PROTOTYPING PHASE 2
FINAL REPORT

Project Acronym:	**PREFORMA**
Grant Agreement number:	**619568**
Project Title:	**PREservation FORMATs for culture information/e-archives**

veraPDF

Revision: final

Authors:

Carl Wilson (Open Preservation Foundation)
Becky McGuinness (Open Preservation Foundation)
Joachim Jung (Open Preservation Foundation)
Duff Johnson (PDF Association)
Boris Doubrov (Dual Lab)

Dissemination Level		
P	**Public**	**X**

1 INTRODUCTION

During the PREFORMA Prototyping phase, suppliers are expected to provide software prototypes that fulfil the requirements of the PREFORMA project, to demonstrate the results of their development work, and to provide explanations and documentation (manuals) on how the developed software can effectively be used in archiving scenarios at memory institutions regardless of their size and the file type they make use of.

Following the same approach used last year, during the Second Prototyping Phase the plan for releases is as follows:

- Frequent releases: monthly;
- Intermediate releases: end of July 2016 and end of October 2016.

The intermediate release shall contain two parts:

- A functionally stable release
- A report which
    - Describes
        - In more detail the respective release;
        - The time line along with the current position (on time, delayed, ahead)
        - How suppliers managed to provide the required functionality (so far);
        - What is still missing compared to the original specifications and what is the plan to implement it.
    - Provides basic information to be used by PREFORMA WP8 in their deliverables to be submitted to the EC, reporting the work done by both suppliers and PREFORMA consortium members during the prototyping phase.

2 PROTOTYPING PHASE 2 - FINAL REPORT

1. Details
Type of Organisation: Not for profit foundation
Registered Name of Organisation: Open Preservation Foundation
Registered Address: 66 Lincoln's Inn Fields
Town/ City: London
Postcode: WC2A 3LH
County:
Country: United Kingdom
Report Author:
Telephone Number: +44 01937 546013
E-mail Address:
Project Name: veraPDF
Report Type: Prototyping Phase 2 - Final Report
Total Contract Price [euro]: 669,525
Start Date: 13 April 2015
End Date: 31 December 2016
Sub-contractors: Dual Lab, KEEP SOLUTIONS, Digital Preservation Coalition

1. Description of the release and progress compared to the last intermediate release

Please provide the PREFORMA consortium with a concise overview of the progress since the last intermediate release, and of the functionalities that are available at the time of this report.

Please highlight

- *The progress compared to the July 2016 release (intermediate release of the second prototyping phase)*
- *how you are addressing the comments received from the PREFORMA consortium*
- *what are your plans to progress further.*

Since version 0.18 in June most of the development effort has been spent replacing the PDF Box based PDF parser and validation model implementations. This is effectively redeveloping existing functionality so the changes to the released packages aren't as visible as new features. Version 0.26 of veraPDF will be released on November 9th, the release details will be found on GitHub: https://github.com/veraPDF/veraPDF-library/releases/tag/v0.26.1. A feature comparison with the 0.18 intermediate release is given below:

Feature	0.26	0.18
Validation Model Support	Greenfield & PDFBox PDF 1.4 and PDF 1.7	PDFBox PDF 1.4 and PDF 1.7
Validation Rules Support	Greenfield & PDFBox 1b, 1a, 2b, 2a, 2u, 3b, 3a. 3u	PDFBox 1b, 1a, 2b, 2a, 2u, 3b, 3a. 3u
PDF/A Flavours supported	Greenfield & PDFBox 1b, 1a, 2b, 2a, 2u, 3b, 3a. 3u	PDFBox 1b, 1a, 2b, 2a, 2u, 3b, 3a. 3u
Plug ins for PDF feature validation	Supports third party plug ins and reference plug in implementation.	Supports third party plug ins and reference plug in implementation.
Command line interface	Fully functional	Fully functional
Cross platform installer	Yes	Yes
Metadata fixing	CLI and GUI	CLI and GUI
REST interface	http://demo.verapdf.org/	http://demo.verapdf.org/
Reporting formats	XML (MMR report and raw formats), HTML, TEXT with templated reporting engine.	XML (MMR report and raw formats), HTML, TEXT
Policy Checking	Prototype Schematron policy engine	Prototype based on report analysis.
Batch Processing	Batch processing and dedicated batch reporting for	Limited batch processing and only single file at a time reporting.

