ARTEFACT DOCUMENT

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

Challenge Brief

Revision: ver 1.0

Authors:

Bert Lemmens (Packed vzw)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

TABLE OF CONTENTS

1 THE PREFORMA CHALLENGE

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.

(*Digital content*)

Digital content is stored in files with specific file containers and encodings for capturing text, images, sound and moving image, depending on the systems and use cases the files originate from.

These files can be produced by external organizations and transferred to memory institutions, or produced in-house by memory institutions as digital reproductions of items in their collections and holdings.

(*Long-term preservation*)

Long-term preservation of these files requires exact knowledge and control over their technical properties, allowing memory institutions to develop an appropriate preservation strategy for the digital content (e.g. by transforming, re-packaging and emulating these files).

(*Full control*)

To obtain this knowledge and control, preservation files are usually generated using ‘standard’ file formats, which normalize the way digital content is captured in a digital file. Yet, these preservation files are always generated using software that implements one particular interpretation of the ‘human readable’ specifications in the ‘standard’ document. Inevitably, ambiguities in these specifications lead to different interpretations and hence software producing different implementations of the ‘standard’ file format.

In practice, a uniform implementation of a ‘standard’ file format is hard to enforce, because the software producing preservation files is controlled neither by the organization that produces them, nor by the memory institution that has to preserve the file.

2 RESEARCH AND DEVELOPMENT

The PREFORMA *research and development*[1] activities, deployed by the triple helix of suppliers, universities and memory institutions, are conceived as *applied research*[2], exploring critical factors in the quality of standard implementation through development of two strategies that empower memory institutions to gain control over the technical properties of preservation files.

(strategies)

PREFORMA develops two strategies that empower memory institutions to gain control over the technical properties of preservation files:

- develop an open-source conformance checker, and
- establish a healthy ecosystem around an open source 'reference' implementation for specific file formats.

(research topics)

The *basic research*[3] objective, underlying these two strategies, is to explore critical factors in the quality of standard implementation. This involves acquiring knowledge about:

- how to establish a methodology or an objective frame of reference to interpret and implement the standard specifications against the background of the current variations of interpretations and implementations by software vendors; is there a need to consolidate the diverse implementations or is a better approach to centralize the interpretation to a specific implementation (i.e. promote one interpretation and implementation as the standard)?
- given the answer to the first question, how to determine whether a file is what it claims to be, i.e., in this context, what makes a file a valid file, i.e., conform to the "standard"?

[1] The PREFORMA R&D effort is conceived along the *research & development* concept defined in par. 63 of the Frascati Manual:
"(63) *Research and experimental development* (R&D) comprise creative work undertaken on a systematic basis in order to increase the stock of knowledgy, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications."
OECD, Frascati Manual. Proposed standard practice for surveys on reserach and experimental development, PDF file, 2002: 31/ Accessed on 9 June 2014. http://browse.oecdbookshop.org/oecd/pdfs/free/9202081e.pdf

[2] The PREFORMA R&D effort is conceived along the *applied research* concept defined in par. 64 of the Frascati Manual:
"(64) *Basic research* is experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundation of phenomena and observable facts, without any particula application or use in view.
Applied research is also original investigation undertaken in order to acquire new knowledge, It is , however, directed primarily towards a specific practical aim or objective. [...]"
ibid.

[3] The PREFORMA R&D effort is conceived along the *basic research* concept defined in par. 64 of the Frascati Manual:
" (64) *Basic research* is experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundation of phenomena and observable facts, without any particula application or use in view.
Applied research is also original investigation undertaken in order to acquire new knowledge, It is , however, directed primarily towards a specific practical aim or objective. [...]"
ibid.

- how can the open source project continue to be developed and sustained in the short and long run; can an open source community operate as the normative source for the answer to the first and second question?

Besides providing a free and open solution for conformance checking at memory institutions, the development of a conformance checker and a reference implementation are instrumental to these research questions. The output of the research is shared with the stakeholders that contribute to the ecosystem around the reference implementation.

(coordination of research)

Research and development activities are performed within the 'triple helix' of universities, memory institutions and suppliers.

- suppliers are the locus for the development of the conformance checker software and the test files that exemplify the reference implementation
- universities provide knowledge and technology for designing, prototyping and testing the software and the test files,
- memory institutions provide the organizational framework for the R&D activities and establish a network of common interest for sharing the results of the project.

