END OF PHASE 1 REPORT

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

veraPDF

Revision: final

Authors:

Ed Fay (Open Preservation Foundation)
Duff Johnson (PDF Association)

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

1 INTRODUCTION

The purpose of the end of phase report is to ensure that contractors have performed the procured R&D services as specified in the framework agreement. Please describe the work undertaken during phase1, including what work was completed and why this was important. Please complete this form as fully as possible.

This report must be submitted within 14 days of the completion/ termination of the phase. You are advised that satisfactory completion of this report forms part of the contract.

Reports should be submitted by email to the following email addresses:

- Peter Pharow: phw@idmt.fraunhofer.de
- Claudio Prandoni: prandoni@promoter.it

The objectives of reporting:

- To create an understanding of the work undertaken and its success in meeting the projects agreed objectives.
- Also, to provide the company with a comprehensive report to share with stakeholders and those that may help further commercialisation.

The report should be completed by the contractor, with input from any sub-contractors or project partners as appropriate. Please answer, wherever possible, on behalf of the business units, divisions, companies or other legal entities involved in the work. If this is not possible, please specify the organisation to which your answers refer.

Please answer the questions in the spaces provided. Try to answer fully, but keep your answers succinct and no longer than necessary to clearly explain them. When describing technical solutions, please regard your audience as being someone familiar with the technology, but not an expert. The report may be done in narrative alone. However, diagrams or pictures may be added where these aid clarity within the restriction on the page limit of a total of 20 sides of A4.

Because the true impact of an R&D project often takes several years to emerge, we may approach you for up to six years after project completion to follow up on the questions in this report. Your co-operation with any such follow up work is greatly valued.

2 END OF PHASE 1 REPORT

1. Details
Type of Organisation: Non-profit foundation
Registered Name of Organisation: Open Preservation Foundation
Registered Address: c/o The British Library
Town/ City: Boston Spa, Wetherby
Postcode: LS23 7BQ
County: West Yorkshire
Country: United Kingdom
Report Author: Ed Fay
Telephone Number: +44 (0) 1937 546 013
E-mail Address: ed@openpreservation.org
Project Name: veraPDF
Report Type: End of Phase 1 Report
Total Contract Price [euro]: 56,350 (excl. VAT)
Start Date: 01 November 2014
End Date: 28 February 2015
Sub-contractors: PDF Association, Dual Lab

2. At the outset of this piece of work, what were your aims and objectives?

Please provide a concise overview of the supplier's project objectives and of what was expected during the first design phase as agreed in the PREFORMA Tender Form and in the Negotiation Protocol with the supplier.

[limit: 1 page]

To produce:

- Functional and technical specifications for a definitive PDF/A Validator (including an analysis of development options based on the specified licencing model);
- Community engagement and communications plan;
- Test cases derived from PDF/A specifications and an analysis of existing test corpora;
- Demonstrator based on PDFBox.

3. Please provide a summary off the outputs of this piece of work and relate these to the original objectives. How do the outputs address the challenge of this PCP?

Please provide the PREFORMA consortium with a concise overview of the progress of the work expected to be done in the first design phase, relating it to the original objectives and to the requirements defined in the PREFORMA Challenge Brief.

[limit: 1 page]

- Functional and technical specifications for a definitive PDF/A Validator

  The functional and technical specifications delivered in our final report contain all design elements required for developing PDF/A Validation software as well as the other components identified by PREFORMA. We have provided detailed descriptions of: all Conformance Checker components, extensions, interfaces, and integrations; technical architecture (including domain model and API); validation model and profile format; machine-readable report format; policy profile format; report template format; test framework; internationalization; and integration mechanism for third-party tools.

  The veraPDF consortium also provided annexes containing: the communications plan; technical milestones for phase 2; analysis of existing test corpora; PDFBox feasibility study; license compatibility report; links to the proof-of-concept demonstrator; and further analysis of ICC profile and font validation requirements.

  These deliverables meet the requirements identified in the PREFORMA challenge.

- Community engagement and communications plan

  The veraPDF consortium's community engagement report includes an analysis of stakeholders in industry, memory institutions, and other organisations. We provide an update on community building progress in phase 1, anticipated community engagement in phase 2, and contribution guidelines for requirements, corpora, and code. The communications plan details specific community engagement objectives and the channels available to the veraPDF consortium.

  These activities demonstrate progress towards leveraging existing communities, including members of the relevant ISO standards committees. They also create the basis for the larger veraPDF community to be developed in phases 2 and 3.

- Test cases derived from PDF/A specifications and an analysis of existing test corpora

  Detailed analysis of the PDF/A specifications has identified and catalogued 800+ test cases and mapped these against the coverage of existing corpora. This set, along with other edge-cases as may be identified during phase 2, determines the objective frame of reference for PDF/A Validation and forms the basis for the development of authoritative validation software.

