veraPDF: open source PDF/A validation through pragmatic partnership

R. McGuinness
Open Preservation Foundation
c/o The British Library, Boston Spa,
Wetherby,
United Kingdom
becky@openpreservation.org

C. Wilson
Open Preservation Foundation
c/o The British Library, Boston Spa,
Wetherby
United Kingdom
carl@openpreservation.org

D. Johnson
Association for Digital Document
Standards e.V.
PDF Association,
Neue Kantstrasse 14
14057 Berlin, Germany
duff.johnson@pdfa.org

B. Doubrov
Dual Lab spr.
Clos du Parnasse 12C, Brussels,
Belgium
boris.doubrov@duallab.com

ABSTRACT

veraPDF [1] is an open source industry-supported PDF/A conformance checker developed by the members of the veraPDF consortium. The software validates all current parts and conformance levels of ISO 19005 (PDF/A[1]). This paper makes the case for open source format validation and describes the project's approach to building and testing the software. It also explores how the unique partnership between cultural heritage organisations and PDF industry has created an active open source community.

KEYWORDS

Validation, conformance checker, open source, digital preservation, file formats, PDF/A, standards

1 THE PREFORMA CHALLENGE

veraPDF is one of three conformance checkers developed with funding from the PREFORMA [2] (PREservation FORMAts for culture information/e-archives) project. PREFORMA's aim is to empower cultural heritage institutions to gain control over the technical properties of preservation files. To achieve this they issued a call for tender [2], alongside a challenge brief [3] for the development of open source conformance checkers for document (PDF/A), audio-visual (Matroska, LPCM, FFV1), and image formats (TIFF). In addition to the delivering software, the three 'suppliers' i.e. the company or consortium selected to carry out the work, were directed to "establish a healthy ecosystem around an open source 'reference' implementation of specific file formats [3]".

veraPDF comprises four components:

- An **implementation checker**, which validates all parts and conformance levels of the PDF/A specifications;
- A **policy checker**, which allows users to implement additional custom checks to enforce institutional policy with respect to the format;
- A **reporter**, which processes the results, producing reports, both human readable and machine parsable; and
- A **metadata fixer**, which repairs metadata in files based on conformance with the standard.

PREFORMA also requested the development of an integrating shell capable of controlling all the PREFORMA-project conformance checkers, and potentially others that implement the same architecture and APIs.

The schedule for the conformance checker suppliers was divided into three phases:

1. A four month competitive design phase where suppliers were funded to submit their designs. At this early stage PREFORMA selected two potential suppliers for each format. PREFORMA evaluated the six proposals and chose one supplier for each of the conformance checkers to develop prototype software.
2. A twenty month prototyping phase starting in April 2015. Only the three selected suppliers progressed to this phase, which also prescribed a two month design review process that took place at the end of 2015. Following a year of development, the final version 1.0 prototypes were released in December 2016.
3. A six month testing phase that is ongoing at the time of writing. During this phase PREFORMA carries out acceptance testing of the delivered prototypes.

[1] http://verapdf.org/
[2] http://www.preforma-project.eu/

2.1 Open Source for Sustainability

Adoption of open source software for digital preservation in cultural heritage organisations is high. A survey[3] conducted by the OPF (Open Preservation Foundation[4]) in 2015 showed that 88% of responding institutions used open source software. Cultural heritage institutions are increasingly both using and contributing to the development of open source tools [4].

PREFORMA did not aim to just develop software, they are just as concerned with establishing active open source communities around the conformance checker projects. The project required suppliers to maintain an up-to-date, public source code repository, and upload monthly software releases to PREFORMA's open source portal[5].

PREFORMA's intention was to ensure that the source code, software and documentation remained available beyond the lifetime of the project. This strategy aligns with the OPF's approach to sustaining open source digital preservation software as described in the OPF software maturity model[6]. The OPF has experience of both sustaining open source projects including PLANETS[7] and SCAPE[8], as well as providing stewardship for standalone tools such as JHOVE[9].

