SCAPE

SCAlable Preservation Environments

Identification of preservation risks in PDF with Apache Preflight

a first impression

Authors

Person	Role	Partner	Contribution
Johan van der Knijff		KB	

Distribution

Person	Role	Partner

Revision History

Version	Status	Author	Date	Changes
0.1		Johan van der Knijff		

Table of Contents

Introduction

The *PDF* format contains a number of features that may make it difficult to access content that is stored in this format in the long term. Examples include (but are not limited to):

- Encryption features, which may either restrict some functionality (copying, printing) or make files inaccessible altogether.
- Multimedia features (embedded multimedia objects may be subject to format obsolescence)
- Reliance on external features (e.g. non-embedded fonts, or references to external documents)

A more exhaustive overview is given here:

http://www.openplanetsfoundation.org/system/files/PDFInventoryPreservationRisks_0_2_0.pdf

and also here:

http://libraries.stackexchange.com/questions/964/what-preservation-risks-are-associated-with-the-pdf-file-format

When *creating* a *PDF*, it is possible to minimise these risks by using one of the *PDF/A* standards, which delineate a number of *PDF* feature profiles that are unlikely to result in any long-term accessibility problems. However, the simple fact is that most *PDF*s that are out there are not *PDF/A*. Nevertheless, for assessing risks in existing collections it would be very helpful to be able to screen *PDF*s for specific 'risky' features, such as encryption or font embedding. Since *PDF/A* was specifically designed to eliminate these 'risky' features, one would expect that *PDF/A* validators (i.e. software tools that check the conformance of a *PDF* file against the *PDF/A* specification) would be able to provide some useful information on this.

Scope and objectives of this work

The aim of this report is to assess the feasibility of using an open-source *PDF/A-1* validator tool for detecting 'risky' features in *PDF* files. More specifically, its objectives are:

- To get a first impression of the *Apache Preflight* (part of *PDFBox*) *PDF/A-1b* validator.
- To investigate if *Apache Preflight* is able to detect unwanted (from a preservation point of view) features in *PDF* files (i.e. *PDF*s that are not necessarily of the *PDF/A* sub-type) such as password protection, encryption and non-embedded fonts.
- To provide a comparison with the *Preflight* module of *Adobe Acrobat* 9.5.

- To decide if doing more work on *Apache Preflight* (more elaborate testing, possible involvement in its development) are worthwhile.

To avoid any confusion, the scope of this work does *not* include testing whether *Apache Preflight* is capable of a full *PDF/A* validation.

Apache Preflight

Apache Preflight is a *PDF/A-1b* validator that is part of *Apache PDFBox*, which is an open-source *Java* library for creating, processing and analysing *PDF* documents. The current stable version of *PDFBox* is 1.7.1. However, the documentation suggests that in this version the *Preflight* module is only accessible by calling it from a *Java* application (it cannot be invoked through a command line interface, and no stand-alone *JAR*s appear to exist). The (officially unreleased) version 1.8.0 *does* include a stand-alone *JAR* that can be called from the command-line, and this is the version that was used for these tests. All tests were done with build #562 of this version, using with the "jar with dependencies".

Preflight wrapper (*pftree*)

Apache Preflight produces output that is fairly unstructured when used from the command line. In order to streamline the testing process, I created a simple custom-built Python wrapper that traverses a user-specified directory tree, launches *Apache Preflight* for each file that has a *.pdf* extension, and then reports the results in XML format. It also reports whether *Apache Preflight* exited with any exceptions. The wrapper code can be found here:

https://github.com/bitsgalore/pftree

All tests were done with *pftree* 0.3.

Acrobat Preflight

The professonal version of *Adobe Acrobat* includes a *Preflight* module that is capable of testing files against of number of pre-defined profiles. This also includes a number of *PDF/A-1* profiles, for which *Acrobat Preflight* uses a third-party plugin that is based on a commercially available *PDF/A* validator by *Callas* Software. *Acrobat Preflight* is used here as a reference for assessing the *Apache Preflight* results. The following software versions were used:

- Adobe Acrobat Professional 9.5.2
- Preflight 9.2.0 (065)

Environment

All tests were done under Windows XP Professional, 5.1.2.2600 SP 3 Build 2600.