- Proof-of-concept demonstrator based on PDFBox

  The demonstrator is available at http://demo.verapdf.org

4. Describe any changes to the original plan in the tender. What was the reason for these changes? Please include any circumstances that aided or impeded the progress of the project and the actions taken to overcome them.

If applicable, explain the reasons for deviations or clarifications from what was agreed in the Tender Form and in the supplier's Negotiation Protocol, or for failing to have achieved critical objectives and the impact on the supplier's project. If applicable, propose corrective actions that will, in case the supplier is invited to the prototype phase, take place either in the prototype phase as such, or in the re-design phase, if applicable.

[limit: 1 page]

We have not introduced any changes to our original plan as agreed following negotiation.

In light of the PREFORMA licensing requirements, as clarified in subsequent correspondence, we will now pursue the greenfield development option outlined in our original proposal of developing a PDF Parser ourselves. This is discussed in detail in the Functional Specification (section 3.1) and Licensing Compatibility Report (Annex E).

We propose the use of the PDFBox PDF Parser (but no other functionality – we specifically exclude the PDFBox preflight engine from the veraPDF Implementation Checker) during the first half of Phase 2 to allow development testing of the Implementation Checker to begin as early as possible. The greenfield solution will be made available by the end of Phase 2.

5. Please provide a short factual summary of the most significant outcomes of your work.

Please provide the PREFORMA consortium with a concise overview of the main results achieved so far. Please refer to the functional and technical specification without repeating too much here.

[limit: 1 page]

The main result of phase 1 is the production of the functional and technical specifications and associated reports as described above, along with initial community building activities amongst the stakeholders represented by the veraPDF consortium partners:

- The PDF Association has established a dedicated PDF Validation Technical Working Group (TWG) comprising most active members of the ISO 19005 and ISO 32000 working groups. The TWG has actively reviewed the functional and technical specifications and has provided substantial input (200+ mailing-list messages to-date) into the corpora analysis. The TWG is preparing input, including clarification questions, that will be submitted for consideration by the responsible ISO working groups at the upcoming (April, 2015) ISO TC 171 SC 2 meetings in San Jose, California.

  See Community Engagement, section 2.2.1 *Industry and Standards* for more detail.

- The Open Preservation Foundation has presented the functional and technical specifications to its members who have actively reviewed and provided feedback on the designs and supplied detailed examples of policy requirements (so far 8 memory institutions have participated directly).

  See Community Engagement, section 2.2.3 *Memory Institutions* for more detail.

6. Describe the innovative aspects of the work, including any new findings or techniques.

[limit: 1 page]

We have established a unique collaboration of industry, ISO committee members, memory institutions, non-profit open-source expertise, and PDF domain expertise to develop designs for a “definitive” PDF/A Validator.

Our approach has already established an innovative collaboration with the potential to influence the market for PDF/A software, while providing immediate benefits to memory institutions through the development of new tools for assuring the quality of cultural heritage collections.

Over time, the veraPDF model has the potential to dramatically reduce the costs associated with ingesting, quality-controlling, and managing PDF documents through normalization of the preservation-ready capabilities of PDF document creating and editing suites worldwide.

7. Describe where the R&D and other operational activities have been performed.

[limit: 1 page]

All Phase 1 activities were carried out at the offices of the consortium partners, which are companies located, based, and legally incorporated in Belgium (Dual Lab), Germany (PDF Association), Portugal (KEEPS) and the United Kingdom (OPF and DPC).

Consistent with the guidance received from Per Elfner on 2014-07-14 by email in response to our question, some work on Phase 1 was performed in the United States and Belarus by contractors and employees of the consortium partners.

Regarding the requirement received from Claudio Prandoni on 2015-03-11 by email that "the majority of the R&D and operational activities related to the PCP contract, including in particular the principal researcher(s) working for the PCP contract, must be performed in EU Member States or FP7 associated countries":

- all R&D work carried out by the principle researchers at each of the partners occurred within the EU;
- overall, 69% of R&D effort (by cost) occurred within the EU;
- overall, 60% of all costs occurred within the EU.

8. Please provide complete and clear information about the allocation of monies paid by the Authority with consideration to the R&D service contract minimum requirement (that more than 50% of the contract value is attributable directly and exclusively to legitimate R&D services).

[limit: 1 page]

The impact of the veraPDF consortium's R&D activities include:

- Increasing knowledge about objectivity in the validation of file formats and processes to establish conformance (the validation model);
- Establishing the requirements of PDF/A validation specifically (the test cases and analysis of existing corpora);
- Producing designs for new applications of this knowledge in the production of software (the functional and technical specifications).

