The PREFORMA Project: Federating Memory Institutions for Better Compliance of Preservation Formats

Linda Cappellato[1], Nicola Ferro[1], Antonella Fresa[2], Magnus Geber[3], Börje Justrell[3], Bert Lemmens[4], Claudio Prandoni[2], and Gianmaria Silvello[1]

[1] Dept. of Information Engineering, University of Padua, Italy
{cappellato,ferro,silvello}@dei.unipd.it
[2] Promoter, Italy
{fresa,prandoni}@promoter.it
[3] National Archives of Sweden, Sweden
{borje.justrell,magnus.geber}@riksarkivet.se
[4] Packed, Centre of Expertise in Digital Heritage, Belgium
bert@packed.be

Abstract. In this paper, we describe the motivations, objectives and organization of the *PREservation FORMAts for culture information/e-archives (PREFORMA)* project, a *Pre-Commercial Procurement (PCP)* project focused on conformity check of ingested files for the long-term preservation.

1 Introduction

Memory institutions, in Europe and elsewhere, are facing a situation when transfers of electronic documents or other electronic media content for long term preservation are continuously increasing. Data are normally stored in specific file formats for documents, images, sound, video etc. that are produced by software from different providers. This software is controlled neither by the institution that produces the files, nor by the institution that holds the archive. There is a risk that data objects meant for preservation, passing through an uncontrolled generative process, can jeopardise the whole preservation exercise.

PREservation FORMAts for culture information/e-archives (PREFORMA)[5] is a *Pre-Commercial Procurement (PCP)* project (2014-2017), co-funded by the European Commission under its FP7-ICT Programme. Its main objective is to give memory institutions full control of the process for testing the conformity of files to be ingested into their archives. This shall be obtained by developing a set of tools which will enable the testing process to happen within an iteration that is under full control of the institutions.

The paper is organized as follows: Section 2 provides an overview of the PCP instrument; Section 3 describes the overall approach adopted by the PREFORMA project for facing the problem of conformance check of file formats for long-term preservation; finally, Section 4 briefly reports on the current status of the project.

[5] http://www.preforma-project.eu/

2 Pre-Commercial Procurement

PCP[6] is an instrument which has been pioneered by the European Commission in the FP7 and is now fully part of Horizon 2020. PCP is appropriate when the required improvements are so technologically demanding that there are no near-to-the-market solutions yet and new R&D is needed. PCP can then be used to compare the pros and cons of alternative competing approaches and to de-risk the most promising innovations step-by-step via solution design, prototyping, development and first product testing.

PCP is a competition-like procurement method which enables public sector bodies to engage with innovative businesses and other interested organisations in development projects, to arrive at innovative solutions that address specific public sector challenges and needs. The new innovative solutions are created through a phased procurement of development contracts to reduce risk.

PCP operates by clustering together stakeholders in a given domain – memory institutions in the case of PREFORMA – which group together in order to face a common technological challenge. The stakeholders consortium is in charge of describing the expected technological solution, specifying its needed features and characteristics, and defining how alternative approaches will be compared and assessed in order to understand their pros and cons. The consortium manages an open call for tenders where technological suppliers can apply and the selected suppliers will go ahead with the design and implementation of the requested technological solution. The consortium is responsible for monitoring the progress of the suppliers work towards the first product testing and for evaluating the final solution developed by the suppliers in order to understand which one best fits with their actual needs.

Several areas are now experimenting the PCP instrument with a range of successful projects: health care and elderly care, transportation, energy, education, safety and public administration [2].

The PCP instrumente has several benefits, among which, to help:

- driving innovation in the public sector;
- reducing fragmentation in the demand of the public sector;
- stimulating the ICT developers to offer more challenging solutions;
- reducing the time-to-market of high-tech solutions.

Moreover, in some areas where there is less abundance of funding, as it is often the case of cultural heritage, grouping together stakeholders may help in promoting the development of required technologies whose costs would not be affordable by the single stakeholder alone.

3 The PREFORMA Approach

PREFORMA aims to establish a set of tools and procedures for gaining full control over the technical properties of digital content intended for long-term preservation by memory institutions [11].

[6] `http://ec.europa.eu/digital-agenda/en/pre-commercial-procurement`

The procurement, following the rules for tenders in public sector, will match the memory institutions professional knowledge and the supplier's skills in development and promotion of products, creating a win-win situation. Joint procurement enables PREFORMA to build a sustainable network of common interest, where the public procurers can remain in contact and cooperate beyond the EC funding period.