Building on its foundations in both the archival and PDF software industry communities, the veraPDF consortium has successfully established a dialog between the PDF industry and cultural heritage organisations through digital preservation and industry events and webinars, an online and social media presence, and the open development approach.

2.1.1 Licensing. PREFORMA mandated that each conformance checker be dual licensed under the GNU General Public License version 3[10] or later and the Mozilla Public License version 2.0[11] or later. All other project outputs such as test data sets and documentation were to be licensed under CC-BY-4[12] The veraPDF test corpus is licensed and is freely reusable for testing PDF/A validators.

3 SELECTING FORMATS FOR VALIDATION

Identifying file formats and determining the extent to which individual files conform to appropriate specifications is an essential step in many digital preservation workflows. It demands both detailed understanding of a file's technical properties, and the format specification. However, cultural heritage organisations do not have the resources to employ in-house, specialist knowledge for every file format which they are obliged to preserve.

According to Artefactual Systems[13], developers of open source digital preservation system Archivematica[14]; "one of the most influential factors in selecting preservation formats is community adoption[15]". PDF/A was one of the file formats selected by PREFORMA because the specification is:

- Complete: final definitive documentation is available;
- Open and accessible: available to anyone, free of charge, or for a one-off nominal fee, so it can be reused without any limitations; and
- widely used by cultural heritage institutions and national archives[16].

While there are commercial PDF/A validators available no complete or authoritative open source implementations existed.

4 PARTNERING WITH INDUSTRY

The veraPDF consortium brings together a network of stakeholders with complementary perspectives. Led by the OPF and the PDF Association[17], the industry trade group for developers of PDF software, with partners Dual Lab[18], DPC[19] and KEEP SOLUTIONS[20], the veraPDF consortium combines digital preservation and cultural heritage expertise with document industry backing and comprehensive access to the PDF/A standardization process.

4.1 The veraPDF Approach

In order to fully understand the format specification, the veraPDF consortium decided to create a test corpus covering all parts and conformance levels of the PDF/A standards. The idea was that creating test files would highlight any lack of understanding on the veraPDF consortium's part or reveal ambiguities in the standards.

When there is a problem interpreting the specification the veraPDF consortium brings the issue to the PDF Association's PDF Validation Technical Working Group (TWG), comprised of PDF technology experts. It was established to analyse PDF validation issues in a transparent fashion and also connect veraPDF to the ISO committee responsible for PDF/A.

The ISO working group responsible for PDF/A (ISO TC171 SC2 WG5)[21] had previously decided that the existing ISO specifications for archival PDF (PDF/A-1, PDF/A-2 and PDF/A-3) would not be revised, even in the light of ambiguities

[3] http://www.openpreservation.org/documents/public/OPFDigitalPreservationCommunitySurvey2015.pdf (6.0)
[4] http://openpreservation.org
[5] http://www.preforma-project.eu/open-source-portal.html
[6] http://openpreservation.org/technology/principles/software-maturity/
[7] http://www.planets-project.eu/
[8] http://scape-project.eu/
[9] http://jhove.openpreservation.org/
[10] https://www.gnu.org/licenses/gpl-3.0.html
[11] https://www.mozilla.org/en-US/MPL/2.0/
[12] https://creativecommons.org/licenses/by/4.0/
[13] https://www.artefactual.com/
[14] https://www.archivematica.org/en/
[15] https://blogs.loc.gov/thesignal/2012/10/archivematica-and-the-open-source-mindset-for-digital-preservation-systems/
[16] http://experienceworkshop.preforma-project.eu/wp-content/uploads/2016/11/20161122_BERT_CHALLENGEBRIEF.pdf
[17] http://www.pdfa.org
[18] http://www.duallab.com/
[19] http://www.dpconline.org/
[20] https://www.keep.pt/en
[21] https://www.iso.org/committee/337930.html

uncovered by the veraPDF consortium's project. This is understandable as changing a specification undermines any software developed against the previous version. Instead the PDF Validation TWG asked WG5 to review the ambiguities uncovered during the development of veraPDF in the context of the development of a PDF Association Technical Note. PDF Association Technical Notes have a good track-record of adoption by the industry. The development of veraPDF has also directly influenced the standardisation process, with several issues raised leading to enhancements in a forthcoming new part for PDF/A. As such, the PDF Validation TWG was able to influence the development of next-generation PDF/A specifications.

veraPDF has thus served as a vehicle enabling cultural heritage organisations to directly influence the standardisation process, benefitting both the digital preservation community and industry.