Test data

The Archivist's PDF Cabinet of Horrors dataset

This is a suite of small, simple test files that were created especialy for this work. Each file contains one 'risky' feature, with focus on the following feature classes:

- Encryption
- Multimedia
- Scripts
- Fonts
- File attachments
- External references
- Byte corruption

The odd one out in this dataset is *test_fontArialNotEmbedded.pdf*, which is a more realistic (larger and more complex) document.

Link:

http://www.opf-labs.org/format-corpus/pdfCabinetOfHorrors/

A detailed description of this dataset can be found here:

http://www.opf-labs.org/format-corpus/pdfCabinetOfHorrors/readme.md

OOJackson dataset

These are test files that were created with *OpenOffice* 3.2 by Andy Jackson. Its focus is mainly on encryption. Link:

http://www.opf-labs.org/format-corpus/office-examples/OpenOffice.org%203.2.0%20OSX/

A more detailed description of the dataset is here:

http://www.opf-labs.org/format-corpus/office-examples/OpenOffice.org%203.2.0%20OSX/README.md

Test procedure

Apache Preflight

I ran the *pftree* wrapper on each data set. This resulted in two XML files containing the *Apache Preflight* output for the *PDF Cabinet of Horrors* and *OOJackson* datasets, respectively. (I then processed these files further with a script that extracts all error codes and messages.)

Acrobat Preflight

For the *Acrobat Preflight* analysis I set up a simple batch process in *Acrobat* that can be re-created in the following way:

1. Go to *Advanced/Document Processing/Batch Processing ...*
2. Create *Preflight* sequence for *Verify compliance with PDF/A-1b*
3. Select *PDF Report* in output settings
4. Run on all *PDF*s in dataset

This results, for each analysed file, in a *PDF*-formatted report with detailed results of the validation outcome. For this analysis we're only really interested in the contents of each report's *Summary* section, which are reproduced in Annex 1.

Which *PDF/A* violations are important?

For a meaningful interpretation of these results, it is important to note that *PDF/A* validation can -in theory- produce 3 types of violations with respect to the standard [PDFLib, 2009]:

1. ISO 19005-1 violations, i.e. one or more requirements of the PDF/A standard itself are not met. This category includes problems with XMP extension schemas since these are defined in the PDF/A standard.
2. XMP 2004 violations, i.e. predefined XMP properties are not used according to the XMP 2004 specification which is referenced by PDF/A. XMP syntax problems also fall in this category.
3. PDF 1.4 violations, i.e. some requirements of the PDF reference are not met. This includes various implementation limits which are documented in the PDF 1.4 reference and must adhered to by conforming PDF/A documents. While blatant violations could result in unusable files (e.g. a page cannot be displayed), there are many subtle requirements in PDF 1.4 which must be met for PDF/A conformance and should therefore be checked by a validator.

Here, we are only interested in a subset of 1st category. Also note that the last category (PDF 1.4 violations) is (as far as I'm aware) not covered by *Apache Preflight* at all.

Both *Apache Preflight* and *Acrobat Preflight* validate a file against the *PDF/A-1* specification and, as such, will report on any features that are not in accordance with *PDF/A-1*. As a result, for most *PDF*s that were not created using the *PDF/A* profile, these validators will report error messages on the absence of *PDF/A*-specific metadata, compressed object streams and device colors. These are not particularly important within the scope of this work, and for clarity these error messages are not reproduced in the sections below. For the sake of brevity the *Acrobat Preflight* output is not shown in the main text either. The full results of both tools are included in Annex 1 of this report.

Results *PDF Cabinet of Horrors* dataset

Encryption

Apache Preflight was able to identify all documents with encrypted content. Documents that require a password to open the file gave the following error:

- 1.0: Syntax error, Error (CryptographyException) while creating security handler for decryption: Error: The supplied password does not match either the owner or user password in the document.

