PROTOTYPING PHASE 2 INTERMEDIATE REPORT

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

veraPDF

Revision: final

Authors:

Joachim Jung Open Preservation Foundation
Becky McGuinness Open Preservation Foundation
Carl Wilson Open Preservation Foundation
Duff Johnson PDF Association
Boris Dubrov Dual Labs
……

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, if possible even more organised release compared with the respective predecessor versions
- A report which
    - Describes
        - More in detail the respective release;
        - The timeline 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 which 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 - IMMEDIATE 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 – Intermediate Report Total Contract Price [euro]: 669, 525 Start Date: 13 April 2015 End Date: 31 December 2016 Partners: PDF Association, Digital Preservation Coalition Sub-contractors: Dual Lab, KEEP Solutions

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

Please provide the PREFORMA consortium with a concise overview of the releases developed so far, and of the functionalities that are available at the time of this report. Please highlight

- *which is the progress compared to the October 2015 release (final release of the first prototyping phase)*
- *how are you addressing the comments received from the PREFORMA consortium*
- *which are your plans how to progress further.*

Feel free to refer to any other document you provided so far, when appropriate, by providing the link.

Version 0.18 of veraPDF was released at the end of June, the release details can be found on GitHub: https://github.com/veraPDF/veraPDF-library/releases/tag/v0.18.1. A feature comparison with the 0.6 intermediate release is given below:

Feature	0.18	0.6
Validation Model Support	PDF 1.4 and PDF 1.7	PDF 1.4
Validation Rules Support	1b, 1a, 2b, 2a, 2u, 3b, 3a. 3u	1b
PDF/A Flavours supported	1b, 1a, 2b, 2a, 2u, 3b, 3a. 3u	1b
Plug ins for PDF feature validation	Supports third party plug ins and reference plug in implementation.	Prototype
Command line interface	Fully functional	Prototype
Cross platform installer	Yes	Yes
Metadata fixing	CLI and GUI	GUI prototype
REST interface	http://demo.verapdf.org/	No
Reporting formats	XML (MMR report and raw formats), HTML, TEXT	XML (MMR format), HTML
Policy Checking	Prototype based on report analysis.	None

The full set of veraPDF release notes can be found on GitHub: https://github.com/veraPDF/veraPDF-library/blob/integration/RELEASENOTES.md

The PREFORMA consortium’s main concern regarding the first intermediate release was the lack of a command line interface. This was released in October 15th and has undergone significant development since. It is now feature complete.

Over the final 6 months of development we will be to:

- Complete development of our greenfield PDF parser implementation, this will replace the current PDF Box dependency.
- Carry out functional, performance and reliability testing of the greenfield PDF parser.
- Remove other Apache licensed code in line with PREFORMA's licensing requirement, e.g. our logging is Apache based / PDF Box compatible.
- 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 better separation of services and user interface, support for feature reporting, and metadata fixing.

2. Testing
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. The 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/. 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've also performed institutional testing on real world data sets that can't be provided because of IPR issues. Testing with heritage sector organisations focuses on reliability, performance and usability rather than testing the validators functionality against the PDF/A specifications. The British Library's large scale testing meant that we fixed serious performance issues with large files containing high quality images and problems with text layer fonts.

3. Dissemination and community building

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

Recent events/conferences

- 7 April - PREFORMA Open Source Workshop, Stockholm
- 1-2 June - OPF AGM, veraPDF overview and progress report, The Hague
- 14-15 June - PDF Days Europe, veraPDF presentation and demonstration, Berlin
- 13-16 June - Open Repositories, veraPDF presentation and demonstration, Dublin
- 23 June - Re:Format - What is file format obsolescence and does it really exist?, York

Webinar

- 14 June - Pre Commercial Procurement for the long-term Preservation of Digital Cultural Heritage

Press releases

- 1 April - http://verapdf.org/2016/04/01/verapdf-0-12-released-alongside-first-version-of-wiki-validation-rules/
- 5 May - http://verapdf.org/2016/05/05/verapdf-0-14-released-with-launch-of-demo-website/
- 3 June - http://verapdf.org/2016/06/03/verapdf-0-16-released-with-full-support-for-all-pdfa-parts-and-conformance-levels/
- 8 July - http://verapdf.org/2016/07/08/verapdf-0-18-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

- 71 followers

veraPDF news

- 144 subscribers

The PDF Association: PDF Validation TWG

- 55 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
How did you progress in setting up an open source community around the developed tools?

