the life of your family will be: Answers from 1"much worse" to 5"much better" than now (Private	and the life of your family will be: Answers from 1" much worse" to 5"much better" than now (Private	1"much	1" much	1
739	question included in the CBOS survey: "Can you describe your political opinions? Please,	question included in the CBOS survey: " Can you describe your political opinions? Please,	"Can	" Can	1
813	"very bad" to 5 "very good" (Country satisfaction); Do you think that in a year your life and the life of your family	"very bad" to 5 "very good" (Country satisfaction); Do you th ink that in a year your life and the life of your family	think	th ink	1
815	family live? Answers from 1 "very bad" to 5 "very good" (Private satisfaction). Controls include gender, age, age-	family live? Answers from 1 "very bad" to 5 "very good" (P rivate satisfaction). Controls include gender, age, age-	(Private	(P rivate	1
870	The following questions were asked: How do you assess current economic situation in Poland? Answers from 1 "very bad" to 5 "very good" (Country satisfaction); Do you	The following questions were asked: How do you assess current economic situation in Poland? Answers from 1 "very bad" to 5 "ve ry good" (Country satisfaction); Do you	"very	"ve ry	1
871	think that in a year your life and the life of your family will be: Answers from 1"much worse" to 5"much better" than now (Private expectations); How do you and your family	think that in a year your life and the life of your family will be: Answers from 1"mu ch worse" to 5"much better" than now (Priv ate expectations); How do you and your family			2
872	live? Answers from 1 "very bad" to 5 "very good" (Private satisfaction). Year dummies included. *, ** and *** denote significance at the 10, 5 and 1% levels respectively.	live? Answers from 1 "very bad" to 5 "very good" (Private sati sfaction). Year dummies included. *, ** and *** denote significance at the 10, 5 and 1% levels respectively.	satisfaction).	sati sfaction).	1
926	Poland? Answers from 1 "very bad" to 5 "very good" (Country satisfaction); Do you think that in	Poland? Answers from 1 "very bad" to 5 "very good " (Country satisfaction); Do you think that in	good"	good "	1
979	"very bad" to 5 "very good" (Country satisfaction); Do you think that in a year your life and the life of your family	"very bad" to 5 "very good" (Country satisfaction); Do you th ink that in a year your life and the life of your family	think	th ink	1
981	family live? Answers from 1 "very bad" to 5 "very good" (Private satisfaction). Reference income is calculated by	family live? Answers from 1 "very bad" to 5 "very good" (Pri vate satisfaction). Reference income is calculated by	(Private	(Pri vate	1
1121	Milanovic, B. (1998). 'Income, inequality, and poverty during the transition from planned to	Milanovic, B. (1998). ' Income, inequality, and poverty during the transition from planned to	'Income,	' Income,	1
1314	"very bad" to 5 "very good" (Country satisfaction); Do you think that in a year your life and the life of your family	"very bad" to 5 "very good" (Country satisfaction); Do you th ink that in a year your life and the life of your family	think	th ink	1
1538	Answers from 1 "very bad" to 5 "very good" (Country satisfaction); Do you think that in a year your life	Answers from 1 "very bad" to 5 "very good" (Country sa tisfaction); Do you think that in a year your life	satisfaction);	sa tisfaction);	1
1539	and the life of your family will be: Answers from 1"much worse" to 5"much better" than now (Private	and the life of your family will be: Answers from 1" much worse" to 5"much better" than now (Private	1"much	1" much	1
1588	"very bad" to 5 "very good" (Country satisfaction) .Controls include gender, age, age-squared,	"very bad" to 5 "very good" (Country satisfacti on) .Controls include gender, age, age-squared,	satisfaction)	satisfacti on)	1

How it looks like if the lines were shifted

Compared PDF Files

Original File	Number of Lines Original	Migrated File	Number of Lines Original Migration	PDF Twins	Different Lines	Levenshtein Distance Full PDF	Line Irregulaties (lines shifting etc.)
original	5	migration	6	false	4	2	very likely

Found differences in PDF Files

Difference in Line Number	Original Line	Migrated Line	Different Word (Original)	Different Word (Migration)	Levenshtein Distance
2	The second line is not supposed to be empty, but is in the migrated PDF.				72
3	The program should detect this, as there are too many differences from that line on.	The second line is not supposed to be empty, but is in the migrated PDF.			62
4	Let's see if this works.	The program should detect this, as there are too many differences from that line on.			69
5		Let's see if this works.			24

Some other PDF conversion problems

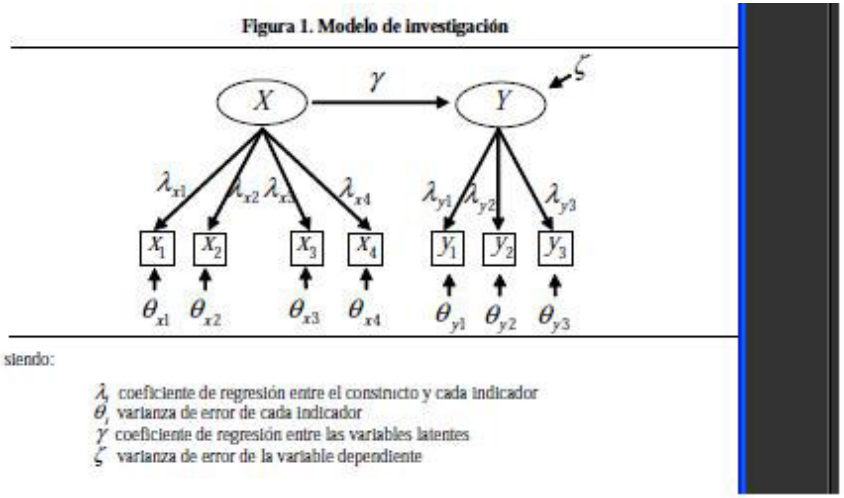
"PDF files may be broken in so many creative ways" (Duff Johnson, 2014)