The main objective of the project is the development and deployment of an open source software licensed reference implementation for file format standards aimed for any memory institution (or other organisation with a preservation task) wishing to check conformance with a specific standard. This reference implementation, called the *conformance checker* will consist of a set of modular tools, which will be validated against specific implementations of specifications of standards relevant to the PREFORMA project and used by the European memory institutions for preserving their different kind of data objects.

In order to demonstrate effectiveness (and refine) these conformance checkers, they will be developed in an iterative process with multiple releases and with a number of experiments with "real" data sets (files) from memory institutions during each iteration.

A conformance checker:

- verifies whether a file has been produced according to the specifications of a standard file format, and hence,
- verifies whether a file matches the acceptance criteria for long-term preservation by the memory institution,
- reports in human and machine readable format which properties deviate from the standard specification and acceptance criteria, and
- performs automated fixes for simple deviations in the metadata of the preservation file.

The conformance checker software developed by PREFORMA is intended for use within the *Open Archival Information System (OAIS)* Reference Framework [6] and development is guided by the user requirements provided by the memory institutions that are part of the PREFORMA consortium. The conformance checker facilitates memory institutions in obtaining sufficient control of the information in an OAIS Archive, provided to the level needed to ensure Long Term Preservation [12]. In particular, the conformance check enables implementation of the following OAIS functions [12]: (i) *Quality assurance* at ingestion, validating the successful transfer of the *Submission Information Package (SIP)* to the temporary storage area; (ii) *Generate AIP* at ingestion, transforming one or more SIPs into one or more *Archival Information Packages (AIPs)* that conform to the Archives data formatting standards and documentation standards; and, (iii) *Archival Information Update* at ingestion, providing a mechanism for updating (repackaging, transformation) the contents of the Archive.

The media types addressed by PREFORMA are: (i) *text* for establishing a reference implementation for PDF/A [7–9]; (ii) *images* for establishing a reference implementation for uncompressed TIFF [4, 5]; and, (iii) *audio-video* for establishing a reference implementation for an audiovisual preservation file, using

FFV1[7], Dirac[8] or JPEG2000 [10] for encoding video or moving image, uncompressed LPCM [3] for encoding sound and MKV[9] or OGG[10] for wrapping audio- and video-streams in one file.

The experience of applying the reference implementation to the format standards and other observations from the development will be used by PREFORMA to give feedback to relevant standardization organizations and other relevant stakeholder groups (e.g. legislators and other suppliers). Such feedbacks can be used to detail the precise interpretation of the specification during the creation of new versions of the standards.

4 Status of PREFORMA

PREFORMA has been launched in January 2014 and, on April 4th, 2014, it organised an Information Day[11] event in Brussels to present the call for tender, which has been launched as part of the PCP.

The call for tender opened on June 12th, 2014 and closed on August 12th, 2014 with a budget of 2,805,000 euros. 16 high-quality proposals have been submitted to the call out of which 6 have been selected[12] for continuing with the subsequent first design phase, which started in November 2014.

The first design phase will terminated in late February 2015 and the 6 suppliers which won the tender have been evaluated again according to well-defined criteria [1] in order to select those worth continuing with the subsequent development phase. The outcomes of the first evaluation phase have been reported on early March 2015 during a workshop which will be organized in Brussels[13].

After having analyzed the technical and functional specifications submitted by the six groups that completed the design phase, PREFORMA chose the three consortia awarded with contracts for the prototyping phase, which will last until December 2016.

The three awardees are the veraPDF Consortium (led by the Open Preservation Foundation and the PDF Association), who is working on the compliance checker for the PDF/A standard for documents; Easy Innova, who is working on the TIFF standard for digital still images; and MediaArea, who is working on a set of open source standards for moving images, namely: the Matroska wrapper, the FFv1 video codec and LPCM for audio streams.

Connected to their work in PREFORMA Easy Innova and their partner University of Basel are leading an initiative to create an ISO Standard to optimize the TIFF format definition for archival purposes, the so called TI/A Standard Initiative[14].

[7] http://www.ffmpeg.org/~michael/ffv1.html
[8] http://diracvideo.org/
[9] http://www.matroska.org/
[10] https://xiph.org/ogg/
[11] http://www.digitalmeetsculture.net/article/follow-up-of-the-preforma-information-day/
[12] http://www.preforma-project.eu/successful-proposals.html
[13] http://www.preforma-project.eu/design-phase-1-final-workshop.html
[14] http://www.ti-a.org/

The first releases of the conformance checkers, developed by the three suppliers during the first part of the prototyping phase, are now available for download on the Open Source Portal section of the PREFORMA website[15] . This section provides an overview and references to each open source project, acting as an entry point for all interested suppliers and memory institutions and allowing easy navigation to all externally hosted resources.

