

# DELIVERABLE

**Project Acronym:** PREFORMA

**Grant Agreement number:** 619568

**Project Title:** PREservation FORMAts for culture information/e-archives

## D3.7.1 Initial version of Sustainability and Exploitation Plan

**Revision:** version 1.0 – FINAL

### Authors:

Claudio Prandoni (PROMOTER)  
Antonella Fresa (PROMOTER)  
Pietro Masi (PROMOTER)  
Valentina Bachi (PROMOTER)

### Contributors:

All PREFORMA partners  
The PREFORMA suppliers (veraPDF consortium, Easy Innova, MediaArea)

### Reviewers:

Börje Justrell (Riksarkivet)  
Melanie Imming (LIBER)

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	

## Revision History

Revision	Date	Author	Organisation	Description
0.1	10/11/2016	Claudio Prandoni	Promoter	First structure and outline
0.2	29/11/2016	Claudio Prandoni, Antonella Fresa, Pietro Masi, Valentina Bachi	Promoter	Revised structure and added content to Sections 2, 4, 5
0.3	21/12/2016	Claudio Prandoni	Promoter	Updated draft
0.4	11/01/2017	Claudio Prandoni	Promoter	Integrated input by Easy Innova
0.5	12/01/2017	Claudio Prandoni	Promoter	Integrated input by Melanie Imming
0.6	13/01/2017	Claudio Prandoni	Promoter	Intergated input by MediaArea and veraPDF consortium
0.7	17/01/2017	Claudio Prandoni	Promoter	Added conclusions and next steps
0.8	19/01/2017	Claudio Prandoni	Promoter	Updated draft including the Annexes and integrating the comments from Borje Justrell and Antonella Fresa
1.0	27/01/2017	Claudio Prandoni	Promoter	Final version

### Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

## TABLE OF CONTENTS

EXECUTIVE SUMMARY .....	4
1. INTRODUCTION .....	5
1.1 OPPORTUNITIES OFFERED BY PREFORMA .....	5
1.2 THE OPEN SOURCE MODEL .....	6
2. MAIN PREFORMA OUTCOMES.....	8
2.1 VERAPDF.....	8
2.2 DPF MANAGER.....	10
2.3 MEDIACONCH.....	12
3. OPEN SOURCE COMMUNITY DEVELOPMENT .....	14
3.1 VERAPDF.....	14
3.2 DPF MANAGER.....	16
3.3 MEDIACONCH.....	19
4. IMPACT OF PREFORMA ON THE STAKEHOLDERS .....	23
4.1 IMPACT ON THE MEMORY INSTITUTIONS .....	23
4.2 IMPACT ON COMPANIES AND TECHNOLOGY PROVIDERS.....	28
4.3 IMPACT ON OTHER DOMAINS.....	31
4.4 IMPACT ON THE STANDARDISATION ACTIVITIES.....	32
4.4.1 <i>Electronic documents: PDF/A and PDF/A-next</i> .....	32
4.4.2 <i>Still images: TIFF and TI/A</i> .....	33
4.4.3 <i>Audiovisual files: Matroska and Ffv1</i> .....	35
5. SUSTAINABILITY, FUTURE USE AND EXPLOITATION PLANS .....	37
5.1 SUSTAINABILITY OF THE PREFORMA RESULTS .....	37
5.1.1 <i>Sustainability of the PREFORMA network</i> .....	37
5.1.2 <i>Sustainability of the PREFORMA infrastructure and data</i> .....	38
5.1.3 <i>Sustainability of the PREFORMA tools</i> .....	40
5.2 ROUTE TO MARKET OF THE PREFORMA TOOLS.....	42
5.2.1 <i>Market Analysis</i> .....	42
5.2.2 <i>Market Segmentation</i> .....	46
5.2.3 <i>Competition</i> .....	49
5.2.4 <i>Costs and revenues analysis</i> .....	50
6. NEXT STEPS AND CONCLUSION.....	54
ANNEX 1: QUESTIONNAIRE ON IMPACT ASSESSMENT AND EXPLOITATION OF PREFORMA RESULTS.....	58
ANNEX 2: IMPACT ASSESSMENT INTERVIEW MODEL.....	64

## EXECUTIVE SUMMARY

This deliverable describes the initial version of the sustainability and exploitation plan of PREFORMA, including a preliminary impact assessment conducted through an analysis of the level of activity generated in the community, of the test cases carried out so far and of the results of a survey circulated to the memory institutions participating to the project, both as partners and as associate partners within the network of common interest.

The following communities have been considered in relation to their expected engagement in exploiting the PREFORMA applications:

- the open-source community of researchers and developers interested in contributing to the code;
- memory institutions willing to integrate the PREFORMA software in their infrastructure;
- the community of enterprises interested in PREFORMA tools and in developing services around them (first of all the suppliers working in the project);
- the standardization bodies looking for feedback on how to improve and advance the specifications of the standard file formats.

This deliverable features seven Chapters and two Annexes.

- Chapter 1 presents a brief introduction, summarizing who are the main stakeholder groups and the rationale behind the open source approach followed in the project.
- Chapter 2 introduces the main outcomes of the project.
- Chapter 3 reports on the status of the development of the open source community around the tools developed in PREFORMA.
- Chapter 4 presents a first analysis of the impact generated so far on the main stakeholder groups.
- Chapter 5 provides some preliminary ideas about how to sustain the PREFORMA network after the project end and how to bring the developed tools to the market.
- Chapter 6 presents the outcomes of a survey carried out among the partner of the project and the members of the PREFORMA community to evaluate how the PREFORMA outcomes are being perceived by the community.
- Chapter 7 provides the conclusion to the deliverable and the actions planned for the last year of the project.
- Annex 1 contains the questionnaire that has been circulated to the partners and to the members of the community to assess the impact achieved so far and to identify possible strategies to reuse and exploit the results of the project.
- Annex 2 contains the model of interviews that are being carried out by LIBER among the digital libraries that are member of the Association.

# 1. INTRODUCTION

## 1.1 OPPORTUNITIES OFFERED BY PREFORMA

PREFORMA targets a wide range of stakeholders who might be interested to contribute to the definition of the requirements and to take part in the assessment and the exploitation of the core component and modules developed in the project. Among these, the main target groups are represented by:

- **Developers and enterprises.** PREFORMA targets all enterprises contributing code to the PREFORMA tools and developing services based on these tools, first and foremost the enterprises participating in the tender but subsequently also other developers that are interested to contribute to the code and to implement the reference implementations in production software. Enterprises and public sector organisations may dispose of practical tools to facilitate the production of documents aligned with the preservation demands of the memory institutions archiving their content.
- **Memory institutions** and cultural heritage organisations coordinating or representing them, that are involved in (or planning) digital culture initiatives. PREFORMA targets memory institutions deploying the core components and modules in their digitisation and preservation workflows. The main expected impact of PREFORMA is to reduce curation and preservation costs and to improve curation and preservation capacity and competences in public organisations, including small archives.
- **Research organisations.** PREFORMA targets research organisations with a scientific interest in identification, validation and normalisation of digital files and who may provide technical and expert advice to cultural stakeholders.
- **Standardisation bodies** maintaining the technical specifications of the preservation formats covered in PREFORMA and interested in the efforts made in PREFORMA towards the standardisation of file formats and the improvement of the standard specifications.
- **Funding agencies**, such as Ministries of Culture and national/regional administrations, that own and manage digitisation programmes and may endorse the use of the PREFORMA tools in the digitisation process.

In addition to the main target groups defined above, PREFORMA aims to achieve an impact also on other communities that go beyond the core cultural heritage sector and OS systems developers, e.g. medical institutions and health care in general, content management system vendors/users, industrial archives, institutional repositories, publishers, the web archiving community, etc.

PREFORMA serves through the PCP mechanism a well-defined public need, i.e. to purchase an open source solution to give to memory institutions and to other interested parties the full control over the implementation of the standard formats of their archive.

The availability of such open source project offers to European service and technology providers the possibility to serve this market with more competitive products, opening a new area of added-value services. The suppliers and the members of the open source project will be therefore able to offer support and consultancy services to the memory institutions, for future the adaptation,

improvements, integration and deployment of the PREFORMA tools into their legacy systems and into their networked environments.

## 1.2 THE OPEN SOURCE MODEL

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 is 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 is made available on an open platform,
- all file formats researched in the PREFORMA project are 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 are released under the open access license Creative Commons CC-BY v4.0 and in open file formats, i.e. an open standard as defined in the European Interoperability Framework for Pan-European eGovernment Service (version 1.0 2004).

Successful risk management, sustainability, and longevity of software are critical in PREFORMA.

First, to **minimise risks** all necessary rights need to be obtained by the suppliers to allow provision of open source software on the PREFORMA Open Source Portal under the two specific PREFORMA licences (“GPLv3 or later” and “MPLv2 or later”). This way risks are minimised since the software can be redistributed (in a cascade) for use in other organisations without any additional restriction.

Second, **sustainability** is promoted through the fundamental requirement in PREFORMA which implies that suppliers must provide on the Open Source Portal the source code, the executables and the complete build environment for all open source software under the two specific PREFORMA licences.

Third, to promote **longevity of software**, the inclusion of the “or later” in the PREFORMA licenses as implies that all open source software provided on the Open Source Portal can (in the future) be re-used and incorporated in other software, which will be provided under new versions of the PREFORMA licences.

This open-source approach ensures 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/communities.

The design of the Open Source Software (OSS) components allows for integration and deployment in new software systems that may be used as a pre-system at the memory institution, but also may be integrated in already existing (legacy) systems that are already used at memory institutions.

As the developed code for the file format tools will be provided under licensing conditions appropriate for very long life-cycles, it will minimise the risk of not being future proof and therefore support public sector organisations with a transparent and effective tool for validation of conformance.

## 2. MAIN PREFORMA OUTCOMES

The most significant outcomes of PREFORMA can be split into two main areas.

Outcomes related to the **software development and related activities**. This includes:

- the three **conformance checkers** developed by the suppliers in the PCP (veraPDF, DPF Manager and MediaConch);
- the **interoperability framework/API** (also called PREFORMA Shell) which allows the integration of different conformance checkers into one single application;
- the activities carried out to advance and improve the **standard specifications** of the preservation formats.

Outcomes related to the **coordination actions** put in place by the project. This includes:

- the **PREFORMA Challenge Brief** which summarises the requirements set by the memory institutions and included in the tender;
- **test files and test cases** collected from the institutions participating in the project, either as partner of PREFORMA or as member of the community;
- the **Open Source Portal** where all the open source software developed in the project is made available by the suppliers;
- the creation of a **knowledge base** where to store and maintain the most relevant results and material, including a glossary, all the publications related to the project, the dissemination material produced and the presentations and video recordings of all the conferences, workshops, webinars and training sessions organised by PREFORMA;
- the **PREFORMA community** as a whole, more than 400 people so far from all over the world working together on improving the long-term preservation of digital data.

The key outcomes in the light of the future use and exploitation of the results of PREFORMA are of course those related to the software development activities.

It has to be noted that for each tool a dedicated webpage has been developed and integrated in the PREFORMA website with a downloadable info sheet, links to all the resources and related material and a sub-page where it is possible to browse and download the software releases. This ensures people can quickly see what a tool is about before deciding to invest resources into its uptake. Proper online documentation also aimed at the non-technical stakeholders is a key element for the uptake of open source tools like those developed in PREFORMA.