All other access restrictions resulted in:

- 1.4.2: Trailer Syntax error, The trailer dictionary contains Encrypt

However, *Preflight* doesn't give any specific information on which specific access restrictions apply (e.g. printing, copying, text access). In this regard it behaves similarly to *Acrobat Preflight* (although *Acrobat* simply produces *no output at all* in case of a 'file open' password!).

Multimedia

Both validators successfully identified all documents with embedded multimedia content, which resulted in the following errors for *Apache Preflight*:

- 1.2.9: Body Syntax error, EmbeddedFile entry is present in a FileSpecification dictionary

- 5.2.1: Forbidden field in an annotation definition, The subtype isn't authorized : Screen

Fonts

The detection of non-embedded fonts turned out to be problematic for *Apache Preflight*. A simple test file with 1 single font that is not embedded resulted in the following errors:

- 3.1.3: Invalid Font definition, They are more than one FontFile

- 3.3.1: Glyph error, The character "84" in the font program "TimesNewRomanPSMT"is missing from the Charater Encoding.

Although *Apache Preflight* did pick up a font-related issue here, the reported error messages are confusing and do not reflect the actual problem. A more serious problem is that for a more complex document, *Preflight* didn't pick up a non-embedded font at all. I reported this using the project's issue tracker:

https://issues.apache.org/jira/browse/PDFBOX-1449

One of the software's main authors (Eric Leleu) explained that the missing font is not reported because of another *PDF/A* violation (related to the used color space) in the document. The effect is that any further processing of the page on which the error occurs is stopped, with the result that the non-embedded font is not detected. He also indicated that he will search for ways to increase the number of errors reported by *Apache Preflight*.

File attachments

File attachments were detected by both validators, and produced the following error messages with *Apache Preflight*:

- 1.2.9: Body Syntax error, EmbeddedFile entry is present in a FileSpecification dictionary

- 1.4.7: Trailer Syntax error, EmbeddedFile entry is present in the Names dictionary

Acrobat Preflight also reported a *Document contains JavaScripts* message for this file. Since the *Apache Preflight* output didn't mention *JavaScript* at all, I opened the file in a hex editor, which revealed that it indeed contained *JavaScript*.

Scripts

Apache Preflight reported the following error for a *PDF* with JavaScript:

- 6.2.5: Action is forbidden, The action JavaScript is forbidden

However, for another file that contained both a file attachment and some accompanying *JavaScript*, *Apache Preflight* didn't detect the *JavaScript* (see above). Although I'm not sure about the underlying reason, this could be just another example of *Apache Preflight* giving up on processing a page after encountering another error (in this case the embedded file attachment).

External references

A reference to another document was detected by both validators, and resulted in the following error with *Apache Preflight*:

- 6.2.5: Action is forbidden, The action Launch is forbidden

Use of the *Web Capture* feature produced this error (followed by a whole range of other errors that are related to the imported content):

- 6.2.4: Action is forbidden, "A" must not be used in a Field dictionary

(*Web Capture* content was detected by *Acrobat Preflight* as well.)

Byte corruption

A file with one byte missing from the comment line following the file header resulted in the following error messages with *Apache Preflight*:

- 1.1: Body Syntax error, Second line must contains at least 4 bytes greater than 127

- 1.0: Syntax error, Error: Expected an integer type, actual='ref'

The file was reported as 'damaged' by *Acrobat Preflight*.

Results *OOJackson* dataset

The results of the *OOJackson* dataset were similar to those of the *PDF Cabinet of Horrors* one, so they are not reproduced here. For completeness they are also included in Annex 1. One noteworthy issue (though unrelated to the scope of this work) is that *Apache Preflight* reported the following error for all files that were created in *OpenOffice*:

- 1.1: Body Syntax error, Second line must contains at least 4 bytes greater than 127

I initially thought that this was somehow related to incorrect handling of line terminators by *Apache Preflight* (`0x 0A` - `0x 0D` - `0x 0A 0D`), however the experiments I did on this were ultimately inconclusive. (Changing the terminator of the second line from `0x 0A` to `0x 0A 0D` made the error go away, but so did changing the byte just *before* the terminator from `0x 9F` to `0x C3`.)

