Why PDF/A validation matters
(even if you don’t have PDF/A)

Johan van der Knijff
'encryption_openpassword' is protected. Please enter a Document Open Password.

Enter Password: [ ]

OK  |  Cancel
This is a test document. It only contains some text.

Creative Commons CC0: Public Domain Dedication. To the extent possible under law, Johan van der Knijff has waived all copyright and related or neighboring rights to this work.
der Knijff has waived all copyright and related or neighboring rights to this work.

Embedded Quicktime video by Misty De Meo, taken from the OPF format corpus. The licensing info there states “created using ffmpeg’s builtin "testsre" test pattern, which is presumed to be unlicensed. ffmpeg’s code is GPLv2 licensed”.
Sorry, browsing is not allowed in this directory.
Going to the movies!

AVI file with the Controller

"Wassup"
SUPERFRIENDS

AVI file w/out the controller
On creator machine

This document uses the Calisto MT font.

On Linux machine

This document uses the Calisto MT font.
Mode Navigation: Click on:
Headings below to navigate model:
A1 - Overview As Built
A2 - Front As Built
A3 - Right As Built
A4 - Rear As Built
A5 - Left As Built
A6 - Front Proposed
A7 - Right Proposed
A8 - Rear Proposed
A9 - Left Proposed
A10 - Section 1
A11 - Section 2
A12 - Section 3
A13 - Section 4
A14 - Credits
A15 - Contact Us
A16 - Visit our Website
A17 - Back to the Gallery
WARNING
```
C:\> c:/qpdf/bin/qpdf --check encryption_nocopy.pdf
checking encryption_nocopy.pdf
PDF Version: 1.7
R = 4
P = -1044
User password =
extract for accessibility: allowed
extract for any purpose: not allowed
print low resolution: allowed
print high resolution: allowed
modify document assembly: not allowed
modify forms: allowed
modify annotations: allowed
modify other: allowed
modify anything: not allowed
stream encryption method: AESv2
string encryption method: AESv2
file encryption method: AESv2
File is linearized
ERROR: linearized file contains an uncompressed object after a compressed one in a cross-reference stream
WARNING: page 0 has shared identifier entries
WARNING: page 0: shared object 25: in hint table but not computed list
```
<table>
<thead>
<tr>
<th>name</th>
<th>type</th>
<th>emb</th>
<th>sub</th>
<th>uni</th>
<th>object ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdobeCorpID-MyriadRg</td>
<td>Type 1</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>13</td>
</tr>
<tr>
<td>ComicSansMS</td>
<td>TrueType</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>8</td>
</tr>
<tr>
<td>Tahoma</td>
<td>TrueType</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>11</td>
</tr>
<tr>
<td>Haettenschweiler</td>
<td>TrueType</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>16</td>
</tr>
<tr>
<td>TimesNewRoman</td>
<td>TrueType</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>18</td>
</tr>
<tr>
<td>Helvetica</td>
<td>Type 1</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>35</td>
</tr>
<tr>
<td>e:\preflightTests&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Identify potential preservation risks of a PDF –any PDF!- by assessing against PDF/A standard.
How?

Use PDF/A validator!
JPEG Images Displayed in Acrobat

In Acrobat files, bitmap images are displayed in a floating window for easy viewing.

Considerations:
- Images can be linked.
- Graphics is used in its original format.
- Floating window can be resized or maximized (click and drag to make full-screen).
- Graphics is displayed at the float window's magnification or Acrobat document's zoom level.
- Not applicable if the image is embedded.
- Supported with Acrobat Pro.

Preflight

Profiles

Verify compliance with PDF/A-1a
Verifies compliance with PDF/A-1a for the current document.

Verify compliance with PDF/A-1b
Verify compliance with PDF/A-1a
Verify compliance with PDF/A-2a
Verify compliance with PDF/A-2b
Verify compliance with PDF/A-2u
Verify compliance with PDF/A-3a
Verify compliance with PDF/A-3b
Verify compliance with PDF/A-3u

PDF/E compliance
- Convert to PDF/E-1 (sRGB)
- Remove PDF/E information
- Verify compliance with PDF/E-1

PDF/VT compliance

Further Options
- Run Preflight checks for visible layers only
- Preflight only pages 1 to 2

Tools | Sign | Comment
Preflight profile "Verify compliance with PDF/A-1a" found the following errors:

Pages 1 - 2 from "Jpeg_linked.pdf"