	CLI, with ability to analyse PDFs in archive formats supported by Java's native compression library e.g. zip, tar.gz	

The full set of veraPDF release notes can be found on GitHub:
https://github.com/veraPDF/veraPDF-library/blob/integration/RELEASENOTES.md
Over the final 2 months of development we will:

- Carry out functional, performance and reliability testing of the greenfield PDF parser.
- Developing greenfield metadata fixing and feature reporting.
- Develop plugins that wrap open source validation / reporting tools for formats used within PDF/A, e.g. fonts, JPEG 2000, ICC colour profiles.
- Perform real world testing against 3rd party / institutional datasets, please note that many of these data sets cannot be shared due to license restrictions.
- Validate the test suite files via comparison with commercial products and discussions within the PDF Validation TWG.
- Produce the PREFORMA shell incorporating all 3 conformance checkers.
- Create platform specific installation packages.
- Finish development of the veraPDF REST services and web interface, specifically support for feature reporting, and metadata fixing and Greenfield implementation.

2. Datasets used to test the release

Please provide the PREFORMA consortium with a detailed description of the datasets that have been used to test the release (own, memory institutions, external, etc.), and the respective purpose of testing.

Both the PDF Box and Greenfield veraPDF releases and all integration branch merges are tested against:

- Our own synthetic test corpus for all PDF/A flavours: https://github.com/veraPDF/veraPDF-corpus
- The Isartor PDF/A-1b test suite: http://www.pdfa.org/2011/08/isartor-test-suite/
- The BFO PDF/A-2 test suite: https://github.com/bfosupport/pdfa-testsuite

The test results for each build are published here: http://tests.verapdf.org/. An example showing the results of the two implementations side by side is here: http://tests.verapdf.org/0.25.16/ These are all purpose produced data sets designed to test PDF/A validation functionality. The veraPDF test corpus comprises over 1,500 PDF files created by the consortium as a comprehensive PDF/A validation suite.

We targeted organisations that archive PDF/A files in accordance with institutional practice and those that archive PDF/A and other PDF files at scale. We approached 12 organisations and received feedback from seven ranging from national archives to university libraries. This diversity meant the software was tested in different working environments.

- British Library
- Koninklijke Bibliotheek (National Library of the Netherlands)
- Parliamentary Archives
- Leibniz-Informationszentrum Wirtschaft (German National Library of Economics)
- University of Yale
- University of Sheffield
- University of Edinburgh

The first testing round focussed on:

- experience of installing and running the tool (via the CLI, the GUI or both)
- response to how the tool works for you in your organisation
- how effective and useful the tool is for validating PDF/A and other PDF files
- how the tool performs against other conformance or validation tools that your organisation may have used
- any bugs or crashes encountered while using the tool.

A number of volunteers contributed issues to the veraPDF Github Issue tracker, in particular with regard to stability and performance improvements. This has led directly to the optimisations in memory usage and better handling of memory exceptions, enabling reliable batch processing.

3. Dissemination and community building

Please provide the PREFORMA consortium with the list of dissemination activities that you have undertaken to promote your open source project (webpages, blogs, newsletters, press releases, papers, presentations, etc.).

Please describe any potential long-term collaborations/partnerships entered into, by listing the organisation/s and the role they played in the project.

How did you progress in setting up an open source community around the developed tools?

Partnerships
KEEPS, Archivematica (as previous reports)

Web presence
Updates to the veraPDF website: http://verapdf.org/.
New documentation and getting started webpages: http://docs.verapdf.org/

Mailing list
New mailing list set up http://lists.verapdf.org/listinfo/users

Recent events/conferences

- 28 September - Presentation to the European Court of Auditors, Luxembourg
- 5 October - iPRES 2016 Workshop: Quality Standards for Preserving Digital Cultural Heritage, Bern
- 23 November - PREFORMA Experience Workshop, Berlin
- 25 November - PDF Day Australia, veraPDF presentation and demonstration, Sydney

Articles
'veraPDF: Building an open source, industry supported PDF/A validator for cultural heritage institutions' *Digital Library Perspectives Journal: Special issue on digital preservation tools and partnerships*

Webinars

- (8 September - DPF Manager webinar)
- (15 September - MediaConch webinar)
- 22 September - veraPDF webinar