Also, further tests resulted in a crash of Apache Preflight for one file. More details and a link to the file can be found here:

https://issues.apache.org/jira/browse/PDFBOX-1465

Discussion

Detection of unwanted features

The tests with the two datasets demonstrate that *Apache Preflight* is able to identify simple *PDF*s with problematic features in most cases. However, it appears to be less successful at this for more complex documents that violate the *PDF/A-1b* specification in multiple ways. For such documents, encountered errors may stop *Preflight* from processing pages any further, and as a result features such as non-embedded fonts or *JavaScript* are not reported. This is not a problem if *Apache Preflight* is used to do a strict *PDF/A* compliance check (which is what the software was designed for in the first place). On the other hand, it makes the software less suitable for profiling use cases such as the one described here (especially with large, complex documents).

The principal author of *Apache Preflight* has indicated that he will search for ways to increase the number of errors reported by the software, so this may well improve in the future.

Also, the error messages reported by *Apache Preflight* in case of non-embedded fonts are slightly confusing, as they don't really reflect the actual problem.

In the (limited) tests here, the *Preflight* module of *Adobe Acrobat* was more successful at detecting non-embedded fonts and *JavaScript*, especially for more complex (i.e. real-world) documents. Although *Acrobat* is not well suited to large-scale processing, its *Preflight* module is based on the *pdfaPilot* software by *Callas*, and for the purposes described in this report this software might be an interesting (though non-free and non-open-source) alternative. Some other similar products are mentioned in [PDFLib, 2009]. Needless to say, further testing would be needed to determine the utility of any of these products.

Scalability: stability and performance

Since the primary focus of this study is on *Apache Preflight*'s overall ability to detect unwanted features in *PDF*, I didn't do any of the larger-scale tests that would be needed to properly assess its scalability. Nevertheless, it is possible to make some observations from these results.

Incidentally the program did crash on one of the first *PDF*s that I used for trying out the tool; see the link below for more details:

https://issues.apache.org/jira/browse/PDFBOX-1465

In addition to this I also ran *Apache Preflight* on the 'Bavaria test suite', which is a corpus of 85 *PDF* files that was used in the *Bavaria Report* on *PDF/A* validation [PDFLib, 2009]. A link is provided here:

http://www.pdflib.com/knowledge-base/pdfa/validation-report/

Preflight raised an exception for 5 out of 85 files in the dataset (6%), which indicates that at this stage it may simply not be sufficiently stable or mature for operational use.

Conclusions and recommendations

Although *Apache Preflight* shows some great promise for detecting features in *PDF* that are unwanted in archival settings, use of this software for these purposes is not feasible at this stage yet. The main reasons for this are:

1. The fact that once a violation of the *PDF/A-1b* is detected, this may stop *Preflight* from any further processing of that page.

2. Lack of sufficient stability.

The first issue isn't really a shortcoming of *Preflight*. The software was really designed to validate files against *all* aspects of *PDF/A*, whereas the objective of this study is to check for individual aspects of the standard. So to a large extent we're trying to use *Preflight* for something it was wasn't designed for in the first place.

In contrast to this, the stability issue will also limit *Preflight*'s usability for conventional *PDF/A* validation. However, we should keep in mind here that the development of *Apache Preflight* is still in its early stages, and that all tests were based on a version that has not even been released officially.

The software hasn't attracted a lot of attention from the archival community so far. Since *PDF* is one of the predominant formats that many archives and libraries are dealing with, an active involvement of our community in *Apache Preflight* (development, testing) could be a wortwhile long-term investment.

In the short to medium run, a number of commercial software products exist for *PDF/A* validation, and some of these are probably able to do detection of unwanted or risky *PDF* features already. As an example, *Adobe Acrobat*'s *Preflight* module was able to detect all problematic features in the test data used here. *Acrobat* is not very scalable, but its *PDF/A* validation is based on a commercially sold *PDF/A* validator tool (and some other, similar tools exist as well). Further tests using a number of such tools could be the subject of a follow-up to this work.