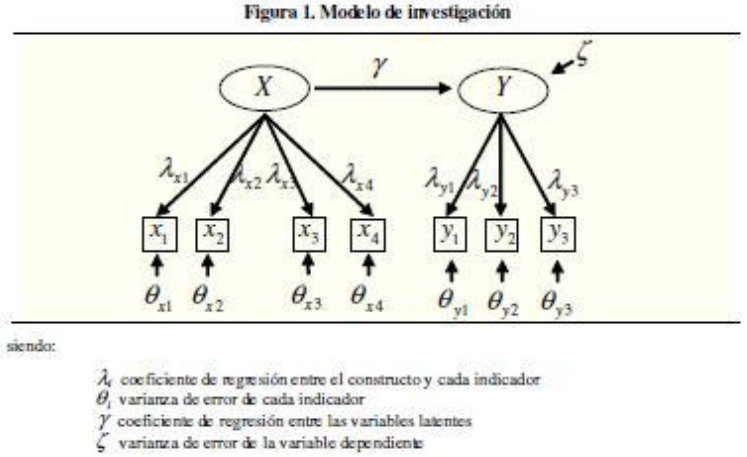
Font issues

Presentation during the PDF Technical Conference in Basel, 2012

PDF-PDF/A causes visual difference
why is this yellow now?

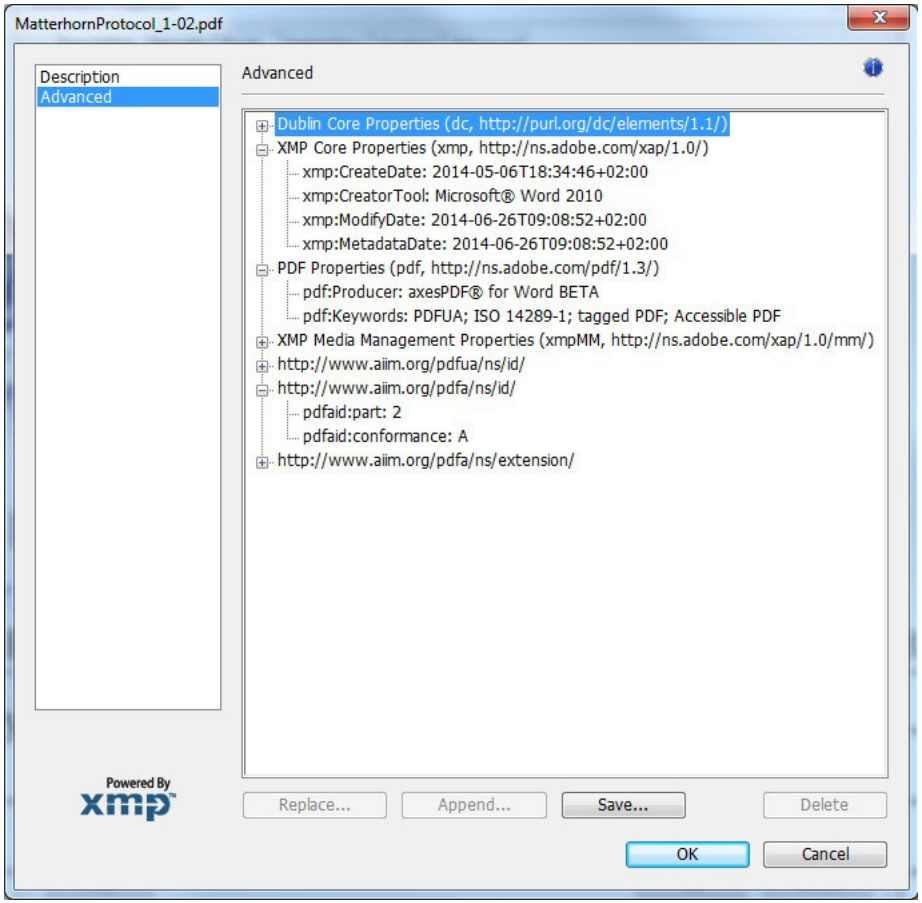
The white one is the (malformed) original PDF, the yellow one the PDF/A-migration





The first PDF/A-2a I‘ve ever run into

Source: http://www.pdfa.org/publication/the-matterhorn-protocol-1/



Planned activities (quality check)

- Some automated check on visual equality via matchbox or some other tool that can compare images.

- Enhance PDF preservation planning (lifelong task?)

Questions?

Yvonne Friese
ZBW, Kiel

Project management
Digital Curation
y.friese@zbw.eu