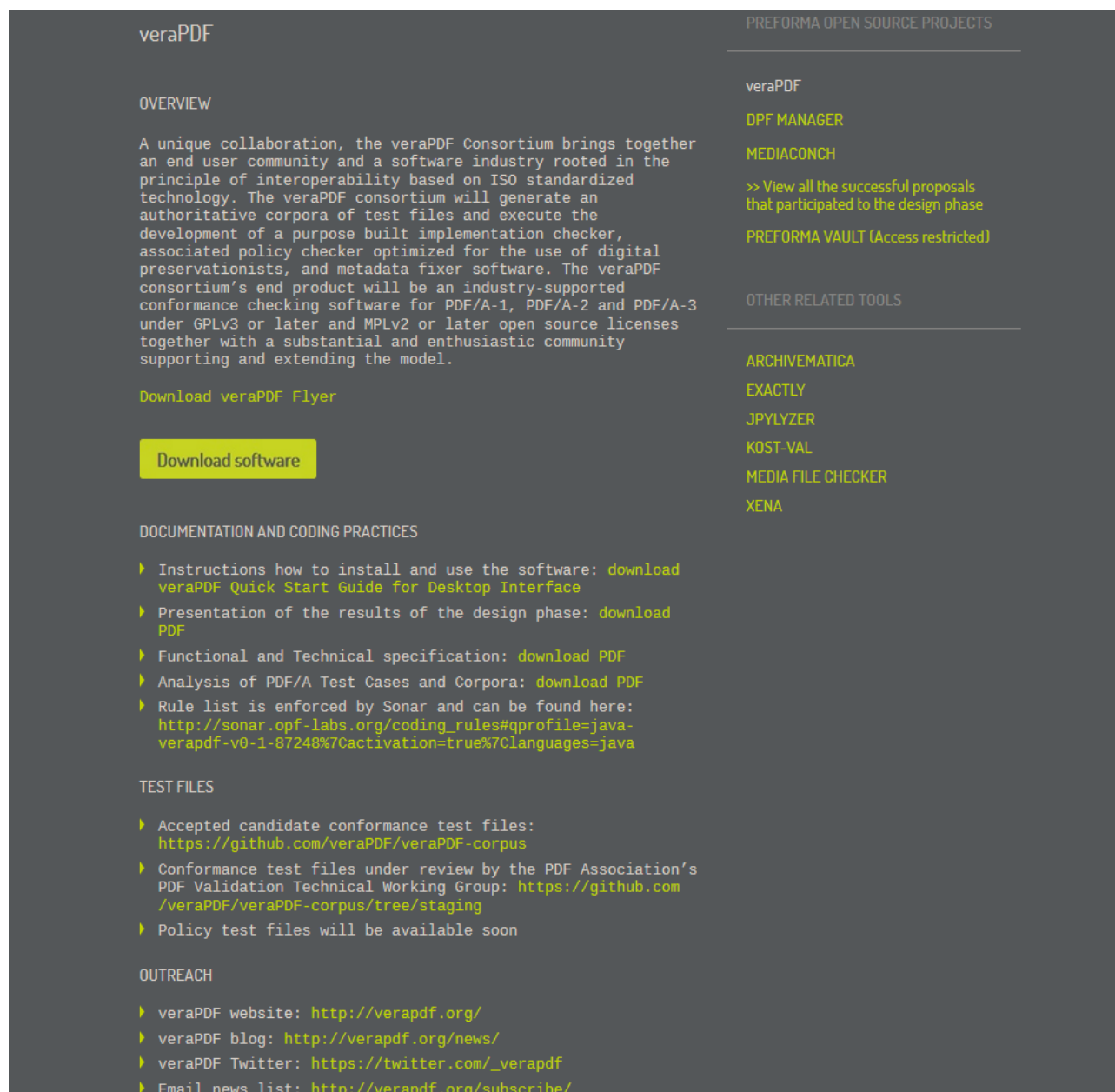
Here below is a summary of the main results achieved by the three suppliers at the end of the prototyping phase.

### 2.1 VERAPDF

veraPDF allows users to evaluate files claiming conformance with all parts and conformance levels of ISO 19005 (PDF/A) for conformance with the PDF/A standard. PDF feature reporting and a customisable policy checker allows organisations to enforce institutional policy beyond the scope



of PDF/A. Additionally, the software performs simple metadata repairs, for example, adding the PDF/A flag to files that only lack the flag in order to conform to the standard.



The screenshot shows the veraPDF project page on the PREFORMA Open Source Portal. The page is divided into two main columns. The left column contains the following sections:

- veraPDF**
- OVERVIEW**

A unique collaboration, the veraPDF Consortium brings together an end user community and a software industry rooted in the principle of interoperability based on ISO standardized technology. The veraPDF consortium will generate an authoritative corpora of test files and execute the development of a purpose built implementation checker, associated policy checker optimized for the use of digital preservationists, and metadata fixer software. The veraPDF consortium's end product will be an industry-supported conformance checking software for PDF/A-1, PDF/A-2 and PDF/A-3 under GPLv3 or later and MPLv2 or later open source licenses together with a substantial and enthusiastic community supporting and extending the model.

[Download veraPDF Flyer](#)

[Download software](#)
- DOCUMENTATION AND CODING PRACTICES**
  - Instructions how to install and use the software: [download veraPDF Quick Start Guide for Desktop Interface](#)
  - Presentation of the results of the design phase: [download PDF](#)
  - Functional and Technical specification: [download PDF](#)
  - Analysis of PDF/A Test Cases and Corpora: [download PDF](#)
  - Rule list is enforced by Sonar and can be found here: [http://sonar.opf-labs.org/coding\\_rules#qprofile=java-verapdf-v0-1-87248%7Cactivation=true%7Clanguages=java](http://sonar.opf-labs.org/coding_rules#qprofile=java-verapdf-v0-1-87248%7Cactivation=true%7Clanguages=java)
- TEST FILES**
  - Accepted candidate conformance test files: <https://github.com/veraPDF/veraPDF-corpus>
  - Conformance test files under review by the PDF Association's PDF Validation Technical Working Group: <https://github.com/veraPDF/veraPDF-corpus/tree/staging>
  - Policy test files will be available soon
- OUTREACH**
  - veraPDF website: <http://verapdf.org/>
  - veraPDF blog: <http://verapdf.org/news/>
  - veraPDF Twitter: [https://twitter.com/\\_verapdf](https://twitter.com/_verapdf)
  - Email news list: <http://verapdf.org/subscribe/>

The right column contains the following sections:

- PREFORMA OPEN SOURCE PROJECTS**
  - veraPDF
  - [DPF MANAGER](#)
  - [MEDIACONCH](#)
  - [» View all the successful proposals that participated to the design phase](#)
  - [PREFORMA VAULT \(Access restricted\)](#)
- OTHER RELATED TOOLS**
  - [ARCHIVEMATICA](#)
  - [EXACTLY](#)
  - [JPYLYZER](#)
  - [KOST-VAL](#)
  - [MEDIA FILE CHECKER](#)
  - [XENA](#)

*Fig. 1. veraPDF webpage on the PREFORMA Open Source Portal - snapshot*

The veraPDF consortium directly contributes to ISO standards development by driving a process of identifying and driving resolutions to ambiguities in the specification through the PDF Association's PDF Validation Technical Working Group (TWG), an industry committee that works directly with the ISO standards body (ISO TC 171 SC 2 WG 5) responsible for developing PDF/A. As such, the ISO WG has discussed each of the ambiguities identified by the veraPDF consortium's effort. and

in most cases, approved some interpretation of the text. Although the WG has decided not to amend the existing PDF/A specifications, it is updating the current PDF/A standards development process, most notably in the development of PDF/A-next, the next generation of PDF/A.

A formal industry publication resolving ambiguities in the existing PDF/A specifications will be published in a PDF Association Technical Note in the first half of 2017.

The most significant outcomes of the veraPDF consortium's work include:

- Industry acceptance of the project's purpose, technical, and functional design.
- Successful development of a test-suite and software under the supervision of a large and diverse group of implementers from across the industry, including all industry-leaders in PDF/A development.
- Successful engagement with the ISO community responsible for development of both PDF/A and PDF (ISO 32000) specifications, resulting in significant influence on the future development of both.
- Successful establishment in principle of an industry-supported validation project, thus opening the possibility of similar projects in the future (for example, validation of the next part of PDF/A, or other archival-related specifications such as PDF/E and PDF/UA).
- Successful implementation of a PDF/A conformance checker, a feature extraction engine for further policy checks and a metadata fixer as defined by the tender functional and technical specifications.
- Successful implementation of a greenfield PDF parser, currently the only open-source PDF parser under MPLv2+/GLPv3+ license.

The veraPDF project has established a dialog between the PDF industry and the cultural heritage / digital preservation communities. The heritage community cannot expect, or be expected, to provide expertise across all formats they are responsible for preserving. Active dialog and co-operative projects with the industry sectors where specialist expertise exists is essential to the preservation of digital heritage.

## **2.2 DPF MANAGER**

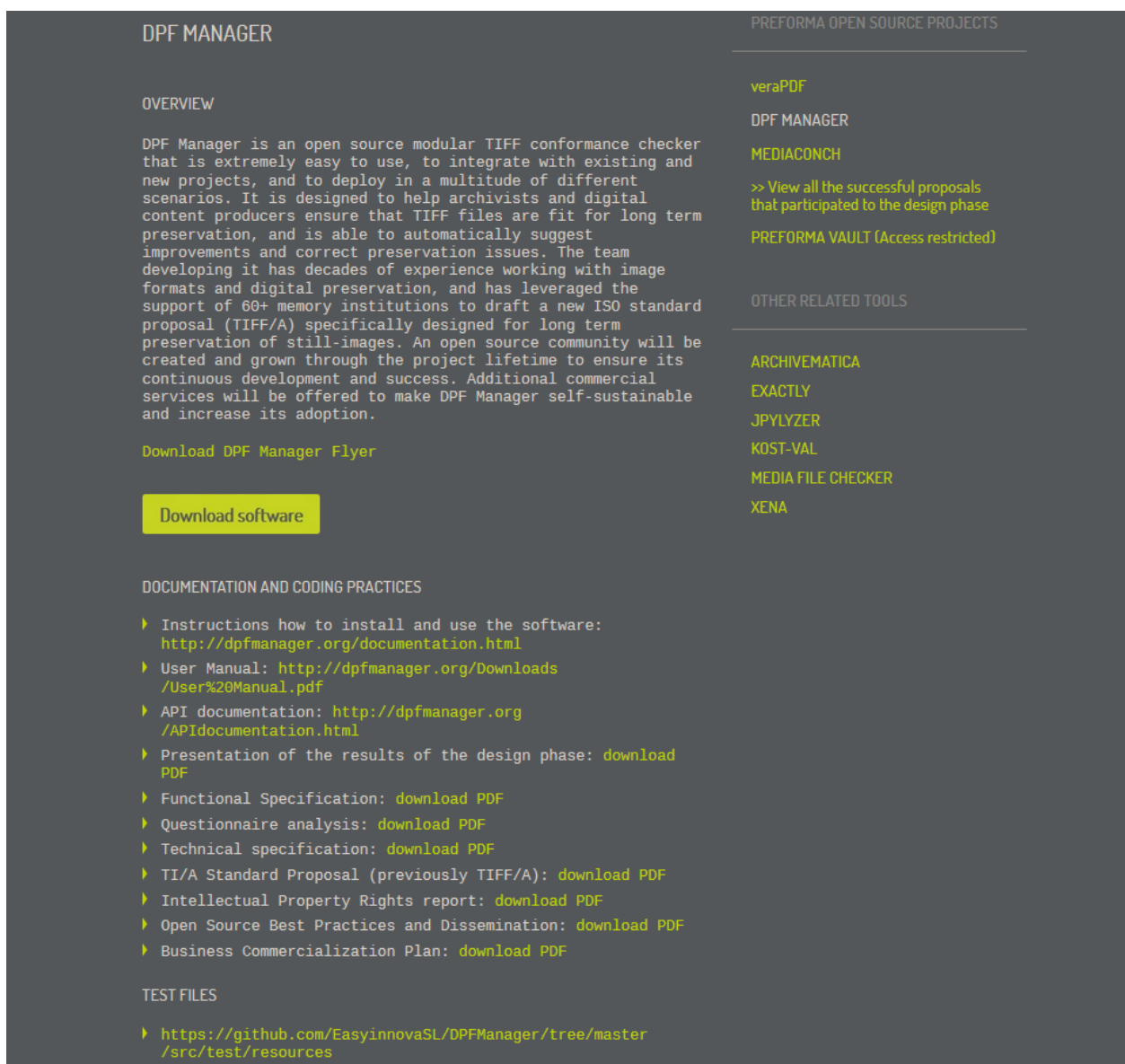
DPF Manager is an open source multiplatform application and framework designed to give memory institutions full control over the conformity tests of TIFF images to be created, migrated and ingested into archives. DPF Manager allows archivists to know the state of their TIFF images in terms of preservation.