For each veraPDF software release we have now uploaded the following zip archives to the PREFORMA open source portal:

- one containing the cross platform installer and start-up shell files for Windows and bash;
- another containing the full source of veraPDF and its dependencies with the exception of log4j (see below); and
- three archives containing the build tools needed to compile the software, Java JDK and Maven, with example scripts to install the tools and compile the code.

The build environment comes in 3 forms, one for linux, another for MacOS and a final one for Windows.

There is an issue with the source archive concerning building log4j from source. This was built using Java 1.4 and does not compile using Java 1.6 or more recent versions. veraPDF requires Java 1.7 or more recent making it impossible to build the veraPDF code and log4j using the same JDK. Our temporary solution has been to package three pre-compiled jars for the log4j modules that won't build. The permanent solution will be to remove the log4j dependency, inherited from PDFBox, when we merge the greenfield parser development later this year.

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: http://staging.verapdf.org/. At the same time we've introduced a dedicated documentation site: http://docs.verapdf.org/ that's GitHub pages based: https://github.com/veraPDF/veraPDF.github.io. This will mean that external contributions to the documentation will be made via GitHub pull requests, in the same manner as contributions to the source code.

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 the existing standards contributing as well as the way on how to support emerging standards.

Since our last report in March, the ISO committee for PDF/A (ISO TC 171 SC 2 WG 5) met in Ghent, Belgium in May 2016. As in the previous meeting, a number of PDF Validation TWG members attended the ISO meetings in Ghent.

Continuing on our success in Basel in November, 2015, the PDF Validation TWG submitted another ten points of ambiguity in PDF/A to the ISO WG for consideration and resolution. As before, the established commenting mechanism was used to address questions of interpretation and provide recommendations for future PDF/A specifications. These included:

- The interpretation of the corrigendum 2 to ISO 19005-1, which contains a special clause to exclude resources unreferenced from the corresponding content stream from further requirements.
- The Charset and CIDSet entries remain a sore-spot. The keys in question are deprecated from ISO 32000-2, and thus do not affect PDF/A-next. However, the requirement remains for PDF/A-2 and PDF/A-3. It will be left to an industry Application Note to provide a universal reference for relaxing these unnecessary and problematic requirements.
- Clarification on whether XMP metadata streams in PDF/A-1 must be uncompressed. The TWG's interpretation was accepted, and the ISO WG added an additional clarification: that XMP packages don't need to conform to XMP or even XML.
- An ambiguity over whether the requirement pertains to the file-format or to a means of comparing real values. The ISO WG decided that non-zero values less than the minimal one are not allowed in PDF/A-2 (and PDFA-3) on purpose.
- Clause 6.1.13 in ISO 19005-2 copies the list of limits from ISO 32000-1 and lists them explicitly. However, the word "approximately" was dropped, and so the definition of the limits differs between ISO 32000-1 and PDF/A-2, creating an untenable situation for processors encountering files that may exceed these limits. The ISO WG elected to leave the matter as-is because although differing from the base specification for PDF the actual requirement for PDF/A-2 was itself not ambiguous.
- Regarding the "shall" requirement in all three parts of PDF/A to comply with either predefined schemas from the XMP specifications or with an extension schema, the ISO WG accepted the PDF Validation TWG's recommendation for PDF/A-next.
- Regarding the value and practicality of the requirement in PDF/A-2 and PDF/A-3 to record user actions in the xmpMM:History property, the ISO WG accepted the PDF Validation TWG's recommendation for PDF/A-next but highlighted that the parameters field is still required in xmpMM:History for conformance with PDF/A-2 and PDF/A-3.