Funding

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

References

PDFLib: Bavaria Report on PDF/A Validation Accuracy. Link: http://www.pdflib.com/fileadmin/pdflib/pdf/pdfa/2009-05-04-Bavaria-report-on-PDFA-validation-accuracy.pdf

Test data

PDF Cabinet of Horrors Corpus:

http://www.opf-labs.org/format-corpus/pdfCabinetOfHorrors/

OOJackson Corpus:

http://www.opf-labs.org/format-corpus/office-examples/OpenOffice.org%203.2.0%20OSX/

Bavaria test suite:

http://www.pdflib.com/fileadmin/pdflib/Bavaria/2009-04-03-Bavaria-pdfa.zip

Annex 1: unedited output of Apache Preflight and Acrobat Preflight

Results *PDF Cabinet of Horrors* dataset

Encryption

encryption_openpassword.pdf

Requires password to open the file.

Apache Preflight

- 1.0: Syntax error, Error (CryptographyException) while creating security handler for decryption: Error: The supplied password does not match either the owner or user password in the document.

Acrobat Preflight

Acrobat cannot open the file (no output is produced, although Acrobat does report a warning message).

encryption_nocopy.pdf

Requires password to copy document contents. #### Apache Preflight + 1.4.2: Trailer Syntax error, The trailer dictionary contains Encrypt + 7.1.1: Error on MetaData, No type defined for {http://ns.adobe.com/xap/1.0/mm/}subject

Acrobat Preflight

- Compressed object streams used
- Device process color used but no PDF/A OutputIntent (4 matches on 1 page)
- Document is encrypted
- Encrypt key present in file trailer
- PDF/A entry missing
- XMP property not predefined and no extension schema present

encryption_noprinting.pdf

Requires password for printing.

Apache Preflight

- 1.4.2: Trailer Syntax error, The trailer dictionary contains Encrypt
- 7.1.1: Error on MetaData, No type defined for {http://ns.adobe.com/xap/1.0/mm/}subject

Acrobat Preflight

- Compressed object streams used
- Device process color used but no PDF/A OutputIntent (4 matches on 1 page)
- Document is encrypted
- Encrypt key present in file trailer
- PDF/A entry missing
- XMP property not predefined and no extension schema present

encryption_notextaccess.pdf

Requires password to enable text access for screen reader devices for the visually impaired.

Apache Preflight

- 1.4.2: Trailer Syntax error, The trailer dictionary contains Encrypt
- 7.1.1: Error on MetaData, No type defined for {http://ns.adobe.com/xap/1.0/mm/}subject

Acrobat Preflight

- Compressed object streams used
- Device process color used but no PDF/A OutputIntent (4 matches on 1 page)
- Document is encrypted
- Encrypt key present in file trailer
- PDF/A entry missing
- XMP property not predefined and no extension schema present

Multimedia

embedded_video_avi.pdf

Contains embedded *AVI* movie.

Apache Preflight

- 1.2.9: Body Syntax error, EmbeddedFile entry is present in a FileSpecification dictionary
- 5.2.1: Forbidden field in an annotation definition, The subtype isn't authorized : Screen
- 1.4.6: Trailer Syntax error, ID is different in the first and the last trailer
- 7.1.1: Error on MetaData, No type defined for {http://ns.adobe.com/xap/1.0/mm/}subject

Acrobat Preflight

- Compressed object streams used
- Device process color used but no PDF/A OutputIntent (6 matches on 1 page)
- Incorrect annotation type used (not allowed in PDF/A) (1 match on 1 page)
- PDF contains EF (embedded file) entry
- PDF/A entry missing
- XMP property not predefined and no extension schema present

embedded_video_quicktime.pdf

Contains embedded *Quicktime* movie.