DPF Manager main features include:

- TIFF file identification among all the different TIFF standards.
- Validation of the conformance of a TIFF file to a specific format specification. This normative can be defined by an ISO standard or it can be a specific acceptance criterion based on locally defined policy rules.
- Fixing of the TIFF file while preserving the image representation, in order to make it more suitable for long-term preservation.

- User and machine-readable report in different formats, including the data object structure and metadata as well as the validation result.

DPF manager can be run through a Graphical User Interface (GUI), a command-line interface or as a Client-Server application. In addition, an online web conformance checker has been also developed which interacts with a server instance of the DPF Manager.



**DPF MANAGER**

**OVERVIEW**

DPF Manager is an open source modular TIFF conformance checker that is extremely easy to use, to integrate with existing and new projects, and to deploy in a multitude of different scenarios. It is designed to help archivists and digital content producers ensure that TIFF files are fit for long term preservation, and is able to automatically suggest improvements and correct preservation issues. The team developing it has decades of experience working with image formats and digital preservation, and has leveraged the support of 60+ memory institutions to draft a new ISO standard proposal (TIFF/A) specifically designed for long term preservation of still-images. An open source community will be created and grown through the project lifetime to ensure its continuous development and success. Additional commercial services will be offered to make DPF Manager self-sustainable and increase its adoption.

[Download DPF Manager Flyer](#)

[Download software](#)

**DOCUMENTATION AND CODING PRACTICES**

- ▶ Instructions how to install and use the software: <http://dpfmanager.org/documentation.html>
- ▶ User Manual: <http://dpfmanager.org/Downloads/User%20Manual.pdf>
- ▶ API documentation: <http://dpfmanager.org/APIdocumentation.html>
- ▶ Presentation of the results of the design phase: [download PDF](#)
- ▶ Functional Specification: [download PDF](#)
- ▶ Questionnaire analysis: [download PDF](#)
- ▶ Technical specification: [download PDF](#)
- ▶ TI/A Standard Proposal (previously TIFF/A): [download PDF](#)
- ▶ Intellectual Property Rights report: [download PDF](#)
- ▶ Open Source Best Practices and Dissemination: [download PDF](#)
- ▶ Business Commercialization Plan: [download PDF](#)

**TEST FILES**

- ▶ <https://github.com/EasyinnovaSL/DPFManager/tree/master/src/test/resources>

**PREFORMA OPEN SOURCE PROJECTS**

- [veraPDF](#)
- [DPF MANAGER](#)
- [MEDIACONCH](#)
- [» View all the successful proposals that participated to the design phase](#)
- [PREFORMA VAULT \(Access restricted\)](#)
- OTHER RELATED TOOLS**
- [ARCHIVEMATICA](#)
- [EXACTLY](#)
- [JPYLYZER](#)
- [KOST-VAL](#)
- [MEDIA FILE CHECKER](#)
- [XENA](#)

*Fig.2. DPF Manager webpage on the PREFORMA Open Source Portal - snapshot*

The application can communicate with other applications and it is interoperable with other conformance checkers, invoking automatically the appropriate conformance checker to validate the input files. Moreover, it can manage multiple conformance checkers instances for the same file format providing high performance and scalability.

The DPF manager can be used also as a framework, ready to be integrated with other applications or frameworks via API. In order to facilitate the integration, the DPF manager has been included in the Maven package repository.

Finally, another significant outcome of the project is the research carried out to create an ISO-standardized recommendation aimed at ensuring the long-term digital preservation of TIFF files (TI/A – Tagged Image for Archival). The aim of this recommendation is to identify the profile that TIFF files should comply with and the elements and metadata that need to be included to ensure that the files can be reproduced exactly the same way using various software in years to come. The TI/A draft proposal, which has been brought to a pre-final stage, was submitted to the ISO TC171 in October 2016, to start the official standardisation process and it will be discussed in the next working group meeting to be held in Sydney in December 2016.

## **2.3 MEDIACONCH**

MediaConch is complex, extensible suite of open source software that helps information professionals validate audiovisual files and check files against institutional policies. MediaConch consists of an implementation checker, policy checker, reporter, and fixer that targets preservation-level audiovisual files (specifically Matroska, Linear Pulse Code Modulation (LPCM) and FF Video Codec 1 (FFV1)) for use in memory institutions. MediaConch provides detailed and batch level conformance checking for these formats and information for most other file formats.

MediaConch is ready for integration and use in a myriad of different cultural heritage institutions -- whether the institution prefers a command line interface, a graphical user interface, a RESTful endpoint, or a web-based interface. The underlying structure, built by extending the MediaInfo and MediaTrace libraries, produce a new software that fulfills the needs institutions have for quality control and file validation checking beyond which has previously existed for Matroska-wrapped FFV1 and LPCM files. MediaConch can be used at any stage in an OAIS-compliant workflow model.

Through the partnership with Artefactual, MediaConch will automatically gain a large influx of potential users through integration into Archivematica. This expands reach of the software without requiring memory institutions to specifically seek out and download the software, and figure out how to integrate it into their systems. Integration into Archivematica should be complete by end of January 2017, with the next release (planned for late Spring 2017) incorporating MediaConch into the microservices suite.

MediaConch's policy creation and sharing platform allows users to create a policy based on an existing video file report, the ability for users to compare two video files, and the ability for users to add complex logic to their policies. Users are able to create policies and use MediaConch for many formats outside of the ones required by the project (Matroska, FFV1, and LPCM), which is an asset for memory institutions that do not perform normalization and collect a large variety of formats.

The formation of CELLAR within IETF has brought together a strong community of open source developers and preservationists interested in the development of these standards, with significant work contributed by the MediaArea team, the original format designers, and anyone else with stake or interest in the formats. The work on CELLAR and the outreach and community-building work done by the MediaArea team helps solidify trust around these formats as ideal preservation formats.

In addition to outreach, MediaArea has supported the development of supplemental CRC integration and checking to expand on the fixity checks that natively existed in Matroska and FFV1. This includes specification refinements within FFV1, FFV1 decoder and encoder improvements, and bug corrections or bug ticket filing. Matroska and EBML specifications have received major updates that clarify the specification to ensure optimal valid file creation and removes ambiguity in usage.

## MEDIACONCH

### OVERVIEW

MediaConch (CONformance CHECKing for audiovisual files) is an extensible, open source software project consisting of an implementation checker, policy checker, reporter and fixer that targets preservation-level audiovisual files (specifically Matroska, Linear Pulse Code Modulation (LPCM) and FF Video Codec 1 (FFV1)) for use in memory institutions, providing detailed and batch-level conformance checking via an adaptable and flexible application program interface accessible by the command line, a graphical user interface, or a web-based shell. MediaConch is currently being developed by the MediaArea team, notable for the creation of open source media checker software, MediaInfo. Furthermore, the MediaArea team is dedicated to the further development of the standardization of the Matroska and FFV1 formats to ensure their longevity as a recommended digital preservation file format.

[Download MediaConch Flyer](#)

[Download software](#)

### DOCUMENTATION AND CODING PRACTICES

- Instructions how to install and use the software: <https://mediaarea.net/MediaConch/documentation.html>
- Test Demonstrations: <https://mediaarea.net/MediaConch/demo.html>
- EBML Specification: <https://github.com/MediaArea/ebml-specification/commits/master>
- FFV1 Specification: <https://github.com/FFmpeg/FFV1/commits/master>
- Presentation of the results of the design phase: [download PDF](#)
- Functional and technical specification: [download PDF](#)
- Conformance Check Registry & Expression Snapshot: [download PDF](#)
- Discovery interviews: [download PDF](#)
- Discovery questionnaire: [download PDF](#)
- Standardization exercises: [download PDF](#)

### TEST FILES

- Test files directory: [https://github.com/MediaArea/MediaConch\\_SampleFiles](https://github.com/MediaArea/MediaConch_SampleFiles)
- Demo with files: <https://github.com/MediaArea/MediaConch/tree/master/Demo>

## PREFORMA OPEN SOURCE PROJECTS

[veraPDF](#)

[DPF MANAGER](#)

[MEDIACONCH](#)

[>> View all the successful proposals that participated to the design phase](#)

[PREFORMA VAULT \(Access restricted\)](#)

## OTHER RELATED TOOLS

[ARCHIVEMATICA](#)

[EXACTLY](#)

[JPYLYZER](#)

[KOST-VAL](#)

[MEDIA FILE CHECKER](#)

[XENA](#)

Fig.3. MediaConch webpage on the PREFORMA Open Source Portal - snapshot

### 3. OPEN SOURCE COMMUNITY DEVELOPMENT

One of the main challenges of PREFORMA is the establishment of an open source community of people interested in maintaining and improving the tools developed by the project in the long-term. This network should involve users and developers outside the PREFORMA consortium to contribute to the development phase and it should become, at the end of the project, a sustainable open user community surrounding the applications provided by PREFORMA. The main task of this community is to ensure the long-term maintenance of the tools, encourage both future use and future development and contribute to the standardisation process.

The main stakeholders involved include:

- **developers** controlling the production of preservation files and contributing code to the PREFORMA tools, thus aiming at improving the software according to the needs of the user community.
- **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.

The three suppliers selected by the PREFORMA partners to work on the design and developing of the prototypes worked a lot towards the establishment of such a community. Even if it is still a bit early to assess the progress of this work and to measure the impact generated, to track the evolution of a successful ecosystem around the tools procured, in the following paragraphs report a first analysis of what has been done and of the level of activity generated so far in the community.