- PDF document is not compliant with PDF/A-1a (2005)
- Structured PDF: Structure tree root entry missing
- Author mismatch between Document Info and XMP Metadata
- Document is encrypted
- Encrypt key present in file trailer
- MarkInfo missing
- Named action with a value other than standard page navigation used
- PDF/A entry missing
- Syntax problem: Indirect object "obj" keyword not followed by an EOL marker
- Annotation has C entry but no PDF/A OutputIntent present (6 matches on 2 pages)
- Annotation has no Flags entry (6 matches on 2 pages)
- Annotation not set to print (6 matches on 2 pages)
- Incorrect annotation type used (not allowed in PDF/A) (4 matches on 1 page)
- Incorrect font process color used but no PDF/A OutputIntent (59 matches on 2 pages)
- Font not embedded (and text rendering mode not 3) (20 matches on 1 page)

Overview

Preflight information
Preflight
Invalid Font definition, CourierNewPSMT: FontFile entry is missing from FontDescriptor

Error on MetaData, PDF/A identification schema http://www.aiim.org/pdfa/ns/id/ is missing

Action is forbidden, GoToPage isn't authorized as named action

Trailer Syntax error, The trailer dictionary contains Encrypt
Policy assessment
<?xml version="1.0"?>
<!-- Schematron rules for policy-based validation of PDF, based on output of Apache Preflight. -->
- <s:schema xmlns:s="http://purl.oclc.org/dsdl/schematron">
  - <s:pattern name="Checks for encryption">
    - <s:rule context="/preflight/errors/error">
      <s:assert test="not(code = '1.0' and contains(details,'password'))">Open password</s:assert>
      <s:assert test="not(code = '1.4.2')">Encryption</s:assert>
    </s:rule>
  </s:pattern>
</s:schema>
Test corpus

~ 15,000 PDFs from Govdocs1 dataset

http://digitalcorpora.org/corpora/govdocs
Failed tests assessment
HTTP Error 404. The requested resource is not found.
Main challenges

1. Font issues
2. Conformance to ISO 32000
3. Ground truth
<?xml version="1.0" encoding="UTF-8"?>
<profile model="org.verapdf.model.PDFA1a"
xmlns="http://www.verapdf.org/ValidationProfile">
  <name>ISO 19005-1:2005 - 6.2.4 Images - Alternates</name>
  <description/>
  <creator>veraPDF Consortium</creator>
  <created>2015-06-16T22:22:45Z</created>
  <hash>sha-1 hash code</hash>
  <rules>
    - <rule object="PDXImage" id="6-2-4-t01">
      <description>An Image dictionary shall not contain the Alternates key</description>
      <test>Alternates_size == 0</test>
      <error>
        <message>Alternates key is present in the Image dictionary</message>
      </error>
    </rule>
    - <reference>
      <specification>ISO19005-1</specification>
      <clause>6.2.4</clause>
    </reference>
  </rules>
</profile>
ISO PDF (ISO 32000)

- PDF/A
  - archive
  - ISO 19005 (PDF 1.4)
- PDF/E
  - engineering
  - ISO 24517 (PDF 1.6)
- PDF/X
  - graphic arts
  - ISO 15930 (PDF 1.4 PDF 1.6)
- PDF/UA
  - accessibility
- PDF/H
  - healthcare
- AIIM --> ISO
veraPDF Corpus

The repository contains the veraPDF test corpus for PDF/A specifications (Versions 1B, 1A, 2B, 2U, 2A, 3B, 3U, 3A) as well as a number of additional test files for ISO 32000-1.

This test suite complements Isartor and Bavaria test suites and follows their test file pattern:

- all test files are atomic
- they are self-documented via the document outlines
- the naming pattern and the directory structure indicate relevant parts of ISO 19005-1 specification

© 2015 veraPDF Consortium
Why PDF/A validation matters, even if you don’t have PDF/A

JULY 7, 2015 / JOHAN VAN DER KNIJFF

This is the first installment of a 2-part blog (part 2 is here). It was prompted by the upcoming Digital Preservation Coalition briefing *When is a PDF not a PDF?*, for which I was asked to prepare a presentation. My initial idea was to give an overview of the work we did on PDF preservation risk assessment using a PDF/A validator in the SCAPE project. Most of this has already been covered by a series of earlier blog posts. Those blogs very much represent different stages of a work in progress, and I think this...
Image attribution

- Unencrypted by Brennan Novak from the Noun Project

* http://thenounproject.com/*