Apache Preflight

- 1.2.9: Body Syntax error, EmbeddedFile entry is present in a FileSpecification dictionary
- 5.2.1: Forbidden field in an annotation definition, The subtype isn't authorized : Screen
- 1.4.6: Trailer Syntax error, ID is different in the first and the last trailer
- 7.1.1: Error on MetaData, No type defined for {http://ns.adobe.com/xap/1.0/mm/}subject

Acrobat Preflight

- Compressed object streams used
- Device process color used but no PDF/A OutputIntent (7 matches on 1 page)
- Incorrect annotation type used (not allowed in PDF/A) (1 match on 1 page)
- PDF contains EF (embedded file) entry
- PDF/A entry missing
- XMP property not predefined and no extension schema present

Scripts

javascript.pdf

Contains embedded *Javascript*.

Apache Preflight

- 1.1: Body Syntax error, Second line must contains at least 4 bytes greater than 127
- 6.2.5: Action is forbidden, The action JavaScript is forbidden
- 3.1.1: Invalid Font definition, Some required fields are missing from the Font dictionary.

- 3.1.2: Invalid Font definition, FontDescriptor is null or is a AFM Descriptor
- 3.3.1: Glyph error, The character "74" in the font program "Helvetica"is missing from the Charater Encoding.
- 1.4.1: Trailer Syntax error, The trailer dictionary doesn't contain ID
- 7.1: Error on MetaData, Missing Metadata Key in catalog

Acrobat Preflight

- Contains action of type JavaScript
- Device process color used but no PDF/A OutputIntent (1 match on 1 page)
- Document contains JavaScripts
- File header not compliant with PDF/A
- Font not embedded (and text rendering mode not 3) (1 match on 1 page)
- ID in file trailer missing or incomplete
- Metadata missing (XMP)
- PDF/A entry missing

Fonts

text_only_fontsNotEmbedded.pdf

Used fonts are not embedded.

Apache Preflight

- 3.1.3: Invalid Font definition, They are more than one FontFile
- 3.3.1: Glyph error, The character "84" in the font program "TimesNewRomanPSMT"is missing from the Charater Encoding.
- 7.1.1: Error on MetaData, No type defined for {http://ns.adobe.com/xap/1.0/mm/}subject

Acrobat Preflight

- Compressed object streams used
- Device process color used but no PDF/A OutputIntent (4 matches on 1 page)
- Font not embedded (and text rendering mode not 3) (4 matches on 1 page)
- PDF/A entry missing
- XMP property not predefined and no extension schema present

text_only_fontsEmbeddedAll.pdf

Used fonts are embedded.

Apache Preflight

- 7.1.1: Error on MetaData, No type defined for {http://ns.adobe.com/xap/1.0/mm/}subject

Acrobat Preflight

- Compressed object streams used
- Device process color used but no PDF/A OutputIntent (4 matches on 1 page)
- PDF/A entry missing
- XMP property not predefined and no extension schema present

text_only_fontsEmbeddedSubset.pdf

Used fonts are embedded as subset.

Apache Preflight

- 7.1.1: Error on MetaData, No type defined for {http://ns.adobe.com/xap/1.0/mm/}subject

Acrobat Preflight

- Compressed object streams used
- Device process color used but no PDF/A OutputIntent (4 matches on 1 page)
- PDF/A entry missing
- XMP property not predefined and no extension schema present

text_only_pdfa1b.pdf

PDF/A-1b (with embedded fonts).

Apache Preflight

None (valid *PDFA-1b*)

Acrobat Preflight

None (*No problems found*)

test_fontArialNotEmbedded.pdf

Font *Arial* is not embedded (other fonts are).

Apache Preflight

- 2.4.3: Invalid Color space, The operator "g" can't be used without Color Profile
- 7.11: Error on MetaData

Acrobat Preflight

- Compressed object streams used
- Device process color used but no PDF/A OutputIntent (2083 matches on 55 pages)
- Font not embedded (and text rendering mode not 3) (1475 matches on 55 pages)

- PDF/A entry missing

File attachments

fileAttachment.pdf

Contains file attachment.