#### 3.1 VERAPDF

The first phase of community development began in October 2014 with the establishment of the PDF Association's PDF Validation TWG in order to review, discuss and accept test suite files and software functionality. Since that time, substantial progress has been made in winning the involvement of industry-leading software developers such as Adobe Systems, callas software, Datalogics, PDFTron, PDF Tools, PDFlib, Foxit, Adlib Software and others.

Since beginning software development in April 2015, the veraPDF consortium has focused on establishing the technical foundations for a healthy open source community to work on integration of veraPDF. Reasons for this focus included:

- without these foundations, dealing with external contributions is labour intensive and error prone;
- this allows to design a modular and scalable architecture and a working and tested code base which ensures the high quality of the software;
- the infrastructure which is in place allows to make regular software releases, an effective way of raising awareness.



At the same time, the veraPDF consortium has been involved in a variety of dissemination activities to help generate interest, a precursor to community. In particular, the veraPDF consortium has organised and participated in webinars and hands on events for developers and users while continuing to make regular software releases. In addition, the veraPDF consortium has introduced a dedicated documentation site: <http://docs.verapdf.org/> driven by GitHub: <https://github.com/veraPDF/veraPDF.github.io>. This means that external contributions to the documentation will be made via GitHub pull requests, in the same manner as contributions to the source code.

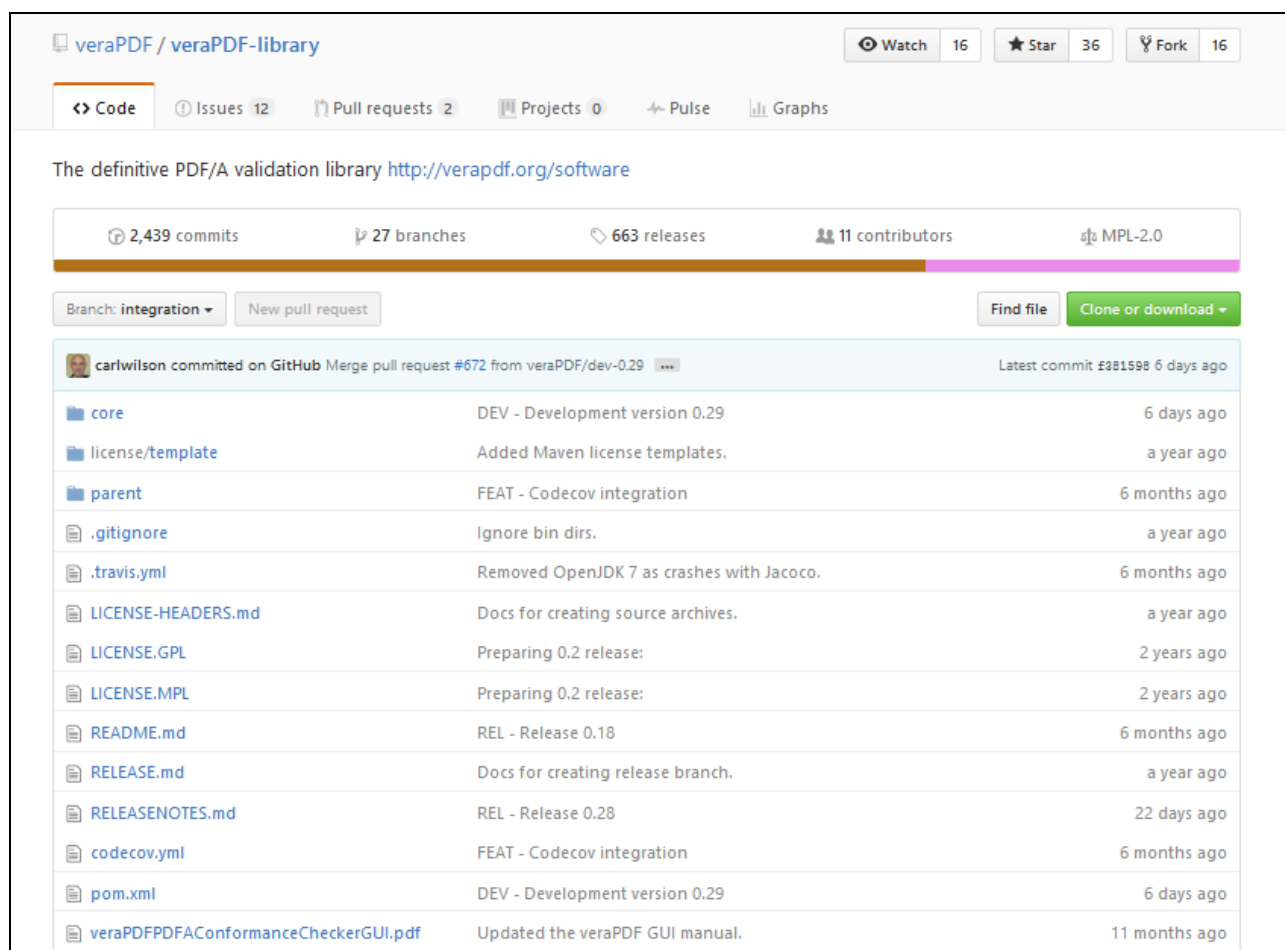
Finally, the veraPDF consortium has tried to engage with and encourage external contributors to the project. The best example to date is an external pull request providing reporting capability to files using command line options: <https://github.com/veraPDF/veraPDFApps/pull/32>.

In this framework, cooperation with the following organisations has been established:

- State and National Library Denmark.
- European Publication Office, one of the three largest collections of PDF document in the world.
- KEEP Solutions (a consortium partner of veraPDF), which has integrated veraPDF into RODA, their software repository product.
- Artefactual Systems (a member of the Open Preservation Foundation), which are planning to integrate veraPDF into their product Archivematica.
- Logius, the digital government service of the Netherlands Ministry of the Interior and Kingdom Relations, which have commissioned a piece of work to develop software based on veraPDF and other open source software that will crawl a domain, identify documents of interest and validate PDF/As against ISO standard 19005 parts 1,2 and 3.

In terms of establishing industry support for veraPDF, many of the most significant discussions in the PDF software developer community occurred between the 58 subject matter experts (SMEs) participating in the PDF Validation TWG via its mailing-list (520 posts) and during the 27 TWG meetings held since December 2014. The results of these discussions were, in many cases, integrated directly into the veraPDF software by the veraPDF consortium's lead developer (Dual Lab), and thus did not involve code contributions from external institutions.

In terms of contributions of non-consortium members to the evaluation of the tool, a total of 246 external institutions downloaded veraPDF more than once and may be considered as evaluators of the software. Some of them (e.g. Isartor and Bavaria) provided test corpora, others tested veraPDF on their own data that couldn't be shared due to IP restrictions. Members of the TWG have also tested veraPDF on their own test sets, but again these couldn't be shared with the consortium. Direct test file contributions are in the 10s while thousands of external files have been used in testing. Finally, there have been very few policy contributions during development but there has been a growing interest among the digital preservation community as the functionality of veraPDF has been highlighted during outreach events.



The screenshot shows the GitHub repository for veraPDF. At the top, it says 'veraPDF / veraPDF-library' with 16 watches, 36 stars, and 16 forks. Below this, there are tabs for Code, Issues (12), Pull requests (2), Projects (0), Pulse, and Graphs. The main heading is 'The definitive PDF/A validation library' with a link to <http://verapdf.org/software>. A progress bar shows 2,439 commits, 27 branches, 663 releases, 11 contributors, and MPL-2.0 license. Below the progress bar, there are buttons for 'Branch: integration', 'New pull request', 'Find file', and 'Clone or download'. A table lists recent commits and files:

File	Description	Time
core	DEV - Development version 0.29	6 days ago
license/template	Added Maven license templates.	a year ago
parent	FEAT - Codecov integration	6 months ago
.gitignore	Ignore bin dirs.	a year ago
.travis.yml	Removed OpenJDK 7 as crashes with Jacoco.	6 months ago
LICENSE-HEADERS.md	Docs for creating source archives.	a year ago
LICENSE.GPL	Preparing 0.2 release:	2 years ago
LICENSE.MPL	Preparing 0.2 release:	2 years ago
README.md	REL - Release 0.18	6 months ago
RELEASE.md	Docs for creating release branch.	a year ago
RELEASENOTES.md	REL - Release 0.28	22 days ago
codecov.yml	FEAT - Codecov integration	6 months ago
pom.xml	DEV - Development version 0.29	6 days ago
veraPDFPDFAConformanceCheckerGUI.pdf	Updated the veraPDF GUI manual.	11 months ago

Fig.4. veraPDF-library GitHub repository - snapshot

In terms of contributions of non-consortium members to the development of the tool, there have been 4 external developers who've contributed code to the veraPDF projects plus 10 external contributions of minor fixes or command line application enhancements. 18 individuals external to veraPDF have submitted 30 issues in total and 5 institutions are actively integrating veraPDF and have asked for help. Many more have expressed and intention to do so or have said they are working on it but there is no evidence as to whether they have or not.

As of January 5, 2017, Google finds 4,950 references to "veraPDF" across the web.

## 3.2 DPF MANAGER

During the prototyping phase, the strategy followed by EasyInnova to build up a community around the DPF Manager was to create first a community around the TI/A initiative (Tagged Image for Archival), which demonstrated to attract a strong interest around it, and then to announce that the DPF Manager will be the first tool that can validate this new recommendation.

TIFF is widely used by memory institutions to preserve their digital images, however, the specification of TIFF is complex and some of its features are proprietary and therefore not suitable for long-term archival purposes. As explained in Section 2.2, the TI/A standard initiative is a group



of experts focusing on the creation of a new ISO recommendation to optimise the TIFF format definition for archival purposes to ensure the long-term preservation of the digital images.

This initiative is being carried out thanks to the valuable collaboration of the Swiss Coordination Centre for the Long-Term Preservation of Electronic Documents (KOST-CECO - <http://www.kost-ceco.ch/>) and some of the biggest and most important Swiss Archives.

The most important image researchers are following closely the progress of the TI/A recommendation and waiting for the final technical specification that is being submitted to the ISO TC-171 technical committee.

The TI/A community is being build up around three online channels (and many offline communications), the TI/A website [ti-a.org](http://ti-a.org), the TI/A twitter account [twitter.com/TI\\_A\\_Standard](https://twitter.com/TI_A_Standard) and the TI/A Intranet for the involved experts [intranet.ti-a.org](http://intranet.ti-a.org).

So far, 80 experts registered in the Intranet from 17 different countries. The TI/A website has received 6.500 visits of more than 4.500 unique visitors in 18 months, with an increment around 25% after the publication of the 1st draft. The twitter account has over 410 followers and 850 tweets have been published related to the TI/A recommendation and digital image preservation with around 170.000 impressions. The trend of the growth in the number of the followers is around 20% per month.

In parallel, Easy Innova communication campaign around the progress in the development of the DPF Manager tool has been oriented to hook early adopters and get feedback from them.

