Validating PDF/A

This article deals with validation software for PDF/A and discusses the PDF/A Competence Center's plans for developing a test suite.

Thomas Merz, author and PDF expert, is president of PDFlib GmbH. PDFlib is located in Munich and has focused on PDF development since 1997. PDFlib's product family provides components for dynamically creating PDF, including PDF/A-1a (tagged PDF) and PDF/A-1b. PDFlib GmbH is a founding member of the PDF/A Competence Center.

In addition, Thomas Merz is involved in the PDF/A Competence Center's Technical Working Group (TWG), which establishes a consolidated technical position on PDF/A issues. The TWG has written several articles, including the PDF/A Technical Notes (published on pdfa.org) that deal with topics like „PDF/A-1 and Namespaces“, „Colors in PDF/A-1“, „Metadata (XMP/Docinfo) in PDF/A-1“ and „Digital Signatures and PDF/A-1“. The TWG also cooperates with the ISO committee, for example in drafting the PDF/A-2 standard and in the development of a PDF/A test suite.

Basics

The PDF/A-1 standard was published in October 2005 as ISO 19005-1. A Corrigendum was published in the second quarter of 2007 to amend the standard and better explain certain issues. The ISO standard itself comprises only 36 pages, but includes references to several additional, more complex documents (for example the PDF Reference 1.4 and specifications for XMP, font formats, ICC and much more).

Conformance to PDF/A involves not only the requirements in the ISO standard itself, but also in the secondary specifications. Currently there exists no implementation that can be used as a reference (documented or software).

The PDF/A-1a standard distinguishes itself through more detailed requirements than PDF/A-1b. PDF/A-1a requires amongst other things „tagged“ PDF and Unicode.

PDF/A conformance requires a significant effort from the producers of PDF software.

Aspects of PDF/A Conformance

PDF/A conformance encompasses several aspects that can apply to a document at different stages of its lifetime:

- **Creation:** This stages deals with the generation of PDF/A conforming documents from various source formats.
- **Correction:** A PDF document may have to be modified in order to achieve conformance with PDF/A.
- **Processing:** Conformance must be preserved when a PDF/A document is modified.
- **Display:** This refers to the presentation of a PDF/A file in accordance with the requirements. Simply displaying a PDF/A file „somehow", as is the case with many viewers, is insufficient.
- **Validation:** It is often necessary to verify that a PDF/A file actually conforms to the standard.

PDF/A-Validation: Products

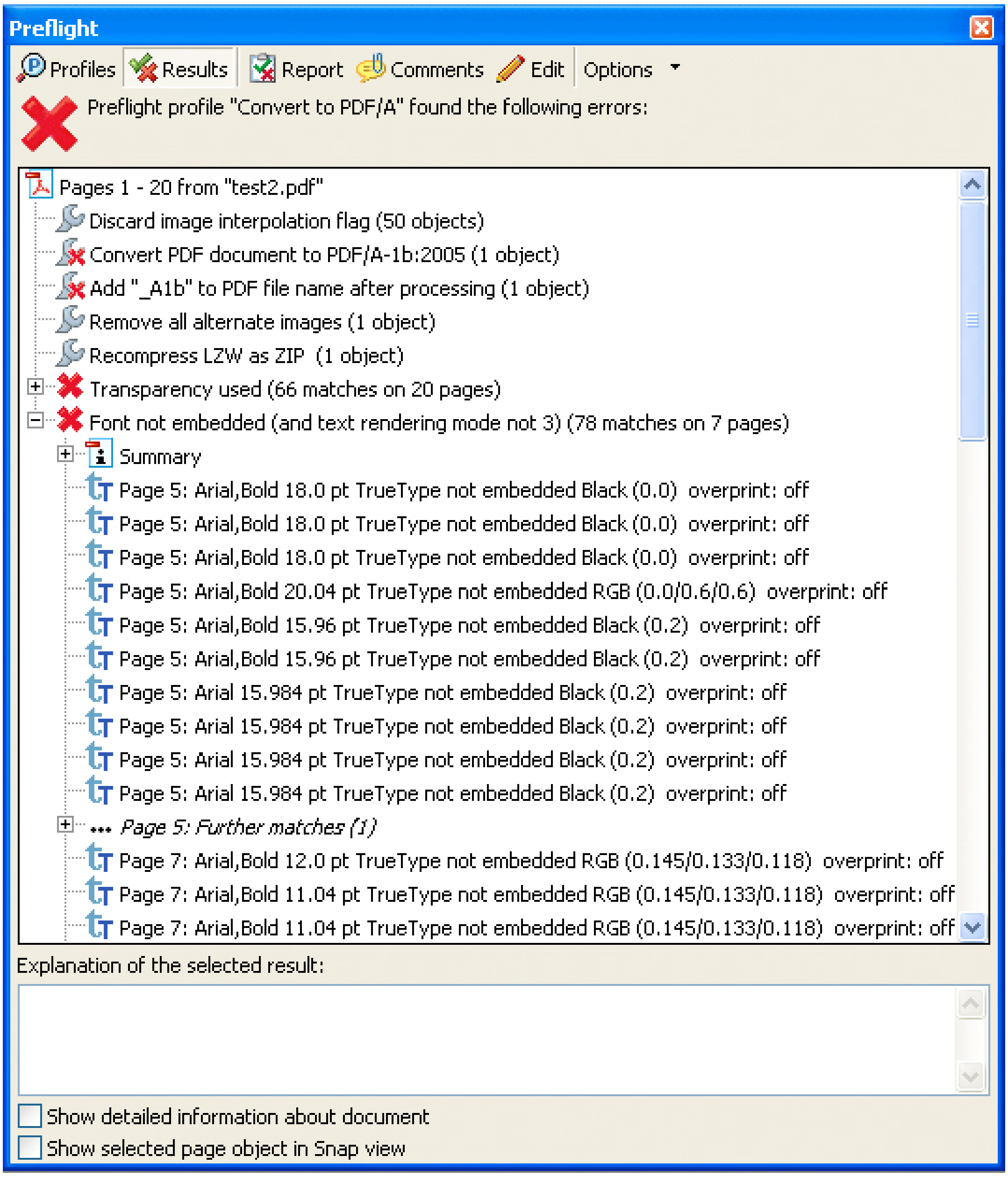
There are several applications available on the market for validating PDF/A documents, including the following validators from PDF/A Competence Center members:

