Life after PREFORMA

The future of veraPDF

Industry-supported
PDF/A validation

Introduction

veraPDF

Overview

We'll be covering:

- the project background to put veraPDF development in context;
- example policy elements with a GUI demonstration;
- some common PDF/A validation issues concerning Unicode and text; and
- our future plans for veraPDF as the PREFORMA project ends.

The PREFORMA project

veraPDF development has been funded by the PREFORMA project

PREservation FORMAts for culture information/e-archives, is a Pre-Commercial Procurement (PCP) project co-funded by the European Commission under its FP7-ICT Programme. The project's main aim is to address the challenge of implementing standardised file formats for preserving digital objects in the long term, giving memory institutions full control over the acceptance and management of preservation files into digital repositories. http://www.preforma-project.eu/

veraPDF

PREFORMA timetable

- Design phase: November 2014 to January 2015
- Suppliers selected: April 2015
- First prototyping phase: April 2015 to October 2015
- Redesign phase: October 2015 to January 2016
- Second prototyping phase: February 2016 to January 2017
- Acceptance testing phase: March 2017 to August 2017
- PREFORMA ends: December 2017

veraPDF

The veraPDF consortium

Led by the

- Open Preservation Foundation
- PDF Association

With partners

- Dual Lab
- KEEP Solutions
- DPC

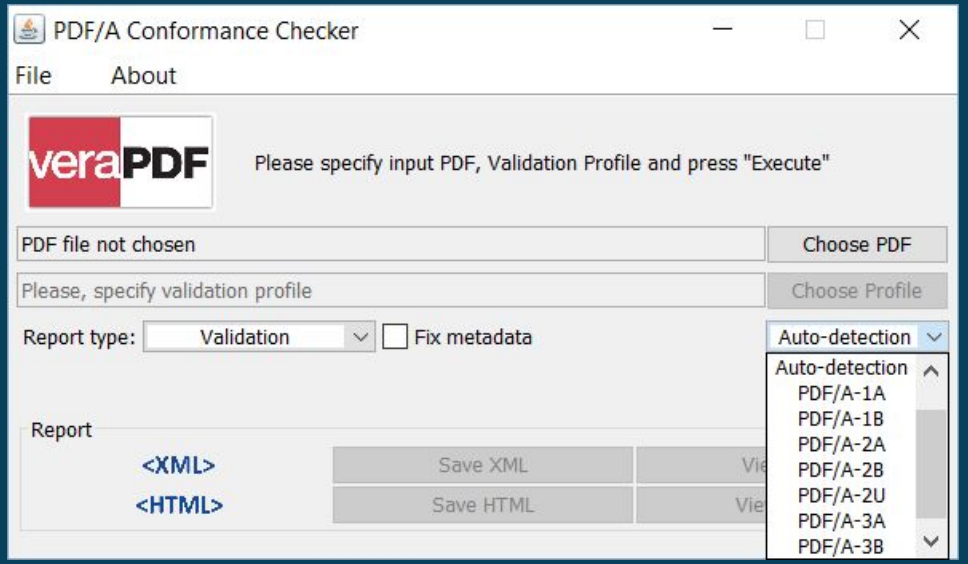
veraPDF

PDF/A validation

Latest release: 1.8 (Aug 17)

Conformance checker: full support for all PDF/A versions (1,2,3) and levels (A,B,U)

GUI - desktop version for single file evaluation
CLI - command line interface targeting large volume batch processing
Web - Online demo web site
Java library - Calling a Java API from custom Java-based applications



veraPDF

Open source for sustainability

Software licenced under dual MPL v2+ / GPL v3+ allowing anyone to download or integrate the software free of charge

Test datasets and documentation licensed under CC-BY-4

Active open source community

veraPDF

Enforcing institutional policy

veraPDF

Policy: restricting acceptance criteria

Institutions may wish to enforce stricter acceptance criteria for the PDF/A documents that they preserve. These criteria will often make demands beyond the PDF/A specifications, for example:

- Disallowing particular image compression types
- Restricting the set of fonts used in documents
- Quality assuring metadata embedded in PDFs

veraPDF

Policy: relaxing acceptance criteria

Alternatively, policy can be used to relax institutional PDF/A acceptance criteria, examples might be:

- Allowing some or any external fonts
- Accepting forms of compression disallowed by PDF/A

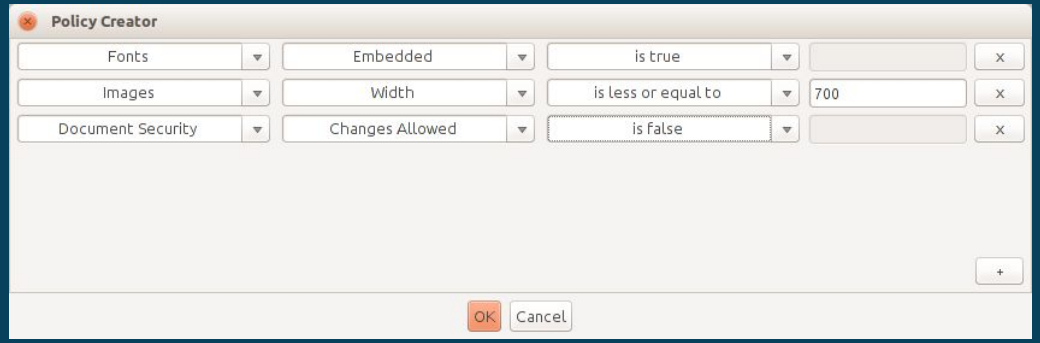
*These are not policy recommendations, just examples. PDFs with these properties are not valid PDF/As but pragmatism might have to prevail. One case might be that there is a known problem with a set of PDFs and the collecting institution cannot change them.

veraPDF

Policy GUI

veraPDF policies are written schematron documents which is a pattern and rule enforcement language for XML documents, written in XML. This requires knowledge of XML, XPath and XQuery.