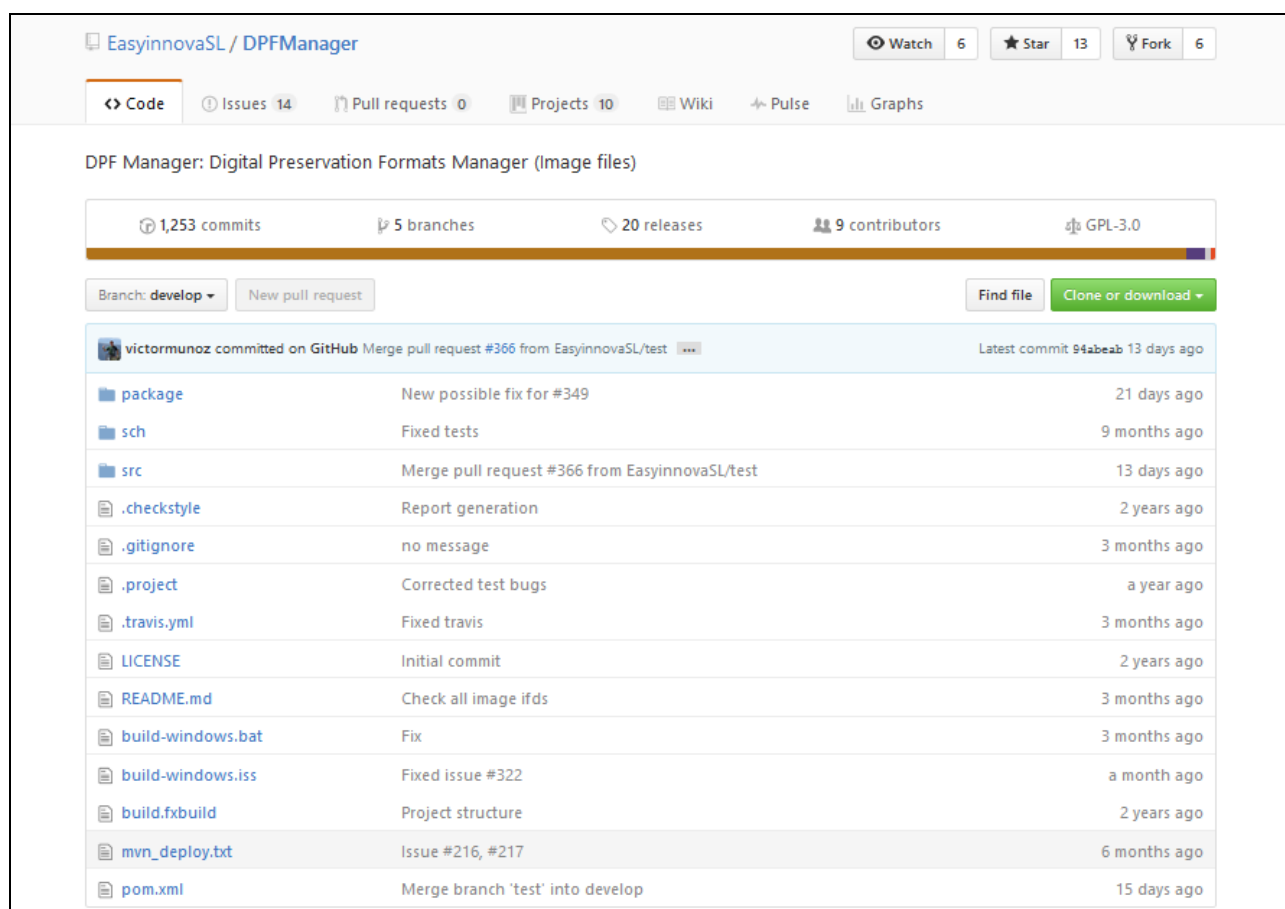
Since October 2015, the DPF Manager website received 10.000 visits (35% of increment in the last 3 months) from more than 4.200 unique visitors, the twitter account has over 280 followers from different countries in Europe (increasing 100% only in the last month) and @dpfmanager tweets have a total amount of 96.889 impressions (+50% in the last 3 months).

From the DPF manager website it is possible to download the software for different Operating Systems with the related documentation, join the developers community to contribute with your code, report bugs using the issue tracker, ask questions to the developers and exchange opinions with other people who tested the tool through a Q&A forum and an IRC channel.

So far, cooperation with the following organisations has been established:

- Digital Humanities Lab at the University of Basel, Switzerland. The Digital Humanities Lab is the TIFF format and digital cultural heritage experts in the DPF Manager team. They have been doing research in long term archival of still-images and video since 1995. They are leading the TI/A Standard initiative and provided access to numerous memory institutions in Switzerland.
- KOST, Switzerland. The Swiss Coordination Centre for the Long-Term Preservation of Electronic Documents (KOST-CECO) is very interested in the TI/A Initiative. KOST contributed to the analysis of almost 4 Million TIFF files in memory institutions' archives.
- Adobe Systems Inc, USA. Adobe is very influent and has a lot of power in the ISO technical committees. Having Adobe on board of the TI/A initiative was crucial and after several discussions by email and a face-to-face meeting in their headquarters (San Jose, USA) the DPF Manager team finally got their support to follow up with the standardization process.

- AENOR, Spain. EasyInnova joined AENOR (Spanish standards organisation, ISO member) and became full members of the AEN/CTN50 committee.
- Hewlett-Packard CDS, Spain, which provided a letter of intent stating their interest on a potential commercial partnership to distribute & integrate DPF Manager into existing document and digital asset management applications.



EasyinnovaSL / DPFManager

Watch 6 Star 13 Fork 6

Code Issues 14 Pull requests 0 Projects 10 Wiki Pulse Graphs

DPF Manager: Digital Preservation Formats Manager (Image files)

1,253 commits 5 branches 20 releases 9 contributors GPL-3.0

Branch: develop New pull request Find file Clone or download

victormunoz committed on GitHub Merge pull request #366 from EasyinnovaSL/test Latest commit 94abeab 13 days ago

package	New possible fix for #349	21 days ago
sch	Fixed tests	9 months ago
src	Merge pull request #366 from EasyinnovaSL/test	13 days ago
.checkstyle	Report generation	2 years ago
.gitignore	no message	3 months ago
.project	Corrected test bugs	a year ago
.travis.yml	Fixed travis	3 months ago
LICENSE	Initial commit	2 years ago
README.md	Check all image ifds	3 months ago
build-windows.bat	Fix	3 months ago
build-windows.iss	Fixed issue #322	a month ago
build.fxbuild	Project structure	2 years ago
mvn_deploy.txt	Issue #216, #217	6 months ago
pom.xml	Merge branch 'test' into develop	15 days ago

*Fig.5. DPF Manager GitHub repository - snapshot*

Since the release of the first prototype, a large list of early adopters used DPF Manager, most of them are anonymous, but some others are known, usually memory institutions but also Universities and companies. All these early adopters used DPF Manager to check their TIFF digital assets, allowing EasyInnova to validate the operation of the tool and identifying bugs to be solved. In particular, from the identified early adopters, more than 4 Million TIFF files have been analysed using DPF Manager. Of course, not all the early adopters sent feedback about their tests, thus, the real number of analysed files is notably higher. When an error appeared, early adopters notified their issues by e-mail, twitter o directly in Github. These issues sometimes attached test files, really useful to check TIFF variations. In particular, more than 40 test files from non-partner early adopters have been received.

It is important to mention that DPF Manager includes functionality where memory institutions can share with EasyInnova the results of their checks, a very useful tool to analyse how the tool is working. A total amount of 14.742 reports of TIFF images analysed were received, which helped identifying TIFF private tags and common errors done by the TIFF writers.

The dissemination campaign to get early adopters provided EasyInnova with interested memory institutions with particular needs. For example, the case of the National Archives of Denmark that requested us some functionalities (i.e. the detection of blank pages) to include in the policy checker of the DPF Manager. Other non-partner institutions provided EasyInnova with more than 10 new policies to be considered in the policy checker.

Now the DPF Manager is developed and maintained by 9 internal contributors that have submitted 1.253 commits in 20 releases. During the project, 250 issues have been submitted, 61 of them from external contributors.

### **3.3 MEDIACONCH**

Feedback and community involvement has been most prominent in MediaArea's standardization efforts for Matroska and FFV1. This aspect of the project involves several distinct communities, such as FFmpeg, Matroska, and the IETF.

In-person meetings such as those held at the IETF, FOSDEM, and VDD has supplied significant initiatives to the projects. At the same time, the recent establishment of the CELLAR (an IETF working group focused on FFV1 and Matroska) provided a new venue for community work and collaboration focused on these selected open media formats. Within this group, with a mailing list of 87 members and growing, the standardization conversation can happen between designers, developers, and archivists.

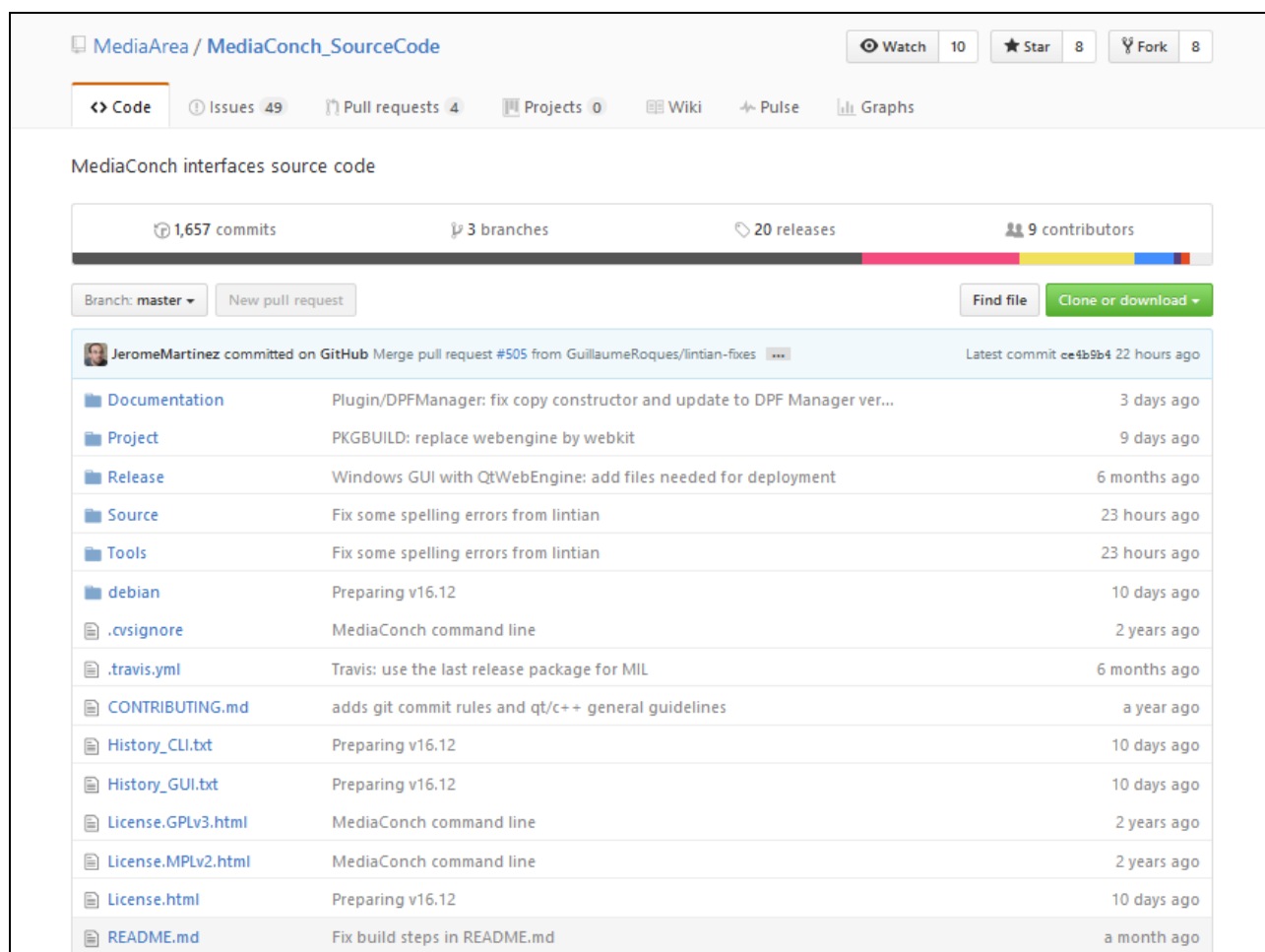
Presently MediaConch is reaching out to archives and vendors that already use Matroska and/or FFV1 in order to document existing practices and encourage shared best practices. With the final release of the prototyping phase, MediaArea aims to expand the user base, and encourage further testing and participation, leveraging on MediaArea's existing software packages, MediaInfo and MediaTrace. MediaInfo already has a very large user base among audiovisual archivists. MediaTrace is used by MoMA (NYC).