5 ESTABLISHING GROUND TRUTH

The consortium's aim when creating test data was to produce a comprehensive ground truth [5] corpus for the PDF/A standards. veraPDF's developers carefully examined each clause in the standards, and developed a formal grammar to describe the requirements in a machine-readable fashion. They then produced validation rules with an accompanying programmatic test for each requirement. PASS and FAIL corpus files were created to test the validator's functionality. Where unable to create a rule, file or test, or when 3rd party validators were observed to disagree, the issue was raised with the TWG. The process is illustrated in Figure 1.

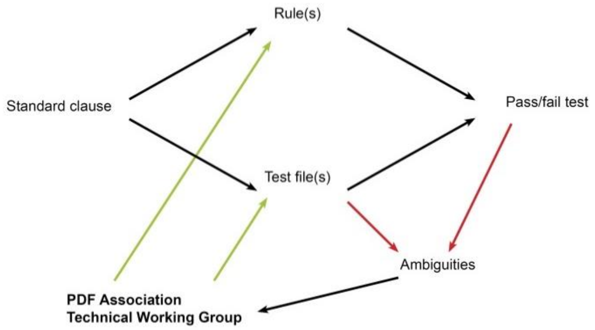

Figure 1: veraPDF test corpus development process

The veraPDF test corpus[22] comprises over 1,500 PDF files. All veraPDF development and release software is tested against the corpus and the results are published online[23]. The veraPDF developers also test against the existing PDF/A test corpora produced by the PDF Association:

- Isartor PDF/A-1b test suite[24]
- The BFO PDF/A-2 test suite[25]

In addition to resolving problems interpreting the standards, the PDF Validation TWG reviewed and approved the final set of validation rules and test files.

OPF and DPC also performed real world testing with their respective members. Focusing on reliability, performance and usability rather than validator functionality, this testing provided valuable feedback for bug fixes and optimisations.

KEEP SOLUTIONS carried out software reviews and tested external integrations by incorporating veraPDF into RODA[26], their digital repository solution.

6 VERAPDF TODAY AND FUTURE PLANS

At the time of writing the veraPDF software is available as version 1.6[27]. Two distinct implementations are available. Early prototypes of veraPDF used Apache PDFBox[28] as its PDF parser, and to implement the veraPDF validation model[29]. Although the veraPDF consortium was aware this was incompatible with PREFORMA's licensing requirements the consortium needed a parser, and obtained dispensation to use PDFBox to test its rule-based approach to validation.

Version 0.26 marked the first dual release in which veraPDF launched its own, purpose-built PDF parser, (the "greenfield" parser), alongside the PDFBox implementation. The greenfield version fully meets the dual licensing requirements with minimal external dependencies.

As well as enabling cultural heritage organisations to ensure their files comply to the PDF/A standards, veraPDF also provides the means to enforce institutional policies. The veraPDF policy checker allows the user to create their own acceptance criteria ("policy") using XML Schematron[30] syntax, enabling organisations to enforce restrictions in line with local policy, but beyond those stated in the PDF/A standards.

[22] https://github.com/veraPDF/veraPDF-corpus
[23] http://tests.verapdf.org/
[24] https://www.pdfa.org/isartor-test-suite/
[25] https://github.com/bfosupport/pdfa-testsuite
[26] http://www.roda-community.org/
[27] http://verapdf.org/2017/06/06/verapdf-1-6-released/
[28] https://pdfbox.apache.org
[29] https://github.com/veraPDF/veraPDF-model-syntax
[30] http://schematron.com/

6.1 Policy Example

Many cultural heritage organisations will not accept PDF/A-3 documents because PDF/A-3 allows the attachment of arbitrary file formats to the archival PDF. This runs contrary to the common approach of limiting the number of accepted formats to minimize preservation risk. The veraPDF policy checker can help organisations identify and analyse files attached to PDF/A-3 documents to ensure they meet acceptance criteria.