In order to make life easier the veraPDF team has developed a policy assistant for the GUI application to help non eXperts.



Policy demonstration

We’ll now demonstrate the policy checker and GUI showing a couple of examples:

- Simple metadata checking, ensuring that the document has a populated author and title field
- Ensuring that attached files are XML only by checking the reported MIME type

veraPDF

PDF/A validation and Unicode

veraPDF

Text extraction risks

- There are no deterministic algorithms to extract text from the generic PDF document:
  - No Unicode mapping: Font glyphs can be referenced directly by their internal glyph IDs in the embedded font program
  - Logical order is not well-defined: text characters can be drawn on the page in an arbitrary order
  - Other risks:
    - Text is represented by line-art or scanned images (OCR needed)
    - Document is encrypted with no permissions to extract text
- This may prevent text extraction and indexing of PDF documents

veraPDF

PDF/A and text extraction

- Level B concerns only visual presentation
- Level U handles risks of mapping text characters to Unicode
- Level A adds to this:
  - the correct logical order
  - the presence of alternative text in cases some glyphs map to Private Unicode Areas (PUA). This typically happens when custom special purpose fonts are used with private glyph collection (eg., company logo or military cartographic symbols)
- *NB. Levels U and A are combined into a single Level A for PDF/A-1*

veraPDF

veraPDF implementation - Level U

- Rule 6.2.11.7-1: The Font dictionary of all fonts shall define the map of all used character codes to Unicode values, either via a **ToUnicode** entry, or other mechanisms as defined in ISO 19005-2:2011, 6.2.11.7.2.
- Rule 6.2.11.7-2: The Unicode values specified in the **ToUnicode** CMap shall all be greater than zero (0), but not equal to either U+FEFF or U+FFFE.

veraPDF

veraPDF implementation - Level A

- Rule 6.2.11.7-3: For any character, regardless of its rendering mode, that is mapped to a code or codes in the Unicode Private Use Area (PUA), an **ActualText** entry as described in ISO 32000-1:2008, 14.9.4 shall be present for this character or a sequence of characters of which such a character is a part.
- Rule 6.7.3-1: The logical structure of the conforming file shall be described by a structure hierarchy rooted in the **StructTreeRoot** entry of the document's **Catalog** dictionary, as described in ISO 32000-1:2008, 14.7.

veraPDF

Life after PREFORMA

veraPDF

Support for ISO 32000-2 (PDF 2.0)

- ISO 32000-2 was published in July 2017 (no longer freely available, should be purchased at ISO web site)
- The new PDF/A standard based on PDF 2.0 will appear some time in 2019(?)
- veraPDF supports PDF 2.0 for feature extraction and policy checks
- Key new features:
  - Low-level syntax changes: UTF 8 encoding for PDF Strings
  - New tagged PDF: completely reworked
  - Up to date security: AES-256 encryption, Unicode passwords, CAdES digital signatures

veraPDF

The future

PREFORMA funding ended in July 2017. veraPDF is transitioning from a funded project to a stand alone open source project.

The OPF and Dual Lab will:

- continue to address bugs reported on GitHub;
- test and merge small pull requests submitted to GitHub; and
- provide answers and support for minor mailing list enquiries.

veraPDF

Future developments?

veraPDF is functionally complete for PDF/A but there is still work to be done, for example:

- developing validators for other parts of the PDF standard
- veraPDF supports a plug in mechanism that can be used to validate other formats embedded or attached to PDF/A documents
- integrating veraPDF with commercial preservation systems and other services
- the creation of more complex policy schematron that reflect collections of real institutional acceptance criteria

veraPDF

Coverage of existing PDF standards

PDF/A
ISO 19005

PDF/UA
ISO 14289

PDF
ISO 32000

PDF/X
ISO 15930

PDF/VT
ISO16612-2

PDF/A-4

PDF/E
ISO 24517

veraPDF

veraPDF ZUGFeRD plugin

ZUGFeRD is an electronic invoicing standard developed in Germany. ZUGFeRD uses PDF/A-3 for invoices where the document is intended for people and an XML attachment provides a machine actionable version of the invoice.

The ZUGFeRD team have developed a plugin that recognises and validates the attached XML invoices: https://github.com/ZUGFeRD

Integration with other services

The veraPDF consortium is in an active development partnership with Logius, the digital government service of the Netherlands Ministry of the Interior and Kingdom Relations (BZK).

Integrating veraPDF and Heritrix to provide a service allowing users to crawl a particular domain report on document compliance, e.g. how many valid PDF/A documents, profiling by PDF version etc. and reporting on MS Office and Openoffice formats found.

Public beta to be announced soon.

veraPDF

Partnership

Memory institutions alone cannot have expertise in-house for every format they collect and preserve.

The PDF family of formats is widely used in digital preservation. PDF/A is just one specification, there are no complete open source validators for the others.

OPF and PDF Association have bridged a gap between memory institutions and industry to create a successful product.

veraPDF

Find out more

- http://verapdf.org/
- users@lists.verapdf.org (Q&A, discussions)
- https://github.com/veraPDF
- http://verapdf.org/subscribe/ (news)
- @_verapdf
- info@verapdf.org

This work was co-funded by the PREFORMA project and European Commission under the FP7-ICT Programme.

veraPDF

PREFORMA