Apache Preflight

- 1.2.9: Body Syntax error, EmbeddedFile entry is present in a FileSpecification dictionary
- 1.4.7: Trailer Syntax error, EmbeddedFile entry is present in the Names dictionary
- 7.1.1: Error on MetaData, No type defined for {http://ns.adobe.com/xap/1.0/mm/}subject

Acrobat Preflight

- Compressed object streams used
- Device process color used but no PDF/A OutputIntent (4 matches on 1 page)
- Document contains JavaScripts
- EmbeddedFiles entry in Names dictionary
- PDF contains EF (embedded file) entry
- PDF/A entry missing
- XMP property not predefined and no extension schema present

Note: the *Document contains JavaScripts* message is unexpected here and needs further analysis!

External references

externalLink.pdf

Contains link to another document.

Apache Preflight

- 6.2.5: Action is forbidden, The action Launch is forbidden
- 7.1.1: Error on MetaData, No type defined for {http://ns.adobe.com/xap/1.0/mm/}subject

Acrobat Preflight

- Compressed object streams used
- Contains action of type Launch
- Device process color used but no PDF/A OutputIntent (5 matches on 1 page)
- PDF/A entry missing

- XMP property not predefined and no extension schema present

webCapture.pdf

Uses Web Capture feature for importing text from a website.

Apache Preflight

- 2.4.3: Invalid Color space, The operator "g" can't be used without Color Profile
- 6.2.4: Action is forbidden, "A" must not be used in a Field dictionary
- 5.1: Missing field in an annotation definition
- 2.4.3: Invalid Color space, The operator "g" can't be used without Color Profile
- 2.2.2: Invalid Graphis transparency, Soft Mask must be null or None
- 2.2.1: Invalid Graphis transparency, Group has a transparency S entry or the S entry is null.
- 3.1.3: Invalid Font definition, They are more than one FontFile
- 3.3.1: Glyph error, The character "79" in the font program "Tahoma"is missing from the Charater Encoding.
- 5.1: Missing field in an annotation definition
- 2.2.1: Invalid Graphis transparency, Group has a transparency S entry or the S entry is null.
- 3.3.1: Glyph error, The character "79" in the font program "Tahoma"is missing from the Charater Encoding.
- 1.4.6: Trailer Syntax error, ID is different in the first and the last trailer
- 7.1.1: Error on MetaData, No type defined for {http://ns.adobe.com/xap/1.0/mm/}subject

Acrobat Preflight

- Annotation has no Flags entry (109 matches on 2 pages)
- Annotation not set to print (109 matches on 2 pages)
- Compressed object streams used
- Device process color used but no PDF/A OutputIntent (13 matches on 3 pages)
- Font not embedded (and text rendering mode not 3) (124 matches on 2 pages)
- Form field does not have appearance dict (1 match on 1 page)
- Has actions (1 match on 1 page)
- PDF/A entry missing Transparency used (soft mask in image) (46 matches on 1 page)

- XMP property not predefined and no extension schema present

Byte corruption

corruptionOneByteMissing.pdf

One byte missing from comment line following file header.

Apache Preflight

- 1.1: Body Syntax error, Second line must contains at least 4 bytes greater than 127
- 1.0: Syntax error, Error: Expected an integer type, actual='ref'

Acrobat Preflight

- Document is damaged and needs repair
- File header not compliant with PDF/A
- Syntax problem: PDF contains data after end of file marker

Results *OOJackson* dataset

Encryption

simple-open-password.pdf

Requires a password to open.

Apache Preflight

- 1.1: Body Syntax error, Second line must contains at least 4 bytes greater than 127
- 1.0: Syntax error, Error (CryptographyException) while creating security handler for decryption: Error: The supplied password does not match either the owner or user password in the document.

Acrobat Preflight

Acrobat cannot open the file (no output is produced, although Acrobat does report a warning message).

simple-open-nocopy-password.pdf

Requires a password to open, and the 'copy' right is restricted.

Apache Preflight

- 1.1: Body Syntax error, Second line must contains at least 4 bytes greater than 127
- 1.0: Syntax error, Error (CryptographyException) while creating security handler for decryption: Error: The supplied password does not match either the owner or user password in the document.