- In PDF/A-1 it’s not clear if any Widget annotation is required to have an annotation dictionary. The WG agreed with the TWG’s interpretation that for PDF/A-1, every button field widget shall have an appearance stream or dictionary.
- The requirement for multiple appearance streams, which applies to all current parts of PDF/A, misses the case when a form (such as a radio button) has multiple widgets associated to it and defined in /Kids array. The TWG proposed to PASS otherwise valid PDF/A documents containing a Widget annotation dictionary with Parent key referring to a parent form field of type Button, and if the value of the N key in this widget annotation dictionary refers to an appearance subdictionary. The ISO WG agreed.
- Some wording pertaining to ICC color spaces is imprecise; the TWG proposed specific replacement text. The ISO WG accepted this interpretation, and the PDF/A-next Project Leader agreed to make this change in the text of PDF/A-next.

The final dispositioned set of comments to the ISO WG was not available at the time this report was due, but the above list includes all items discussed in that set.

Since the Ghent meeting members of the ISO WG, including veraPDF consortium staff, have been awaiting the next Committee Draft (CD) of PDF/A-next from the ISO Project Leader. This document was delivered in mid July. veraPDF consortium members will study it in order to prepare comments for the November, 2016 ISO meetings in Sydney, Australia.

In addition to comments on PDF/A-next, the PDF Validation TWG will seek to generate any remaining ambiguities for discussion and resolution in Sydney.

Updated summary of veraPDF consortium standards-development activities

- 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.
- Following release of the post-Ghent PDF/A-next Committee Draft (CD) we plan to submit additional items for clarification in PDF/A-next for consideration at the WG 5 meeting in Sydney in November, 2016.

We anticipate our collaboration with WG 5 and WG 8 to persist throughout the course of 2016 and into 2017, as we continue to resolve remaining matters of interpretation regarding the specification.

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 generated so far for the memory institutions and for any other relevant target group.

The active use of a tool such as a format validator depends strongly on its stability and feature completeness. For this reason, we have not seen an installation into a working environment yet. However, during our dissemination efforts (see section 3) we have seen strong interest in veraPDF from both the digital preservation and the document industry side. A number of archiving system vendors have also expressed an interest in including veraPDF in their offering. A first approach from outside these communities has come from the Dutch office for standardisation, who have asked for meetings with both the PREFORMA and the veraPDF consortium. This included a request for help with the integration of veraPDF into a public website, which should verify the formats of content of Dutch governmental (and potentially third-party) websites. Negotiations about this are ongoing.

Part of OPF’s mission is the sustainability of project results in the digital preservation field. Since the end of the Planets project, we have sustained and made available the results of a whole set of projects. Over this time, we have created the tools and processes needed to keep software and knowledge collected in research projects available for the long-term. These tools are already in use for the project and will sustain veraPDF results after the project runtime. However, we also have a model to extend such results, which we call stewardship. Currently, this model is applied to JHOVE and we will extend it to veraPDF after the project officially ends.

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 which 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) in a way that the PREFORMA 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.

Comparison with the planned delivery schedule for the milestone M2.4d (15/07/2016) as defined at the end of the design phase:

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	Prototype implementation, not tested against real institutional policy as was planned. This is mainly due to the lack of real requirements from institutions. Based on feedback we’re working on the concept of risk assessment of non-valid PDF/As also.	November 2016
Internationalization	We’ve not worked on this as all of the veraPDF documentation and error reporting is tied to the PDF/A standards which are only in English. We will assess the need for full internationalisation support by October 2016.	By December 2016 if necessary.
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	In development / testing.	As planned
Jpylyzer plug in development	Not complete	September 2016
PREFORMA shell	Not complete, integration has taken a back seat until we considered our validator functionally complete.	Full prototype for iPres workshop start October 2016.