Press releases

- 22 July - http://verapdf.org/2016/07/22/preforma-experience-workshop-improving-long-term-digital-preservation/
- 1 August - http://verapdf.org/2016/08/01/verapdf-0-20-released/
- 31 August - http://verapdf.org/2016/08/31/preforma-webinar-series-september-2016/
- 8 September - http://verapdf.org/2016/09/08/verapdf-0-22-released/
- 26 September -

http://verapdf.org/2016/09/26/recording-and-slides-from-verapdf-webinar-published/
- 27 September - http://verapdf.org/2016/09/27/join-the-verapdf-mailing-list/
- 12 October - http://verapdf.org/2016/10/12/verapdf-0-24-released/
- Sent to ~20 mailing lists and LinkedIn interest groups.
- Regular updates are posted for the industry community and public on pdfa.org and additional information for PDF Association members only at intranet.pdfa.org.

Twitter account https://twitter.com/_verapdf
- 97 followers

veraPDF news
- 160 subscribers

The PDF Association: PDF Validation TWG
- 57 subscribers

4. Open Source approach

Please provide the PREFORMA consortium with a description of how you addressed the relevant open source topics, best practices, and licensing issues identified in the report of the University of Skövde.

Since the last report we have developed the greenfield PDF parser and validation model that replace the Apache licensed PDFBox dependencies we've used until now. At the same time we've removed the Log4J code from our library. This means that release 0.26 will be the first source package that meets the PREFORMA licensing conditions.
We've tried to engage with and encourage external contributors to the project. The best example to date is an external pull request providing reporting capability to files using command line options: https://github.com/veraPDF/veraPDF-apps/pull/32.

We're re-designing the website and revising the content in response to review feedback.

Open Source Best Practises:

- source code, validation profiles, test corpus, and documentation on GitHub: https://github.com/veraPDF;
- Travis-CI for first stage of continuous integration: https://travis-ci.org/verapdf;
- Jenkins server for continuous deployment: http://jenkins.openpreservation.org/view/A-veraPDF/;
- Continuous integration testing against acceptance corpora: http://tests.verapdf.org/;
- continuous deployment of development and release installation packages: http://downloads.verapdf.org/dev/ and http://downloads.verapdf.org/rel/;
- Maven repository for all development and release source, javadoc and jar packages: http://artifactory.openpreservation.org/artifactory/vera-dev-local/;
- signed GitHub tags for all development and release versions: https://github.com/veraPDF/veraPDF-library/tags;
- Codecov for test coverage (moved from Sonar to allow community administration): https://codecov.io/gh/veraPDF/veraPDF-library; and
- Codacy for static code analysis (again moved from Sonar for community administration ease): https://www.codacy.com/app/veraPDF.

5. Standardisation efforts

Please provide the PREFORMA consortium with a description of how you are actively contributing to the standardisation process in your domain, by means of providing feedback on existing standards as well as supporting emerging standards.

The next scheduled meeting of the ISO committee for PDF/A (ISO TC 171 SC 2 WG 5) will be in Sydney, Australia in December, 2016. As in the Ghent meeting in May, 2016, a number of PDF Validation TWG members are expected to attend the Sydney meetings.

Since July 2016 the PDF Validation TWG has made progress towards finalizing the list of ambiguities in PDF/A and detailing the resolution thereof. Discussion continues on a few items already raised, with a few new additions. The complete list (known internally as the "Resolution of Ambiguities" document) will be provided to the ISO committee for review in Sydney in the form of a ready-for-publication PDF Association Technical Note. If approved by the ISO WG, the PDF Association will proceed to publish this Technical Note in early 2017.

The purpose of PDF Association Technical Notes

The ISO WG has already determined and confirmed that the existing ISO specifications for archival PDF (PDF/A-1, PDF/A-2 and PDF/A-3) will not be revised to address the ambiguities identified and resolved by the PDF Validation TWG and ISO WG through the operation of the veraPDF project.

With respect to the ISO standard itself, the TWG's input will be incorporated into the forthcoming 4th part of PDF/A, currently under development, but this leaves the question of how to address the ambiguities identified in previous parts of ISO 19005.

In order to promote industry awareness and acceptance of this work, the PDF Validation TWG will request that the ISO WG issue a formal resolution calling on the PDF Association to publish a Technical Note addressing these ambiguities.

Technical Notes published by the PDF Association and its PDF/A Competence Center have a good track-record of adoption by the industry. Within a year of publication of the six PDF/A Technical Notes published to-date, all major PDF/A vendors had addressed them in their own implementations.