6.2 The Future

The middle of 2017 marks a pivotal moment for the veraPDF project. As PREFORMA's test phase comes to an end, so does the governance and funding they have provided. veraPDF will become an independent, open source project. The OPF, PDF Validation TWG and Dual Lab will continue to address issues raised by users as well as testing and incorporating community contributions. Beyond this, future development will require funding.

The veraPDF consortium is developing plans to establish a community-led steering group, and explore open source business models [31] such as annual subscription or sponsorship, grant funding and consultancy for software maintenance and to extend the software's functionality.

Unaddressed so far is the fact that PDF/A represents a small fraction of the document files cultural heritage organisations ingest; the vast majority are simply PDF (ISO 32000) files.

Developing a conformance checker for the PDF format as a whole is a huge undertaking, requiring many man-years of effort and resources. However, as the veraPDF validation model is format-neutral it can be extended to cover all aspects of PDF as well as PDF subset standards such as the forthcoming PDF/A-next specification, PDF/X (print), PDF/E (engineering), PDF/UA (universal accessibility), related specifications, and even third-party standards.

7 CONCLUSIONS

Cultural heritage organisations cannot be expected to provide expertise across all the formats they are responsible for preserving. Active dialog and collaboration with industry sectors who have relevant specialist expertise [5] is essential to the preservation of digital heritage, and the veraPDF project has proven that such collaborations are possible.

Open source software has properties that offer a compelling case for its use in preservation systems. Source code availability, publication of quality assurance results and the opportunity to play a role in developing and testing software can provide organisations with confidence that is difficult to establish when dealing with a commercial vendor and proprietary solutions.

The work described in this paper shows that collaborations between the cultural heritage sector and industry can be effective in building software that meets the needs of the digital preservation community and incorporates the specialist expertise of other sectors.

ACKNOWLEDGMENTS

The funding for this work was awarded to veraPDF by the PREFORMA project, a Pre-Commercial Procurement[32] (PCP) co-funded by the European Commission under its FP7-ICT Programme under grant agreement no 619568. It runs from 1 January 2014 - 31 December 2017.

REFERENCES

[1] ISO 19005-1:2005 Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1) https://www.iso.org/standard/38920.html, ISO 19005-2:2011 Document management -- Electronic document file format for long-term preservation -- Part 2: Use of ISO 32000-1 (PDF/A-2) https://www.iso.org/standard/50655.html, ISO 19005-2:2011 Document management -- Electronic document file format for long-term preservation -- Part 2: Use of ISO 32000-1 (PDF/A-2) https://www.iso.org/standard/50655.html

[2] Per Elfner. 2007. PREFORMA Invitation to Tender http://www.digitalmeetsculture.net/wp-content/uploads/2014/06/PREFORMA_Invitation-to-Tender_v1.0.pdf

[3] Bert Lemmens. PREFORMA Challenge Brief. http://www.digitalmeetsculture.net/wp-content/uploads/2014/06/PREFORMA_Challenge-Brief_v1.0.pdf

[4] Angela Dappert, Rebecca Squire Guenther, Sébastien Peyrard (Eds.) 2016. *Digital Preservation Metadata for Practitioners: Implementing PREMIS*. 6.2.25 DOI 10.1007/978-3-319-43763-7.

[5] Nicola Ferro. 2004. Proposal for an Evaluation Framework for Compliance Checkers for Long-term Digital Preservation. In *Proceedings of the 12th Italian Research Conference on Digital Libraries* (IRCDL 2016). Florence, Italy.

[31] http://www.digitalmeetsculture.net/wp-content/uploads/2017/03/4.-McLellan-Business-models.pdf

[32] https://ec.europa.eu/digital-single-market/en/pre-commercial-procurement