Acrobat Preflight

Acrobat cannot open the file (no output is produced, although Acrobat does report a warning message).

simple-password-copy.pdf

Encrypted (with default password of ""), no rights restrictions.

Apache Preflight

- 1.1: Body Syntax error, Second line must contains at least 4 bytes greater than 127
- 2.2.1: Invalid Graphis transparency, Group has a transparency S entry or the S entry is null.
- 2.4.1: Invalid Color space, The operator "rg" can't be used with CMYK Profile
- 1.4.2: Trailer Syntax error, The trailer dictionary contains Encrypt
- 7.1: Error on MetaData, Missing Metadata Key in catalog

Acrobat Preflight

- Author mismatch between Document Info and XMP Metadata
- Creation date mismatch between Document Info and XMP Metadata
- Creator mismatch between Document Info and XMP Metadata
- Device process color used but no PDF/A OutputIntent (1 match on 1 page)
- Document is encrypted
- Encrypt key present in file trailer
- Metadata missing (XMP)
- PDF/A entry missing
- Producer mismatch between Document Info and XMP Metadata

simple-password-nocopy.pdf

Encrypted (with default password of ""), and the 'copy' right is restricted.

Apache Preflight

- 1.1: Body Syntax error, Second line must contains at least 4 bytes greater than 127
- 2.2.1: Invalid Graphis transparency, Group has a transparency S entry or the S entry is null.
- 2.4.1: Invalid Color space, The operator "rg" can't be used with CMYK Profile
- 1.4.2: Trailer Syntax error, The trailer dictionary contains Encrypt
- 7.1: Error on MetaData, Missing Metadata Key in catalog

Acrobat Preflight

- Author mismatch between Document Info and XMP Metadata
- Creation date mismatch between Document Info and XMP Metadata

- Creator mismatch between Document Info and XMP Metadata
- Device process color used but no PDF/A OutputIntent (1 match on 1 page)
- Document is encrypted
- Encrypt key present in file trailer
- Metadata missing (XMP)
- PDF/A entry missing
- Producer mismatch between Document Info and XMP Metadata

simple-annotated-in-adobe-x.pdf

PDF generated by OOO and subsequently annotated it in Adobe X.

Apache Preflight

- 5.2.3: Forbidden field in an annotation definition, Annotation uses a Color profile which isn't the same than the profile contained by the OutputIntent
- 5.1: Missing field in an annotation definition
- 2.2.1: Invalid Graphis transparency, Group has a transparency S entry or the S entry is null.
- 2.4.1: Invalid Color space, The operator "rg" can't be used with CMYK Profile
- 7.11: Error on MetaData

Acrobat Preflight

- Annotation has C entry but no PDF/A OutputIntent present (2 matches on 1 page)
- Compressed object streams used
- Device process color used but no PDF/A OutputIntent (5 matches on 1 page)
- PDF/A entry missing
- Transparency used (blend mode not “Normal” nor “Compatible”) (1 match on 1 page)
- Transparency used (filled object with ca value smaller than 1.0) (1 match on 1 page)

Miscellaneous

simple.pdf

Simple basic document on which all other *PDF*s in this datset were based.

Apache Preflight

- 1.1: Body Syntax error, Second line must contains at least 4 bytes greater than 127

- 2.2.1: Invalid Graphis transparency, Group has a transparency S entry or the S entry is null.
- 2.4.1: Invalid Color space, The operator "rg" can't be used with CMYK Profile
- 7.1: Error on MetaData, Missing Metadata Key in catalog

Acrobat Preflight

- Author mismatch between Document Info and XMP Metadata
- Creation date mismatch between Document Info and XMP Metadata
- Creator mismatch between Document Info and XMP Metadata
- Device process color used but no PDF/A OutputIntent (1 match on 1 page)
- Metadata missing (XMP)
- PDF/A entry missing
- Producer mismatch between Document Info and XMP Metadata

simple-PDFA-1a.pdf

PDF/A-1b.

Apache Preflight

- 1.1: Body Syntax error, Second line must contains at least 4 bytes greater than 127

Acrobat Preflight

None (*No problems found*)