The preliminary result of the prototyping phase will be presented and demonstrated during the "Open Source Preservation Workshop Serving the Cultural Heritage", the first in a series of international events planned by PREFORMA. The workshop will take place in Stockholm on April 7, 2016, and will be hosted at the National Library of Sweden[16].

The event will feature keynote presentations by representatives from the PREFORMA project and the open source community, live demonstrations of the three conformance checkers for electronic documents, images and AV files by the suppliers working in the project (veraPDF, Easy Innova, MediaArea) and an informal networking event where all the attendees can share experiences, meet the PREFORMA developers and learn about the tools.

This event is aimed at anyone interested in digital preservation and cultural heritage: developers who want to contribute code to the PREFORMA tools; memory institutions or other cultural heritage organisations involved in (or planning) digital preservation initiatives; standardisation bodies maintaining the technical specifications of preservation file formats; any other person interested in cooperating with us in defining open digital preservation standards.

In addition to the public events, PREFORMA is very much committed to communicate online with its community. To this aim, the project developed two instruments that are constantly updated and monitored to exchange ideas and get feedbacks from all the interested people: the project website[17] and the project blog[18].

Furthermore, fruitful cooperation has been established with many other institutions, organisations and projects that expressed their interest to cooperate with PREFORMA to test the software prototypes and to improve the conformance checkers. In this light, it is worth citing BenchmarkDP[19], an interdisciplinary multi-year research project funded by the Vienna Science and Technology Fund, and Europeana Space[20], an EU-funded CIP Best Practice Network. In the first case, methodologies and approaches of BenchmarkDP could be useful in PREFORMA in establishing an objective frame of reference for the evaluation of the conformance checkers, while in the second case the aim is to test the integration of the open source conformance checkers in the Technical Space developed by Europeana Space.

---

[15] http://www.preforma-project.eu/open-source-portal.html
[16] http://opensourceworkshop.preforma-project.eu/
[17] http://www.preforma-project.eu/
[18] http://www.digitalmeetsculture.net/projects/preforma/
[19] http://benchmark-dp.org/
[20] http://www.europeana-space.eu/

Overall, the experience of PREFORMA is demonstrating that putting in place a joint PCP is very challenging, but it is also offering an opportunity of growth and learning, not only for the procurers (namely the memory institutions) but also for the technical partner who are supporting the implementation of the call.

References

1. Agosti, M., Ferro, N., Lemmens, B., Silvello, G.: Deliverable D8.1 – Competitive Evaluation Strategy. PREFORMA PCP Project, EU 7FP, Contract N. 619568. http://www.digitalmeetsculture.net/wp-content/uploads/2014/12/PREFORMA_D8.1_Competitive-evaluation-strategy_v1.0_no-appendix.pdf (December 2014)
2. European Commission: Innovation Procurement – The power of the public purse. http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=5443 (May 2014)
3. IEC 60958: Digital audio interface - Part 1: General. Standard IEC 60958-1 Ed. 3.1 b:2014 (2014)
4. ISO 12234-2: Electronic still-picture imaging – Removable memory – Part 2: TIFF/EP image data format. Recommendation ISO 12234-2:2001 (2001)
5. ISO 12639: Graphic technology – Prepress digital data exchange – Tag image file format for image technology (TIFF/IT). Recommendation ISO 12639:2004 (2004)
6. ISO 14721: Space data and information transfer systems – Open archival information system (OAIS) – Reference model. Recommendation ISO 14721:2012 (2012)
7. ISO 19005-1: Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1). Recommendation ISO 19005-1:2005 (2005)
8. ISO 19005-2: Document management – Electronic document file format for long-term preservation – Part 2: Use of ISO 32000-1 (PDF/A-2). Recommendation ISO 19005-2:2011 (2011)
9. ISO 19005-3: Document management – Electronic document file format for long-term preservation – Part 3: Use of ISO 32000-1 with support for embedded files (PDF/A-3). Recommendation ISO 19005-3:2012 (2012)
10. ISO/IEC 15444: Information technology – JPEG 2000 image coding system: Core coding system. Recommendation ISO/IEC 15444-1:2004 (2004)
11. Lemmens, B., Elfner, P., Lundell, B., Prandoni, C., Fresa, A.: Deliverable D2.2 – Tender Specifications. PREFORMA PCP Project, EU 7FP, Contract N. 619568. http://www.digitalmeetsculture.net/wp-content/uploads/2014/05/PREFORMA_D2.2_Tender-Specifications_v2.1.pdf (June 2014)
12. The Consultative Committee for Space Data Systems (CCSDS): Reference Model for an Open Archival Information System (OAIS). Magenta Book, Issue 2. Recommended Practice CCSDS 650.0-M-2, http://public.ccsds.org/publications/archive/650x0m2.pdf (June 2012)