Suppliers can take advantage of the interaction with universities and memory institutions to raise their technological level. On the other hand they will benefit from the recommended practices set forth by the network of common interest, since this will reduce fragmentation for their market. They will be able to develop single solutions for a larger market and to be more aware of and closer to the actual demand of their customers.

(open source project)

The use of open file formats and open source software is considered to be fundamental for establishing long-term sustainable preservation workflows at memory institutions. As for PREFORMA, it ensures long-term availability of the conformance checking software and test files, independent of the specific memory institutions and suppliers involved in the PREFORMA PCP.

Therefore the research & development activities of PREFORMA will be organized in a series of open-source projects that each focus on establishing a reference implementation for one particular media type.

3 OPEN SOURCE PROJECTS

The PREFORMA open source projects each address a particular set of standard file formats that are (1) open standards[4], (2) considered appropriate for long term preservation by digital preservationists and (3) relevant for the memory institutions participating in PREFORMA.

These file formats cover three major media types: text, image and audiovisual media.

3.1 TEXT

The open source project on text media researches tools and procedures for establishing a reference implementation for **PDF/A**.

Basic research should involve checking for the existence of PDF/A -functionalities and whether they are implemented in accordance with the specifications for PDF/A. The functionalities to examine could for example be the following:

- the use of annotations, images and digital signatures,
- the use of compression schemes and transparency,
- the use of character sets and fonts,
- the use of interactive features such as executable scripts, forms and navigation tools,
- embedded administrative and structural metadata, and
- dependencies on external resources.

The conformance checker developed in this open-source project validates PDF/A files against all following standard file format specifications:

- ISO (2005). Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1). ISO/TC 171/SC 2, ISO 19005-1:2005.
- ISO (2008). Document management -- Portable document format -- Part 1: PDF 1.7. ISO/TC 171/SC 2, ISO 32000-1:2008.

[4] O**pen standards** are standard specifications that meet the following requirements, defined in the European Interoperability Framework for Pan-European eGovernment Service (version 1.0 2004):

- The standard is adopted and will be **maintained by a not-for-profit organization**, and its ongoing development occurs on the basis of an open decision-making procedure available to all interested parties (consensus or majority decision etc.).
- The standard has been published and the standard specification document is **available either freely or at a nominal charge**. It must be permissible to all to copy, distribute and use it for no fee or at a nominal fee.
- The **intellectual property** - i.e. patents possibly present - of (parts of) the standard is made irrevocably **available on a royalty-free basis**.
- There are no **constraints on the re-use** of the standard

- ISO (2011). Document management -- Electronic document file format for long-term preservation -- Part 2: Use of ISO 32000-1 (PDF/A-2). ISO/TC 171/SC 2, ISO 19005-2:2011.
- ISO (2012). 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). ISO/TC 171/SC 2, ISO 19005-3:2012.

The conformance checker determines if a file is a PDF/A file or if it is something else. It also checks all different conformance levels in **PDF/A-1**, and **PDF/A-2**, and **PDF/A-3** and provides detailed information on the criteria that have not been fulfilled.

3.2 IMAGE

The open source project on image media researches tools and procedures for establishing a reference implementation for **uncompressed TIFF**.

Basic research should involve checking for the existence of TIFF -functionalities and whether they are implemented in accordance with the specifications for TIFF. The functionalities to examine could for example be the following:

- the use of ‘baseline’, ‘extension’ and ‘private tags’,
- the use of color profiles,
- assessment of user-specific acceptance criteria based on technical parameters of the digital image, and
- executing automated fixes for making TIFF files compliant with baseline specifications,

The conformance checker developed in this open-source project validates uncompressed TIFF files against all following standard file format specifications:

- ISO (2001). Electronic still-picture imaging — Removable memory — Part 2: TIFF/EP image data format. ISO/TC 42, ISO 12234-2:2001
- ISO (2004). Graphic Technology -- Prepress digital data exchange -- Tag image file format for image technology (TIFF/IT). ISO/TC 130. ISO 12369:2004

The conformance checker determines if a file is an uncompressed TIFF file or if it is something else. It also checks the use of ‘baseline’, ‘extension’ and ‘private’ tags, as well as user-specific acceptance criteria. The tool provides detailed information on the criteria that have not been fulfilled.

3.3 AUDIOVISUAL

The open source project on audiovisual media researches tools and procedures for establishing a reference implementation for an audiovisual preservation file, using **FFV1**, **Dirac** or **JPEG2000** for encoding video or moving image, **uncompressed LPCM** for encoding sound and **MKV** or **OGG** for wrapping audio- and video-streams in one file.