- Acrobat 8 Preflight (developed by callas software)
- PDF Tools AG: 3-Heights PDF Validator
- LuraTech: LuraDocument PDF Validator
- Seal Systems: PDF Checker
- Intarsys: PDF/A Live!
- callas: pdfaPilot
- callas: pdfInspektor
- Apago: PDF Appraiser (sold byh Actino)

Additional products are also available from other vendors.

Example: Acrobat 8 Preflight

The first example for PDF/A validation is the Acrobat „Preflight" function, developed by Adobe and callas software. The PDF/A functionality found in Acrobat 7 Preflight is based only on a draft version of the standard, the approved PDF/A standard is implemented in Acrobat 8.



Preflight profile "Convert to PDF/A" found the following errors:

```
Pages 1 - 20 from "test2.pdf"
    Discard image interpolation flag (50 objects)
    Convert PDF document to PDF/A-1b:2005 (1 object)
    Add "_A1b" to PDF file name after processing (1 object)
    Remove all alternate images (1 object)
    Recompress LZW as ZIP  (1 object)
  + Transparency used (66 matches on 20 pages)
  - Font not embedded (and text rendering mode not 3) (78 matches on 7 pages)
      + Summary
        Page 5: Arial,Bold 18.0 pt TrueType not embedded Black (0.0)  overprint: off
        Page 5: Arial,Bold 18.0 pt TrueType not embedded Black (0.0)  overprint: off
        Page 5: Arial,Bold 18.0 pt TrueType not embedded Black (0.0)  overprint: off
        Page 5: Arial,Bold 20.04 pt TrueType not embedded RGB (0.0/0.6/0.6)  overprint: off
        Page 5: Arial,Bold 15.96 pt TrueType not embedded Black (0.2)  overprint: off
        Page 5: Arial,Bold 15.96 pt TrueType not embedded Black (0.2)  overprint: off
        Page 5: Arial 15.984 pt TrueType not embedded Black (0.2)  overprint: off
        Page 5: Arial 15.984 pt TrueType not embedded Black (0.2)  overprint: off
        Page 5: Arial 15.984 pt TrueType not embedded Black (0.2)  overprint: off
        Page 5: Arial 15.984 pt TrueType not embedded Black (0.2)  overprint: off
      + ••• Page 5: Further matches (1)
        Page 7: Arial,Bold 12.0 pt TrueType not embedded RGB (0.145/0.133/0.118)  overprint: off
        Page 7: Arial,Bold 11.04 pt TrueType not embedded RGB (0.145/0.133/0.118)  overprint: off
        Page 7: Arial,Bold 11.04 pt TrueType not embedded RGB (0.145/0.133/0.118)  overprint: off
```

Explanation of the selected result:

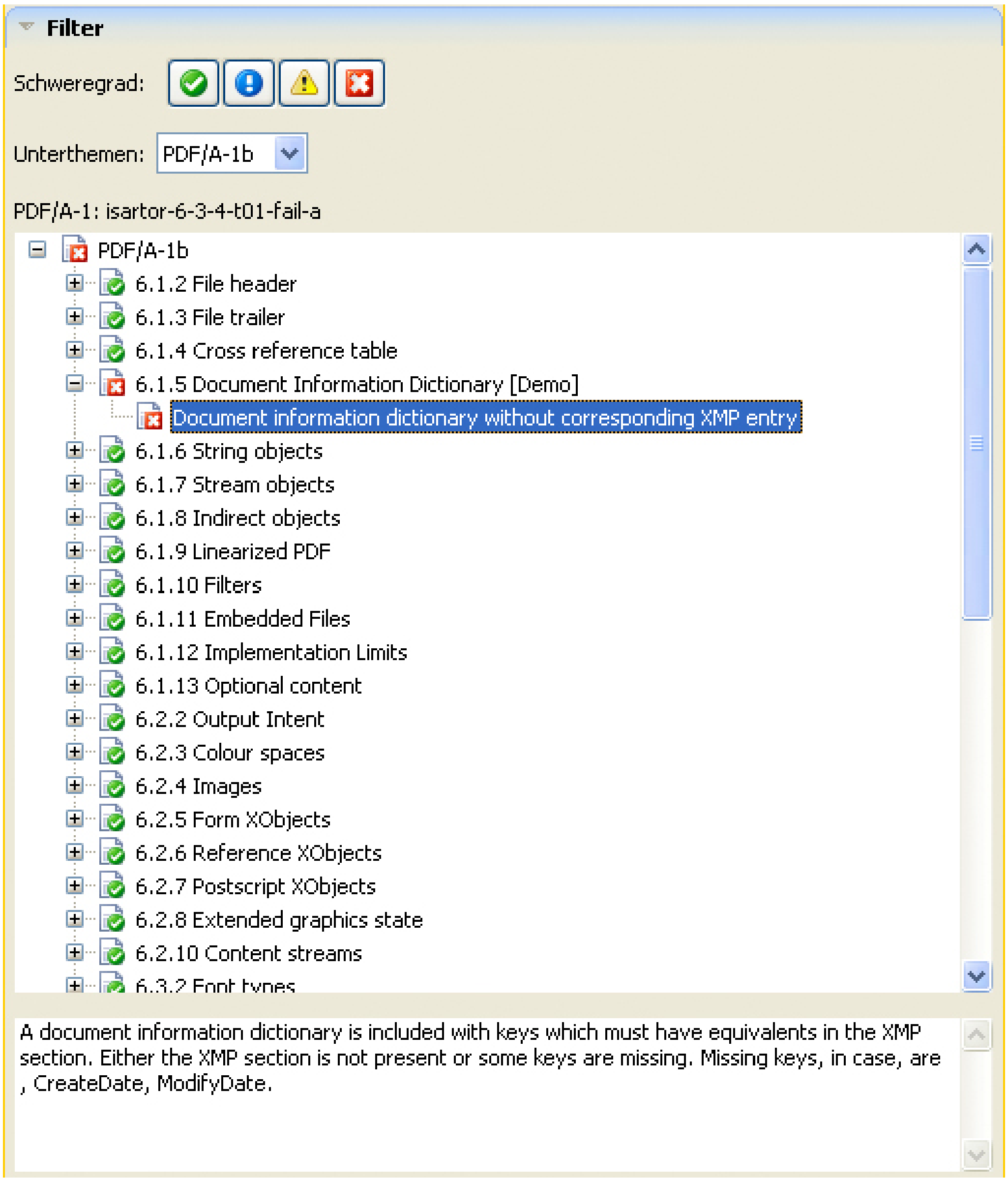
☐ Show detailed information about document

☐ Show selected page object in Snap view

Preflight is capable of validating PDF files on their conformance to PDF/A. The above example shows the results of a validation, where the PDF file does not conform to PDF/A.

Example: Cabaret Stage 3

The Cabaret Stage 3 product (marketed through Intarsys) is a software program for displaying, filling, editing, saving, printing and validating PDF documents.



PDF/A-1: isartor-6-3-4-t01-fail-a

- ✗ PDF/A-1b
  - ✓ 6.1.2 File header
  - ✓ 6.1.3 File trailer
  - ✓ 6.1.4 Cross reference table
  - ✗ 6.1.5 Document Information Dictionary [Demo]
    - ✗ Document information dictionary without corresponding XMP entry
  - ✓ 6.1.6 String objects
  - ✓ 6.1.7 Stream objects
  - ✓ 6.1.8 Indirect objects
  - ✓ 6.1.9 Linearized PDF
  - ✓ 6.1.10 Filters
  - ✓ 6.1.11 Embedded Files
  - ✓ 6.1.12 Implementation Limits
  - ✓ 6.1.13 Optional content
  - ✓ 6.2.2 Output Intent
  - ✓ 6.2.3 Colour spaces
  - ✓ 6.2.4 Images
  - ✓ 6.2.5 Form XObjects
  - ✓ 6.2.6 Reference XObjects
  - ✓ 6.2.7 Postscript XObjects
  - ✓ 6.2.8 Extended graphics state
  - ✓ 6.2.10 Content streams
  - ✓ 6.3.2 Font types

A document information dictionary is included with keys which must have equivalents in the XMP section. Either the XMP section is not present or some keys are missing. Missing keys, in case, are , CreateDate, ModifyDate.