MediaArea is working closely with the FFV1 and Matroska format designers to ensure standardization in line with the formats vision. MediaArea is also working closely with other preservationists, collecting user needs and feedback, to apply to the forthcoming standardized specifications. In addition, MediaArea has interviewed users of MediaConch to gather feedback and implement improved features within the software through filing issues on the code's Github page.

MediaArea encourages community members to get involved through 1-on-1 coaching and instructions, speaking at conferences, and holding workshops about MediaConch and, more generally, about file validation and conformance checking. In addition, both MediaConch and Archivematica (who has integrated MediaConch) led a workshop in training at the Association of Moving Image Archivists to show archivists how to participate, contribute to, and use these open source projects.

So far, cooperation with the following organisations has been established:

- **Artefactual Systems (Archivematica).** The collaboration with Archivematica has been beneficial to both projects and has utilized previously unforeseen aspects of the software. Archivematica integrated both the implementation checker and the policy checker as a method to identify media file qualities and send the media to corresponding micro services according to their preservation requirements. At the AMIA conference in early November, both Archivematica and MediaConch were presented in a workshop along with other open source tools. Artefactual also hosted a poster detailing the integration of MediaConch into Archivematica at this conference.
- **Tate Museum.** During the project MediaArea found an opportunity to collaborate with Pericles, an FP7 project. Within this collaboration, MediaConch developers were hired to present two days of workshops related particularly to the reporting and policy checking aspects of MediaConch. The collaboration continued to develop policies for particular conservation needs at the Tate Museum. This work inspired many extensions to the policy checker such as using MediaTrace data in addition to MediaInfo for policy design as well as adding the ability to create policies upon the shared or distinct characteristics of two different files (such as a preservation and access rendition of the same video).
- **Indiana University.** IU, working with Memnon as a digitization vendor, selected FFV1 and Matroska as preservation formats for a large digitization project. The work of PREFORMA, MediaConch, and the IETF CELLAR working group were contributing factors in this decision. MediaArea has supported IU in their use of MediaConch tools in policy design to help verify and assess the Matroska and FFV1 work as their project is underway.
- **Library of Congress.** MediaArea has informally collaborated with the Library of Congress which leads a project referred to as AS-07 to develop a specification for the archival use of JPEG2000 and MXF. This collaboration has been a helpful guidance to each project as there are obviously few audiovisual projects that so closely integrate standardization procedures with preservation applications. Kate Murray of the Library of Congress presented at the No Time to Wait symposium about the lessons learned from their project. At the IASA conference, presentations of the MXF/JPEG2000 and Matroska/FFV1 projects were presented back-to-back. Also members from each project have participated and contributed in both projects, as such projects benefit from working better together.
- **VIAA / PACKED.** Working with VIAA with the support of PACKED, MediaArea is supporting VIAA's integration of MediaConch. This involves extending MediaConch via its plugin structure to support transcoding as well as policy application and implementation checking.
- **New York Public Library.** New York Public Library has started to integrate MediaConch into their audiovisual preservation and digitization unit's quality control workflow. NYPL has been using MediaConch to check and validate TIFF files in their Digital Imaging Unit.
- **CUNY-TV.** MediaConch usage and policy development work has been done within CUNY-TV's preservation department related to testing QuickTime and audio files.
- **California Audiovisual Preservation Project.** CAVPP has used MediaConch policies to test QuickTime files in their internal workflows.



*Fig.6. MediaConch GitHub repository - snapshot*

MediaConch's Source Code repository has had 18 contributors, including people who opened issues, provided feedback to the software and committed code. Of these, nine are from or related to the MediaArea team. Two are from those outside of the project but within related partner institutions (verwinv, beyo-ra). Seven are from external institutions or independent contributors (pjotrek-b, krm, retokromer, kieranjol, mistydemeo, vascom, dinahhandel).

MediaConch's main (MediaArea/MediaConch) repository has had 22 contributors. Of these, ten are from or related to the MediaArea team (including Artefactual). One is from outside of the project but within related partner institutions (Elorrain). Eleven are from external institutions or independent contributors (pjotrek-b, kimec, kieranjol, t-rapp, joshuatl, victormunoz, PatriciaFalcao, retokromer, vascom, genfhk, clacinak).

The MediaArea team has helped people within institutions create policies (Artefactual, New York Public Library, Tate Museum, Louisiana Public Broadcasting) but only one contributor outside of the MediaArea team has contributed to the public policy resource available via MediaAreaOnline.

As opposed to PDF/A and TIFF, the usage of Matroska and FFV1 is relatively new with most Matroska and FFV1 held by archives being produced by FFmpeg libraries. The MediaConch team worked with original video submissions at the Internet Archive to identify approximately 100,000

unique Matroska uploads which covered a wide variety of Matroska writers, found some obscure Matroska issues, and illuminated the diversity of implementations of Matroska. The team defined some errors but did not identify any occurrence of those so also produced synthetic files of errors to submit to the official Matroska repository. The implementations of FFV1 were more consistent with only two encoders (FFmpeg and libav) so it was more easily possible to produce test files from various snapshots of these encoders through combinations of supported arguments. MediaConch also worked with non-partner institutions, including many of those listed above. Often it was not necessary to exchange the test files themselves (as they are significantly large), but exchanged MediaTrace reports, produced by MediaConch, in order to fully interpret the structure of example files.



## 4. IMPACT OF PREFORMA ON THE STAKEHOLDERS

### 4.1 IMPACT ON THE MEMORY INSTITUTIONS

The main expected impact of PREFORMA is to reduce curation and preservation costs, to improve curation and preservation capacity, to enhance competences in public organisations, including small archives, and to contribute to their independence from individual vendors.

First of all PREFORMA procures an open source software component that can be easily integrated in the existing curation and preservation workflows, enabling memory institutions and other interested parties to make a better assessment of the digital objects they acquire and facilitating the improvement of the preservation infrastructure they already have in place. This will not only improve the preservation efforts of the particular institution, but should also ensure and demonstrate the feasibility of the integration of the tools developed by PREFORMA in existing curation infrastructures.

Secondly, by organizing a joint procurement for a validation platform that facilitates multiple file formats and implementations, memory institutions obtain a flexible and more sustainable validation platform that reduces the future cost of validating new types of digital objects.

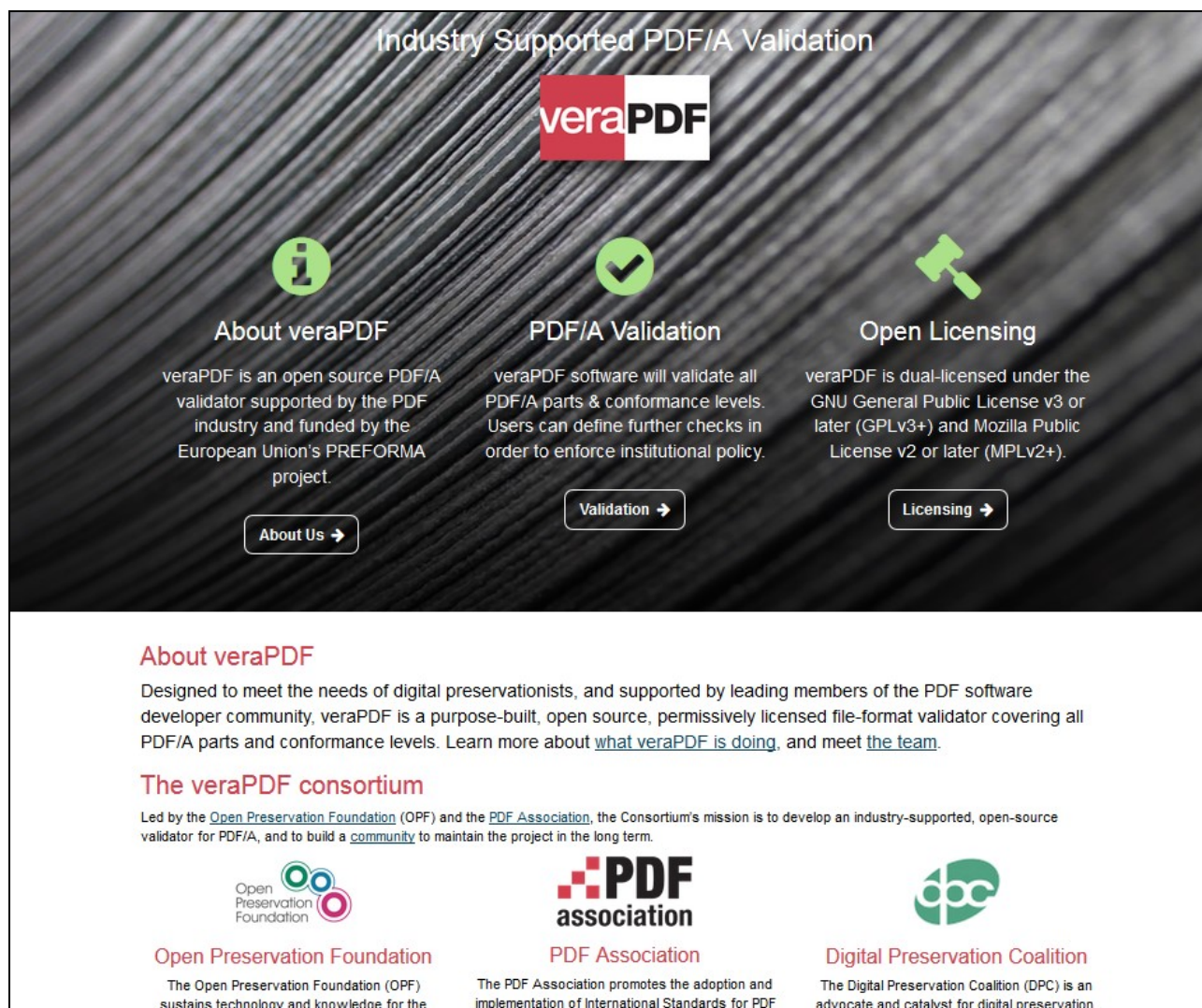
Furthermore, through the user and developer community built by PREFORMA, memory institutions may benefit from research & development procured by other institutions that use the same platform.