Basic research activities involve defining a profile for an audiovisual preservation file that allows for:

- capturing uncompressed or mathematically lossless compressed audio- and video- or image streams,
- preserving the image and sound properties of the 'original' audio-visual resource,
- capturing a comprehensive set of preservation data

Basic research should also involve checking for the existence of standard functionalities and whether they are implemented in accordance with the corresponding specifications.

The conformance checker developed in this open-source project validates one container file format and one video/image codec, chosen by the supplier from the standards mentioned above, and uncompressed LPCM encoded audio. The supplier is asked to select a specific set of standards from the list below.

- MKV: Matroska – Technical Details. http://www.matroska.org/technical/index.html
- OGG: Ogg - Documentation. https://xiph.org/ogg/doc/
- JPEG2000: ISO (2004). Information technology - JPEG 2000 image coding system: Core coding system. ISO/IEC JTC 1/SC 29, ISO/IEC 15444-1:2004
- FFV1: FFV1 Video Codec Specification, http://www.ffmpeg.org/~michael/ffv1.html
- Dirac: Dirac Specification Version 2.2.3 (2008), http://diracvideo.org/download/specification/dirac-spec-latest.pdf
- LPCM: IEC (2014). Digital audio interface - Part 1: General. IEC/TC 100, IEC 60958-1 ed3.1 Consol. with am1: 2014

This selection provides the reference implementation that the files will be checked against. The conformance checker determines if a file is conform the selected standard specifications or if it is something else. It also provides detailed information on the criteria that have not been fulfilled.

4 CONFORMANCE CHECKER

The first strategy researched in PREFORMA is to develop an open-source toolset for conformance checking of digital files, intended for long-term preservation in memory institutions.

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 OAIS Reference Framework[5] and development is guided by the user requirements provided by the memory institutions that are part of the PREFORMA consortium.

4.1 OAIS ENVIRONMENT (REFERENCE FRAMEWORK)

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*[6].

The conformance check enables implementation of the following OAIS functions[7]:

- **Quality assurance** at Ingest, validating (QA results) the successful transfer of the SIP to the temporary storage area.
- **Generate AIP** at Ingest, transforming one or more SIPs into one or more AIPs that conform to the Archive's data formatting standards and documentation standards.
- **Archival Information Update** at Ingest, providing a mechanism for updating (repackaging, transformation) the contents of the Archive.

Additionally, the conformance checker must allow Producers to check whether a file conforms to the technical criteria before submission of a file to an OAIS Archive.

[5] ISO 14721:2012 Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model
[6] CCSDS 650.0-M-2 Reference Model for an Open Archival Information System (OAIS). Magenta Book. Issue 2, PDF file, June 2012: 38. Accessed on 12 May 2014. http://public.ccsds.org/publications/archive/650x0m2.pdf.
[7] CCSDS 650.0-M-2 Reference Model for an Open Archival Information System (OAIS). Magenta Book. Issue 2, PDF file, June 2012: 4-6 – 4-8 . Accessed on 12 May 2014. http://public.ccsds.org/publications/archive/650x0m2.pdf.

4.2 USE CASES (DOCUMENT LIFE CYCLE)

Development of the conformance checker focuses on four use cases that facilitate the interaction between the supplier, academic research and memory institution. They are compliant with the OAIS Reference Model and represent conformance checking procedures at different moments in the life cycle of a preservation file:

- Conformance Checking at **Creation Time**: Producers pro-actively check if technical properties of a file meet the acceptance criteria of an OAIS Archive, e.g. government agencies checking conformance of text documents to be deposited at public archives when the document is made available.
- Conformance Checking at **Transfer time**: Archives check the technical properties of files ingested in the OAIS Archive, assessing whether they meet the acceptance criteria for ingest and conformance to the relevant preservation file formats, e.g. libraries monitor the preservation status of digital publications deposited in their digital repository.
- Conformance Checking at **Digitization time**: Archives check the technical properties of digital representations of collection items, internally or externally produced, if they meet the requirements specified in the digitization tender, e.g. museums doing quality control on the digital representations and documentation, produced by photographers.
- Conformance Checking at **Migration time**: Archives check the technical properties of files that are repackaged or transcoded, following the rules defined in the preservation strategy of the OAIS Archive, e.g. libraries doing quality control when transcoding audiovisual files from a ‘transmission’ to a ‘preservation’ format.

4.3 DEPLOYMENT (ENVIRONMENT)

The conformance checker allows for deployment in different infrastructures and environments.