The breakdown of costs between R&D and other activities in Phase 1 was as follows. Note that veraPDF consortium partners KEEPS and DPC were not resourced in Phase 1, and thus do not appear in the table.

	OPF	**PDFA**	**DL**	**Total**
R&D	€15,750	€2,640	€15,660	**60%**
Communications	€2,325	€7,920	€1,740	**22%**
Management	€3,675	€2,640	[included in R&D contract]	**11%**
Materials and travel	€3,340	€434	€227	**7%**

9. Describe any potential long-term collaborations/ partnerships entered into. Please list the organisation/s and the role they played in the project.

[limit: 1 page]

The Open Preservation Foundation and PDF Association have entered into an alliance through the signing of a Memorandum of Understanding that appoints the organisations as allied or partner organisations respectively. This recognises our mutually beneficial work and identifies opportunities for collaboration, both during PREFORMA and beyond.

In the context of the veraPDF consortium, the Open Preservation Foundation brings memory institutions and expertise in open-source software development while the PDF Association brings PDF software developers, vendors, and a formal liaison to ISO working groups responsible for PDF and PDF/A.

10. Please describe how your organisation has gained from this project. What new business opportunities have been created? Do you expect your organisation to grow as a result of this project?

[limit: 1 page]

Little growth amongst the partners and subcontractors was expected to result from phase 1 alone however the potential described for phase 2 is already manifesting as illustrated below.

Dual Lab have already developed leads and a service contract based on their design work during phase 1 through the PDF Validation TWG. This is indicative of market interest as well as collaborative opportunity, as described in our original proposal.

Numerous PDF Association members have expressed their strong support for this project. The PDF Association has added several new members since the beginning of phase 1; several of these have cited the veraPDF project as a leading rationale for their involvement.

11. Describe the potential for exploiting the work. Please identify any new intellectual property which has been filed or for which filing is anticipated.

In this sub-section, the supplier is expected to describe possible business models, business plans, and business cases based on use cases or scenarios relevant for planning ahead. The business plan should not only cover the PREFORMA phases to come but may also give an indication on how exploitation could look like after the end of the PREFORMA project.

[limit: 1 page]

As discussed elsewhere, the veraPDF consortium's project is directed at the PDF (and by extension, the electronic document) industry as a whole rather than individual businesses or consultancies. As such, we anticipate that veraPDF will be implemented as a component in the product development and marketing strategies of a wide range of vendors. Preliminary indications garnered during Phase 1 support this thesis; the PDF Validation TWG has already attracted the active involvement of industry leaders such as Adobe Systems, callas software, and iText, among others.

veraPDF consortium partner KEEP SOLUTIONS plans to leverage the veraPDF software in the development and delivery of their open source-based solutions and consulting efforts directed at memory institutions.

Please see the original veraPDF tender proposal, section III *Commercial Feasibility and Route to Market* for a longer discussion of the possible business models, plans, and cases.

12. Describe the suitability of the project results for: (a) developing a prototype, and (b) development of test series – in order to facilitate assessments of progress into next phase.

In this sub-section, the supplier is expected to shortly introduce the thought and plans of the supplier consortium on how to proceed with the development work in the next phase. That may include ideas and plans, e.g., for the meta data to address in the prototyping phase, the common platform to demonstrate interoperability between the modules developed inside and outside the PREFORMA project, and other aspects the supplier considers relevant for the prototyping phase.

[limit: 1 page]

The results so far are eminently suitable for both objectives.

(a) developing a prototype

The software design already submitted is fully considered and ready for development work at the initiation of phase 2. We have covered use cases, functional design, and technical architecture in detail, and have staff available within each of the partner organisations to begin work in April 2015.

The metadata to report are described in detail in the Technical Specification, section 4 *Machine-readable Report format* and interoperability is discussed in the Functional Specification, section 2 *Conformance Checker components* and section 3 *Conformance Checker extensions* and the Technical Specification, section 9 *Integration with third-party tools*.

(b) development of a test series

The PDF/A specification analysis and the 800+ test cases already identified suffice as the basis for establishing definitive test corpora covering all parts of PDF/A. The analysis of existing corpora, and the assurance of their creators of support and relicensing of existing files where appropriate, provides a starting point of existing test files which will be added to with new files until the corpora both represent every requirement of PDF/A and instantiate them authoritatively.

The corpora will be openly available for the evaluation of existing software against this objective frame of reference as well as for testing the veraPDF Conformance Checker.

Corpora are discussed in detail in the Community Engagement report, section 3.2 *Corpora*, the Technical Specification, section 6.2 *Test corpora*, Annex C: *PDF/A Test corpus analysis*, and Supplement J: *PDF/A Test corpus analysis.*

13. Open Source approach

In this sub-section, the supplier are asked to describe how they will address the relevant open source topics, the open source licensing, the way to address the open source communities, and the ideas in this respect for the project phases to come.