Involvement in requirements for policy checking, and publishing the requirements and test files for policy checking provides a basis for objective reasoning about different approaches to collection management. This provides the opportunity to consolidate policy within memory institutions and deliver efficiencies to the sector and to depositing organisations such as publishers and other producers of PDF/A, TIFF and AV files. This is only possible once there is an objective frame of reference such as that provided by the PREFORMA project and once policy decisions can be shared and compared.

By sharing requirements and leveraging a better understanding of the file format specifications, policy decisions will be documented and best practices will be published showing objective reasoning across different approaches on the basis of evidence from testing across multiple authoritative and representative corpora.

The active use of a tool such as a format validator depends strongly on its stability and feature completeness. For this reason, during the prototyping phase the work of the suppliers focused most on the development of the prototypes rather than on installing them into working environments. However, during the dissemination campaign, a strong interest in the PREFORMA conformance checking tools has been raised from both the digital preservation and the industry side. A number of archiving system vendors have also expressed an interest in including the tools in their offering.

In the case of **PDF/A**, an example is 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.



*Fig.7. veraPDF website - snapshot*

The Digital Preservation Coalition (DPC) conducted several testing runs by applying the tool to a real-life test case to identify remaining stability issues and other bugs relating to the application of the tool to a problem space. DPC also approached volunteers from a range of heritage and research organisations to solicit feedback on the functionality, stability, and usefulness of veraPDF in a digital preservation context. In addition to asking for feedback on how well the tool works generally, they asked for comment on whether the tool could be integrated into existing institutional workflows. They also directly requested suggestions for improved functionality based on comparison to other validators.

DPC approached 14 organisations and received feedback from 8. Volunteers from round 1 provided feedback on changes and improvements to the tool, while volunteers from round 2 provided insight into their experiences using the tool.



The feedback that veraPDF and PREFORMA 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.

In the case of **TIFF**, the low technical knowledge about TIFF format makes memory institutions often not aware about the lack of conformance of their TIFF files to the standard specifications and about their possible unsuitability for long-term preservation. Thanks to the work carried out with the TI/A initiative, aimed at building a set of ISO-certified recommendations for memory institutions to guarantee the long-term preservability of their TIFF files, memory institutions are now increasingly understanding that they have to validate the conformance of their files to the standard specifications and to the upcoming TI/A ISO recommendation.

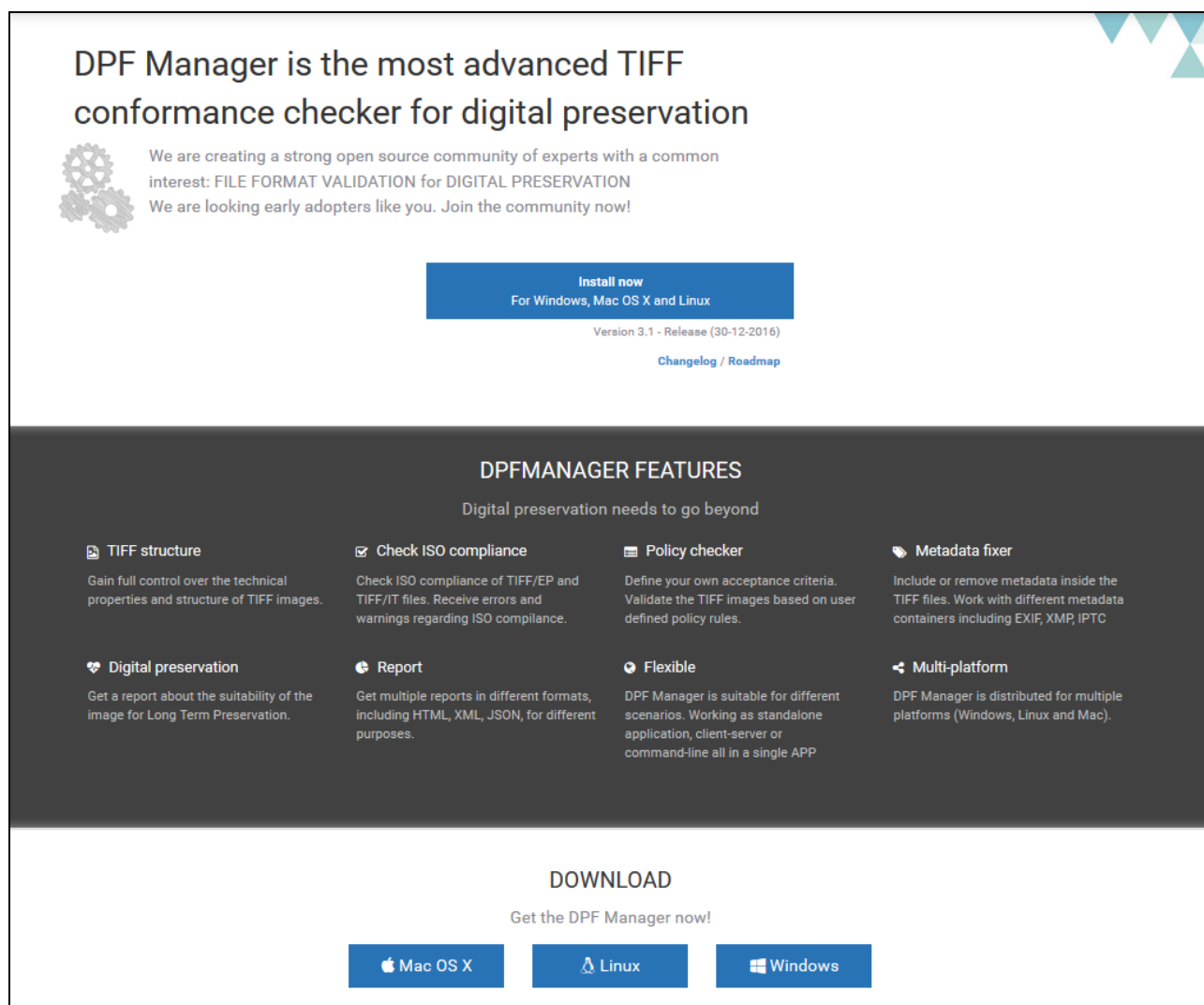
The known early adopters of the DPF Manager already counts several members: National Archives of Sweden, Aquaforest Limited, eASys, Oregon State University Libraries, bj institute, MIT Libraries, MoMu, Hochschule der Künste Bern, Royal Museums of Fine Art of Belgium, Technical University of Viena, National Archives of Denmark, City Council of Stockholm Archive, University of Pittsburgh, University of Illinois, Dartmouth College, New York Public Library, Swiss Federal Archives and the cantonal archives of Basel City and Saint Gallen.

A lot of feedback has been received from these early adopters. Most of the times early adopters only sent feedback when something was not working as desired or when they requested some kind of new feature to be developed. Only a very few times feedback just to congratulate or acknowledge was received.

These early adopters allowed to receive 14.742 reports of files analysed, which helped the DPF Manager team to discover 18 private tags and 16 typical errors in the baseline. The increment of use by early adopters in the last months is notorious, as they analysed more or less the same number of files in 2 months than in the whole year before.

In particular, the DPF Manager has been used intensively in two cases:

- Packed validated the conformance of around 40 thousand TIFF files of scanned paintings for La Fundació Tàpies (<http://www.fundaciotapies.org/>).
- The University of Basel analysed almost 4 Million TIFF files from 3 big memory institutions in Switzerland in order to understand which variants and tags have been used in the past to create TIFF files.



**DPF Manager is the most advanced TIFF conformance checker for digital preservation**

We are creating a strong open source community of experts with a common interest: **FILE FORMAT VALIDATION for DIGITAL PRESERVATION**  
We are looking early adopters like you. Join the community now!

**Install now**  
For Windows, Mac OS X and Linux

Version 3.1 - Release (30-12-2016)  
[Changelog](#) / [Roadmap](#)

### DPFMANAGER FEATURES

Digital preservation needs to go beyond

- TIFF structure**  
Gain full control over the technical properties and structure of TIFF images.
- Check ISO compliance**  
Check ISO compliance of TIFF/EP and TIFF/IT files. Receive errors and warnings regarding ISO compliance.
- Policy checker**  
Define your own acceptance criteria. Validate the TIFF images based on user defined policy rules.
- Metadata fixer**  
Include or remove metadata inside the TIFF files. Work with different metadata containers including EXIF, XMP, IPTC
- Digital preservation**  
Get a report about the suitability of the image for Long Term Preservation.
- Report**  
Get multiple reports in different formats, including HTML, XML, JSON, for different purposes.
- Flexible**  
DPF Manager is suitable for different scenarios. Working as standalone application, client-server or command-line all in a single APP
- Multi-platform**  
DPF Manager is distributed for multiple platforms (Windows, Linux and Mac).

### DOWNLOAD

Get the DPF Manager now!

[Mac OS X](#) [Linux](#) [Windows](#)

*Fig.8. DPF Manager website - snapshot*

Finally, in the case of **AV files**, the No Time To Wait! Symposium organised in Berlin from July 18<sup>th</sup> 20<sup>th</sup> proved to be a very good opportunity for archivists and open source software developers to collaborate and share feedback. Archivists appreciated the opportunity to engage more closely with Matroska and FFV1 and welcomed the opportunity to interact with developers about their needs and workflow specifications. Direct access to the specification authors helped the international archival community better understand the tools and have meaningful twoway interactions.

MediaArea's partnership with Artefactual, Tate Museum, VIAA, New York Public Library, CUNY-TV and California Audiovisual Preservation Project is a first way to measure the impact of MediaConch on the user community.

Archivematica is a popular framework for OAIS-compliant digital preservation with a robust user community. Initial integration of MediaConch is available in the latest Archivematica release, which

has been promoted during Artefactual's Archivemata Camp and a webinar (Video: [https://www.youtube.com/watch?v=ZTG9nlp\\_4oA](https://www.youtube.com/watch?v=ZTG9nlp_4oA)) specifically on the topic.

Through the cooperation with VIAA and PACKED, MediaConch will be used to transcode archival videos into Matroska/FFV1, use the policy checker to assure that the source file and resulting Matroska/FFV1 share significant characteristics, and run the implementation checker on the result.

Through the cooperation with the Tate Museum in the context of their Pericles project, the MediaConch policy checker will be extended to include methods to develop policies that compare two files to each other (for example to compare a source file and result after transcoding to ensure that only certain characteristics are adjusted). This collaboration also led to more advanced policy checking features in order to test certain files for specific hardware support.

New York Public Library has started to integrate MediaConch into their audiovisual preservation and digitization unit's quality control workflow to check and validate TIFF files in their Digital Imaging Unit.

Finally, MediaConch usage and policy development work has been done in cooperation with CUNY-TV and California Audiovisual Preservation Project, which used MediaConch policies to test QuickTime and audio files in their internal workflows.

Functionally, MediaConch can already be expanded to support file formats beyond Matroska, FFV1, and LPCM (and PDF/TIFF via integration of work from the other suppliers). There is potential for MediaConch to become the conformance checking software for *any* audiovisual formats after the completion of this project, not just limited to Matroska and FFV1. MediaConch's local policy creation feature can already be extended out to any format that MediaInfo supports. There has been institutional interest in supporting development within MediaConch to extend conformance checking to support FLAC and other video formats.

Now that the prototyping phase of PREFORMA has been successfully completed, the PREFORMA partners and suppliers will focus more on trying to deploy the prototypes at memory institutions sites in order to evaluate how they work in real environments and to measure the impact of the tools for the memory institutions.