- **PREFORMA Website**: Deployment at the PREFORMA project website, demonstrating the scope and functionality of the tool. The PREFORMA website should be considered as the deliverable for the PREFORMA project.
- **Evaluation framework**: Deployment within an evaluation framework that allows for gathering structured feedback on the conformance checking process. PREFORMA will require deployment within the DIRECT infrastructure for test and evaluation of the tool in the PCP procedure.
- **Stand-alone**: The tool must allow for packaging it in an executable and run it on a PC. This ensures the conformance checker can be used in small-scale institutions without centralized IT infrastructure.
- **Networked**: The tool must allow for deployment in network-based solutions (dedicated server, cloud solutions) for digital repositories.
- Integration in **legacy systems**: The tool must allow for plugging it into proprietary legacy systems via API’s.

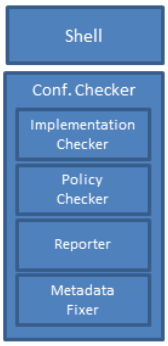
4.4 COMPONENTS (ARCHITECTURE)

The conformance checker allows for modular deployment. Since OAIS archives usually contain multiple file types, the conformance checker should enable checking multiple reference implementations in one operation. For this purpose, the conformance checker allows for integrating other conformance checker components maintained within the PREFORMA ecosystem via one shell.

The design of the conformance checker therefore complies with the following functional architecture that facilitates modular deployment of multiple conformance checkers in one tool.

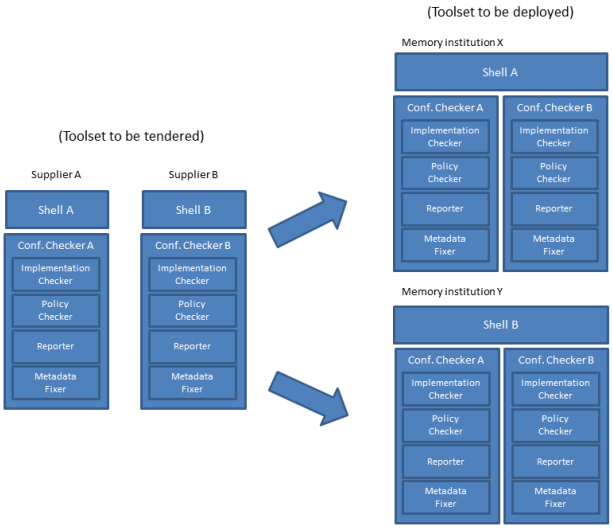
The conformance checker comprises four functional components:

- the **shell**: The conformance checker should interface with other systems through a 'shell' which allows for interfacing multiple conformance checkers at the same time. This might in the future allow integrating the conformance checkers of different suppliers into one application.
- the '**implementation checker'**: which performs a comprehensive check of the standard specifications listed in the standard document.



- the '**policy checker'**: which allows for adding acceptance criteria, always compliant with the standard specifications, that further differentiates the properties of the file. This might for example include limiting conformance to PDF/A-1b. Or limiting the colors-space of TIFF-files to ECI-RGB or aRGB. Or limiting the audiovisual file to 'progressive scanned' video.
- the '**reporter'**: which interprets the output of the implementation checker and policy checker and allows for defining multiple human and machine readable output formats. This might include a well-documented JSON or XML file, a human readable report on which specifications are not fulfilled, or a fool-proof report which also indicates what should be done to fix the non-conformances.
- the '**metadata fixer'**: which allows for simple fixes of the metadata embedded in the file, making them compliant with the standard specification.

Integration of multiple conformance checkers via one shell should be established as follows:



4.5 OPEN-SOURCE DEVELOPMENT

The open-source approach is fundamental for achieving the overall objectives of the PREFORMA challenge. This approach implies that:

- all software developed in the PREFORMA project will be released using established open source development practices with early and frequent releases of developed software and associated artefacts.
- all software developed in the PREFORMA project is licensed under “GPL v3 or later” and “MPL v2 or later”, enabling that anyone that has adopted such software has the right to freely read, use, improve and redistribute the source code for such software.
- all software developed in the PREFORMA project will be made available on an open platform, e.g. GitHub or equivalent.
- all file formats researched in the PREFORMA project will be available under licensing conditions that allow for implementation in open-source software, including allowing for implementation in open source software which is licensed under “GPLv3 or later” and “MPLv2 or later”.
- all files produced in the PREFORMA project will be released under a CC-BY-SA license.