Validation with Cabaret Stage 3: The inspected file does not conform to PDF/A.

PDF/A-Validation: Aspects

Several aspect must be taken into account when verifying the PDF/A conformance of documents. Both the breadth and the depth of the verification are important factors when validating.

Breadth of the Verification:

Validation must take into consideration all of the requirements in the standard. This means that all of the rules, requirements and restrictions in the standard must be verified. Some of the requirements apply to the entire document, for example XMP and color spaces. Other requirements only apply to special data structures, and are not necessarily present in all PDF documents. Font subsets and annotations are two such examples.

Depth of the Verification:

When considering the depth of the verification, it must be decided on how exact the data structure will be examined and to what extent the individual rules will be verified. The following secondary specifications must also be considered:

- Fonts: TrueType/OpenType/PostScript Type 1
- ICC color profiles
- XMP -> RDF -> XML, Namespaces

The combination of different properties can also be relevant:

- For example fonts and form fields: are the fonts that are used in form fields also embedded?

The area of PDF/A-1a and tagged PDF:

- There is quite a degree of latitude here for „meaningful‟ structural information.

Further Aspects

Additional factors play a role in the actual method of validation. Is an interactive or a more batch-oriented environment practical for the verification? The level of detail for reporting errors will differ depending on the intended use. An automatic correction will also often be desirable upon completion of the validation.

PDF/A-Validation: Special Considerations

A PDF/A Validator must take the Corrigendum to the standard into account. This is the case with Acrobat 8, but not with Acrobat 7.

An insufficient breadth or depth in the validation will deliver incomplete and unreliable results.

Different types of failures are possible with validation. A validator could report an error, even though the document is actually PDF/A conforming. The opposite is also possible, when the validator does not report an error even though one is present.

A further problem that could appear during validation is when a valid input document is rejected before the check is even performed. A report that does not provide enough detail can also be problematical.

Test suite for PDF/A

The PDF/A Competence Center and the TWG are preparing a test suite with PDF/A documents. The goal is to create standardized documents that deal with different aspects of PDF/A. The strategy is to then purposely introduce breaches to the specification. The „synthetic test suite‟ will be created in stages, and the breadth and depth of coverage will be documented in the process. An important final step is the comparison with the different validators that are available on the market.

Starting Point:

In focus are (presumed or actual) creation programs, as well as different methods for the manual creation of PDF/A.

How will the TWG confirm the accuracy of the test suite? There will be a check using different validation tools, and a manual inspection using special analysis tools will also be performed.

The combined expert knowledge of the TWG members will be applied for creating the test suite, and an iterative approach will ensure accuracy.

Validating the Validator

The purpose here will be to check different validation products through use of the test suite. The accuracy of the check will increase as the test suite is expanded. The criteria for assessing the validation tools will be based on the following questions:

- Which errors are recognized?
- Which valid elements were incorrectly rejected?
- How meaningful are the failure reports?

Next Steps

Once the test suite is completed, the test procedure will be formalized. The tests should then be conducted by an independent test center. In the end, it will be possible to use the test suite to officially certify validation products.

Practical Recommendations for Validation

A reduction in performance can be expected when validating high volumes of documents. For this reason it is recommended to validate products and workflows, instead of each PDF/A document separately. It will however probably be necessary to validate all externally supplied PDF/A documents individually, since the creation process is usually not known. Don't forget: non-conforming documents cannot always be converted to PDF/A! For example, in order to embed a missing font, the font must of course be present and available.

Thomas Merz, PDFlib GmbH, aoe
Translated from the German original by Roger Reeves, Reeves & Partner GmbH

Appendix: PDFlib GmbH

- Focused on PDF development since 1997
  - Approx. 11,000 licenses in more than 60 countries
  - Hundreds of OEM customers
- PDFlib product family: components for dynamically creating PDF
  - PDF/A-1a (Tagged PDF) und PDF/A-1b
- PLOP: Encryption, optimization, signatures
  - Preservation of PDF/A-1 (if permitted by the standard)
- Text Extraction Toolkt (TET): PDF to Text/XML/RTF
  - Normalization of all text to Unicode
  - Recognition of word boundaries and text flow