See the existing Technical Notes on pdfa.org:

https://www.pdfa.org/publication/pdfa-1-technical-notes/

In the meantime, members of the ISO WG, including veraPDF consortium staff, have reviewed the next Committee Draft (CD) of PDF/A-next, and have prepared comments for review at the December, 2016 ISO meeting.

Summary of veraPDF consortium standards-development activities (updated)

- We have established that existing Parts of PDF/A will not be amended in any way. Any clarifications to existing ambiguities will be addressed in the forthcoming new Part for PDF/A, presently termed "PDF/A-next".
- We have driven awareness of the need for PDF/A-next, and led in its development.
- We plan to submit a request for an ISO WG resolution calling on the PDF Association to publish its Resolution of Ambiguities document as a PDF Association Technical Note.

6. Impact assessment, sustainability, future use and exploitation

Please provide the PREFORMA consortium with a description of your ideas and plans related to the sustainability, future use and exploitation of the results of your project.

Please include some evidence of the impact that the project has generated to date for the memory institutions and for any other relevant target groups.

The case for maintaining and sustaining veraPDF

The feedback we have received from users, including at the latest iPRES event, encourages us to believe that veraPDF is meeting a deeply-felt need on the part of memory institutions, and commercial organizations concerned with the long-term viability of their records.

PDF/A, however, represents a small fraction of the files such organizations process; the vast majority are simply PDF files, and do not claim to conform to PDF/A. Further, PDF/A, however, is only a "use" of ISO 32000, the PDF specification itself. What memory institutions really need is the ability to validate the conformance of PDF itself.

Conformance-checking for the entire PDF format is a massive project, requiring many man-years of effort. In addition, there are other PDF subset specifications of interest to memory institutions, including: PDF/A-next, PDF/E (engineering), PDF/UA (universal accessibility), PRC and more.

As designed, veraPDF may be readily extended to cover all aspects of PDF, PDF subset standards, related specifications, and even third-party standards such as XMP or PRC.

Sustaining veraPDF

Maintaining and developing the software in a professional fashion and maintaining the attention of the industry, requires financial resources and organisational commitment. The veraPDF consortium is planning to develop a revenue-generation system based on the veraPDF software and veraPDF.org to achieve these resources.

The veraPDF Project

In order to fund software maintenance and future development, the veraPDF consortium plans to:

- Create a mechanism to facilitate the aggregation of anonymized test data from veraPDF users
- Create a mechanism to generate conformance reports from the universe of files tested
- Provide memory institutions and commercial (industry) organizations with access to the conformance reports based on an annual subscription
- Provide a means of demonstrating support for the veraPDF project via a "sponsors" page, or similar.

Grants

In addition to activities intended to generate revenue directly, the veraPDF consortium has approached 3rd parties to request grant monies to continue development of the software.

Consulting

The veraPDF consortium have been approached by memory institutions with requests for commercial applications based on veraPDF.

7. Gap analysis and next steps

Please provide the PREFORMA consortium with a description of the status of the work compared to what was planned in the functional and technical specification that you provided at the end of the design phase.

Please highlight critically what it is still missing in the current release and what are your plans to overcome the gaps.

Please include also an updated version of your work plan and a timeline, preferably in a graphical way (GANTT) so that the PREFOMA consortium members now and later can easily compare the status of fulfilling the requirements of the project as well as the level of compliance to your own technical and functional description.

Feature	Status	Schedule
Generic validation model	Complete	As planned
PDF/A validation profiles	Complete	Finished early
Machine-readable reports	Complete	As planned
Metadata fixer	Complete	As planned
Policy profiles	We now have some concrete policy cases from the test corpus meetings. We'll release a customisable policy engine capable of meeting these uses cases.	November 2016
Internationalization	There's been no call for translated versions and the PDF/A specifications are English. We'll assess the requirement for internationalisation during the testing phase.	Possibly 2017
Report templates	PDF reports are still in development and behind schedule.	November 2016
Interfaces	Second prototypes are implemented.	As planned
Third-party plug-in	Complete	As planned
Greenfield parser	Testing	As planned

Jpylyzer plug in development	Not complete, we may be able to implement in 2016, but it might be later.	Unsure
PREFORMA shell	Not complete, integration has taken a back seat until we considered our validator functionally complete.	Full prototype for PREFORMA experience workshop in Berlin Nov 2016.