(long term sustainable preservation workflow)

This open-source approach must ensure that memory institutions will always have access to the required tools for deploying a long-term sustainable preservation workflow, supported and maintained by the associated ecosystems.

(business opportunities)

PREFORMA will advocate a number of business opportunities for the selected technology providers, all in line with presently used Open Source software models:

1. Combinations with other software offerings

Technology providers are invited to use the open source software developed in PREFORMA as a complement to proprietary licensed software products in their portfolio, e.g.

- in combination with text, image or moving image editors, facilitating the production of preservation files
- in combination with digital repositories, facilitating assessment of files being ingested and processed by a Trusted Digital Repository.
- in combination with transcoding software, facilitating validation when migrating files.

2. Offer supplementary proprietary solutions

Technology providers are invited to develop supplements to the open source software and provide them using other licenses during and after the PREFORMA project. This may include:

- additional conformance checkers for other preservation formats that plug into the same environment/ecosystem
- additional reporter modules that facilitate integration of the open-source software in other proprietary software products

3. Selling professional services

Technology providers are invited to provide services for deploying the open source software at memory institutions, e.g. providing

- consulting
- customization
- technical support

(pre-commercial procurement)

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

5 REFERENCE IMPLEMENTATION

The conformance checker authorizes a 'reference implementation' for a standard file format, i.e. an implementation of a 'standard' specification that is to be used as a definitive interpretation for that 'standard' specification.
The second strategy to gain control over the technical properties of preservation files is to establish a **network of common interest** that advances:

- the **adoption** of such a 'reference implementation' by other software applications, and
- continuous **improvement** of the 'standard' specification through engagement in the standardization process.

5.1 OAIS ENVIRONMENT

The network of common interest facilitates following OAIS functions by memory institutions operating an OAIS Archive[8]:

- **Monitor Designated Communities** for Preservation Planning, interacting with Archive Consumers and Producers to track changes in their service requirements and available product technologies.
- **Develop Preservation Strategies and Standards** for preservation planning, developing and recommending strategies and standards, and for assessing risks, to enable the Archive to make informed tradeoffs as it establishes standards, sets policies, and manages its system infrastructure.
- **Establishing Standards and Policies** by the Administration of the Archive system and maintain them.

5.2 STAKEHOLDERS

The network gathers all stakeholders that control different stages in the lifecycle of a preservation file, providing a sustainable and viable ecosystem for the deployment of tools developed by PREFORMA as well as tools adopting the reference implementation. These stakeholders include:

- **developers**, controlling the production of preservation files, e.g. by file editors or transcoders, thus aiming at improving the effectiveness and interoperability of their software.

[8] CCSDS 650.0-M-2 Reference Model for an Open Archival Information System (OAIS). Magenta Book. Issue 2, PDF file, June 2012: 4-12 – 4-15 . Accessed on 12 May 2014. http://public.ccsds.org/publications/archive/650x0m2.pdf.

- **digital preservationists**, controlling the acceptance and management of preservation files in digital repositories, thus aiming at improving the preservation status of the digital collection they maintain and the effectiveness of the ingest procedures.
- **standardization bodies**, maintaining the formal specifications of file formats in standards, thus aiming to improve the specification of the standard.

5.3 ADOPTION

PREFORMA communicates the achievements of the project with developers and digital preservationists, raising interest to integrate the reference implementations in existing software products and digital repositories. PREFORMA establishes appropriate communication procedures to provide stakeholders with technical information facilitating adoption by the corresponding stakeholder.
These procedures include contributions by technology providers, such as:

- providing demonstration files with good and bad samples of the corresponding reference implementation,
- providing comprehensive documentation of the source code, which allows for automated generation of the internal API of the application,
- providing comprehensive documentation of the conformance checker for developers, such as quick start guide, cookbooks and other tutorials,
- online availability at the development platform for technical support to other developers deploying the conformance checker, and
- marketing the reference implementation and conformance checker at conference for professional networks of developers and digital preservationists.

5.4 IMPROVEMENT

PREFORMA researches critical factors in the quality of the ‘reference implementation’ by assessing the files and tools produced by the project. Technical details on precisely what is incorrect will be tailored to the needs of each specific stakeholder group. PREFORMA establishes appropriate feedback procedures for each stakeholder to share the results of this assessment.
These procedures include contributions by technology providers, such as:

- Drafting proposals for changes and additions to the standard specifications, and
- Participating in technical workgroups that maintain a standard specification.