[limit: 1 page]

We discuss our approach to open-source licensing in detail in Annex E: *Licensing compatibility report*, as also discussed above in the answer to question 4. In summary, we will adhere to all requirements of PREFORMA, and our designs ensure that the core, innovative functionality of the Conformance Checker is entirely free of external dependencies. We propose the reuse of existing software to provide generic functionality within the Policy Checker, Reporter, and Shell to provide PREFORMA with the most efficient and effective means of building a fully-featured Conformance Checker for PDF/A.

We discuss our approach to open-source work practices in the original veraPDF tender proposal, section IV *Cohesion with open source development values and objectives* and complement this with a detailed discussion of our community-building activities in the Community Engagement report submitted at the end of phase 1. In summary, we are engaging a broad community in the development of the conformance checker, and have formal mechanisms and experience in establishing long-term community sustainability. Where appropriate, we will also pursue collaborations with existing communities, such as those already working on PDF technology as well as in associated areas such as fonts and colour profiles.

14. Standardisation efforts

In this sub-section, the supplier shall, if applicable, describe how the supplier's project aims at contributing to the exploitation of existing standards relevant to the project aims and goals, or how the supplier consortium has thought about contributing to emerging standards. Maybe the supplier can describe how the consortium is going to address future changes on the existing standards taking into account that the near future will bring new archival standards.

[limit: 1 page]

The PDF Association's PDF Validation Technical Working Group, more fully described in the Community Engagement section of our Phase 1 Report, is already populated with most of the active members of the working groups responsible for ISO 32000 and ISO 19005. The people participating in the TWG are the same people who are working on ISO 32000-2 and considering whether, how, and when to update the PDF/A standard.

The TWG has the right, as a category A liaison to ISO TC 171 SC 2, to create formal comments against draft standards documents. Accordingly, the TWG plans to present at least two questions and proposed clarifications to the normative language to WG 5 regarding interpretation of PDF/A-2 at the next ISO TC 171 SC 2 meeting, this coming April, in San Jose, California.

See the original veraPDF tender proposal, section II *Potential of the Proposed Idea/ Solution/ Technology to Address Future and/ or Wider Challenges in the Area* and Community Engagement, section 2.2.1 *Industry and Standards* for more detail.

15. Provision of data.
In this sub-section, the PREFORMA consortium asks the supplier to provide an overview on how the supplier consortium will work out the different sets of data needed to develop the respective module. This mainly considers the training data to be used internally but also the test data used by the PREFORMA consortium to test the modules and achievements of each of the suppliers working on the same file type. Eventually, demonstration data is needed to allow companies and organizations outside the PREFORMA consortium to spend their effort on developing their own modules but compare them with the PREFORMA modulkes by means of using the same correct and corrupt demonstration data sets. *[limit: 1 page]* As described in detail in our plans for developing comprehensive test corpora which authoritatively instantiate the requirement of the PDF/A specifications we will build training data to lead the development of the conformance checker. The test corpora will be released as demonstration data under open licenses to encourage transparent review of their resolution of ambiguities in existing software and to provide the basis for parallel development of competitive solutions by 3rd party companies and organisations. See Community Engagement, section 3.2 *Corpora* and Technical Specification, section 6.2 *Corpora* for more detail.

16. Please insert additional information that may be pertinent. This may be in the form of text, pictures, diagrams, data, graphs that support the work.
[limit: 1 page] We include various supporting information in the phase 1 specifications, particularly: - test case analysis of PDF/A specifications and existing corpora (Annex C, and Supplement J – also available as a spreadsheet on request); - the full API model (available via http://veraPDF.org/).

17. Describe what ethical aspects you have identified and how this may influence your solution.
[limit: 1 page] We have identified the need for all interfaces to meet accessibility requirements and best practices. The interfaces and reports generated by veraPDF will conform to WCAG 2.0 Level AA.

3 FINANCIAL REPORT

	Unit price	Quantity	Quoted price (€)	Total Price (€)
Labour Price				
1.Project manager (OPF)	49	75	3675	3675
2.Technical architect (OPF)	42	375	15750	15750
3.Community manager (OPF)	31	75	2325	2325
4.Industry Coordinator (PDFA)	60	220	13200	13200
5.Technical architect (DL)	25	300	7500	7500
6.Senior developer (DL)	24	300	7200	7200
7. QA engineer (DL)	18	150	2700	2700
Materials	[varies]	[as needed]	1000	1000
Capital Equipment	n/a	n/a	0	0
Sub Contract	n/a	n/a	0	0
Travel and accommodation	1500	2	3000	3000
Other (specify)				
TOTAL PRICE (excluding VAT)			56350	56350
TOTAL PRICE (including VAT)*			67620	67620