In order to facilitate this, a series of training sessions and hands-on workshops has been planned for the last year of the project in several European Countries (Belgium, Estonia, Germany, Italy, Latvia, Spain, Sweden, and The Netherlands), where people working in memory institutions can bring their files and analyse them with the PREFORMA tools with the help of the PREFORMA partners.



*Fig.9. MediaConch website - snapshot*

## 4.2 IMPACT ON COMPANIES AND TECHNOLOGY PROVIDERS

The need for a validator that implements interpretations of the ISO-standards in such a way that it sets the de facto standard in the European region concerning memory institutions and governmental agencies is in big demand. This comes from a demand of archiving electronic collections that are fast growing in size. Many institutions are beginning to implement e-archive or broader digital asset management systems to handle collections of information objects. The market for consultancy services and program add-ons to a validator product and interpreter of the standard is large and growing in the European area.

The software developed in PREFORMA will have a substantial impact on the technology vendors:

- saving time and money via reduced support costs due to collisions between claims of the standard specifications and variations in quality;
- enhancing efficiency in development due to the ability to address issues systematically and conclusively;

- opening the door to new software implementations leveraging a file-format validated environment;
- driving rapid marketplace adoption by the vast majority of software producers supporting the standard file formats addressed in PREFORMA;

Industry members will be highly motivated to check the progress and features of the implementation checker against their own interpretation of the standard.

As outlined in the preliminary market considerations and business opportunities analysis included in deliverable D2.1, PREFORMA advocates a number of business opportunities for companies and technology providers, all in line with the Open Source software model adopted in PREFORMA:

- Combinations with other software offerings, i.e. using the open source software developed in PREFORMA as a complement to commercial software products in their portfolio.
- Selling optional proprietary extensions, i.e. developing extensions to the open source software and sell them using a closed license.
- Selling professional services, i.e. providing services for deploying the open source software at memory institutions, e.g. consulting, customization, technical support, training, etc.

As a preliminary impact analysis, we surveyed the organisations that are working as suppliers in PREFORMA to understand which main benefits have been gained so far, as these organisations are those that are in the best position to exploit the results of the project.

### **Open Preservation Foundation**

The OPF has formed strong relationships with the veraPDF consortium partners, raising their visibility with memory institutions, industry and service providers. In building the veraPDF community it was seen increased support for open standards and open source software in the digital preservation community. OPF intends to build on these connections to provide reliable open source file format validation for its members and the wider digital preservation community.

### **PDF Association**

The veraPDF consortium has advanced the degree of interoperability in PDF creation and management technology, a benefit to all members of the PDF Association. We believe that non-conforming implementations will tend to become conforming implementations over time, resulting in higher-quality PDF files that are more readily and reliably processed. As reliability is the core value proposition for PDF technology, this development is of material benefit to not only the member companies of the PDF Association, but to all customers and users as well.

### **Digital Preservation Coalition**

PDF files form a major component of the collections of many DPC members. Despite the advent of the PDF/A standard, significant preservation challenges remain in preserving these files. By working closely with DPC members to understand their needs and feeding them back to the development team, the veraPDF consortium ensured that veraPDF solves real challenges for the Coalition. veraPDF is plugging a major gap in the digital preservation toolset, and has the potential to enhance the accuracy and reduce the costs of preservation for the members of the DPC.

### **Dual Lab**

The main benefits so far for Dual Lab included establishing new contacts and partnerships in the area of digital preservation, deeper involvement into open source software development projects and best practices, more active participation in the events and initiatives of the PDF Association. This all potentially leads to new projects in the areas of PDF technologies and digital preservation and will contribute into further growth of the company.

### **KEEP Solutions**

Many of the customers of KEEP Solutions have well-established format policies. PDF/A is an outstanding example of this. However, there has been little technical capability to adequately validate received files against these policies. veraPDF enabled the company to develop a validation process for our main repository software product - RODA. One of our largest clients is already using this task over their corpora. Future developments include enhancement of other repository systems to support validation of PDF during ingest. We plan to expand our product line with risk assessment processes that can inform content holders on their PDF/A collections.

### **Easy Innova**

First of all, thanks to the continuous interaction with memory institutions, now Easy Innova has a better knowledge of their specific needs around image preservation. All this knowledge is essential to develop a useful tool and to deliver in the future useful services around it.

Secondly, the hard work carried out in developing a specific tool for TIFF conformance checking and to create an ISO recommendation for TIFF archival provided the company with an extensive and valuable knowledge. Now Easy Innova can claim they are one of the most expert companies in the world regarding TIFF technical knowledge and image digital preservation, which is also very important from both the research and the business point of view. This prestigious position allows EasyInnova to stay in contact with top research institutions and companies like Adobe Systems Inc.

Thirdly, EasyInnova has now a very large network of contacts, allowing them to quickly get feedback, validate their proposals and disseminate their services. Among these, there is an important list of key partners and brilliant institutions who provided them with valuable complementary knowledge, crucial to achieve their objectives.

Finally, EasyInnova is already preparing a business plan to exploit commercially the results of the project, whose key ideas are anticipated in Section 5. The business idea behind this is based on selling professional services around TIFF, digital preservation and the DPF Manager. In order to implement this business plan, the company is planning to hire 3 to 4 additional people during the first two years of the commercialization phase after the end of the PREFORMA project.

### **MediaArea.net**

With their flagship product, "MediaInfo", MediaArea has been contacted in the past by several people interested in having a conformance checker in addition to classic MediaInfo report, but the cost of the initial development of a conformance checker and the associated interface (CLI, GUI, server) was too high and they never were able to start this project. Before PREFORMA, MediaArea was not able to start the project because people interested in it were ready to pay for the creation of a portion of the product specific to their institutional needs, but were not able to pay the price for creating the first version with core, extensible features due to the high cost of this first version and research associated with the validation of complex file formats such as video. This PREFORMA



project permitted MediaArea to be able to secure the funding of this initial development with a 2-year trajectory.

It is now possible for MediaArea to show to interested people what they are able to do, for both the end user interfaces and the capabilities of testing every single bit of a compressed stream.

MediaArea was able to hire 2 full time employees in France to work on this project during the development phase, and expect to keep them on as full time employees thanks to business opportunities provided by the software produced as a result of this project. They also have developed strong links with several contractors and plan to continue working with them as partners in future business opportunities, such as expanding the conformance checker software to additional formats.

During the prototyping phase, MediaArea was able to connect with some institutions considering extending MediaConch to the support of other formats used by their institutions, e.g. DV (many archives have lot of DV content) or DPX, as well as testing the conformance of MXF and JPEG 2000 (a combination of formats used and recommended by the American Library of Congress).

Thanks to the PREFORMA project, the foundations of the conformance checker are developed and people can now request smaller developments for additional file formats.

### 4.3 IMPACT ON OTHER DOMAINS

In addition to its main target stakeholder groups (memory institutions and cultural heritage organisations, OS systems developers, standardisation bodies), PREFORMA tried to address other communities that might be interested in the results of the project, specifically the web archiving community, the medical and health care sector and the publishers.

The following contacts have been taken so far in this respect:

- Web archiving community through the National Library of Sweden (KB), which is member of IIPC
- AKU ("Archivierung von Krankenunterlagen"), a German Working Group on the aspect of "Archiving of Medical and Patient-related Documentation", being member of the German Association for Medical Informatics, Biometrics, and Epidemiology (GMDS) ([http://www.gmds.de/fachbereiche/informatik/AGsundPGs/AG\\_Archivierung\\_von\\_Krankenuunterlagen/AG\\_AKU.php](http://www.gmds.de/fachbereiche/informatik/AGsundPGs/AG_Archivierung_von_Krankenuunterlagen/AG_AKU.php))
- Other medical-related Working Groups in Europe and in the world through Fraunhofer IDMT, e.g. the European Federation for Medical Informatics (EFMI), UMIT - Private University for Health Sciences, Medical Informatics and Technology, and the International Medical Informatics Association (IMIA)
- The Finnish Working Group on organizing the Finnish Regional Medical Archive infrastructure and the related legal and technical framework
- Publishers and publishers associations through Promoter, e.g. EDItEUR (the international group coordinating development of the standards infrastructure for electronic commerce in the book, e-book and serials sectors, with over 110 members from 25 countries, including North America, most European countries, and the Asia-Pacific region) and AIE

(Associazione Italiana Editori, the trade association of Italian publishers – and foreign ones operating in Italy – of books, journals, and digital publishing products and contents)

The first answers received both from the German WG on medical archiving and from EDItEUR were positive. They are interested in joining forces and they started to encourage their members to test our prototypes and participate in the PREFORMA events. PREFORMA has been also invited to present its achievements at the Archival Day in Jena in December 2016.

In this light, a joint meeting between the PREFORMA consortium and the German Working Group on PCP and PPI has been organised in the occasion of the PREFORMA Experience Workshop in Berlin. Representatives from the WG attended the workshop and a specific session has been planned where the two consortia presented their activities and set the basis for possible future cooperation. Among the various possibilities, particularly interesting is the application of the PREFORMA methodology and results to the health and medical sector, which is currently under investigation together with HELIOS, Germany's largest hospital operator. owning and operating 112 clinics, including seven maximum care hospitals in Erfurt, Berlin-Buch, Duisburg, Wuppertal, Schwerin, Krefeld and Wiesbaden.

## **4.4 IMPACT ON THE STANDARDISATION ACTIVITIES**

As already said, contributing to the standardisation of the preservation formats represents one of the objectives of PREFORMA.

For this reason, the suppliers working in the prototyping phase have been asked to actively engage in interacting with relevant organisations that maintain the standard specifications used by the open source projects. The aim is to provide feedback, resolve technical issues, and contribute to a dialogue for improvement of the technical specifications of standards.

The following paragraphs present what has been done by each consortium and the main results achieved so far.

### **4.4.1 ELECTRONIC DOCUMENTS: PDF/A AND PDF/A-NEXT**

Since the beginning of the prototyping phase of the PREFORMA project, the veraPDF consortium has collaborated with the relevant ISO working groups (ISO TC 171 SC 2 WG 5, for PDF/A and WG 8 for PDF) via the PDF Association's PDF Validation Technical Working Group (TWG).

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.

After the ISO committee meetings held in Basel (November 2015) and Ghent (May 2016), the PDF Validation TWG, including veraPDF consortium staff, has made progress towards finalizing these list of ambiguities and detailing the resolution thereof. The complete list (known internally as the "Resolution of Ambiguities" document) has been provided to the ISO committee for review in Sydney (December 2016).



In order to promote industry awareness and acceptance of this work, the PDF Validation TWG is developing a PDF Association 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.

Following review by the ISO WG, the PDF Association will proceed to publish the PDF Validation TWG's Resolution of Ambiguities document as Technical Note 0010 in the first half of 2017.

The following summarises the main facts developed or decisions taken so far.

- Established that existing Parts of PDF/A will not be amended in any way. Any clarifications to existing ambiguities will be addressed in a forthcoming Part for PDF/A.
- Led the effort to create a new Part of PDF/A, currently termed "PDF/A-next" and driven awareness of its need.
- Contributed feedback from archivists and other interested parties regarding PDF/A-next to the ISO community.
- Established that the Validation TWG may, as a body, comment on drafts of the PDF/A-next specification based on veraPDF and community findings.
- Submitted proposed enhancements for PDF/A-next to address ambiguities, and proposed additional changes following release of the PDF/A-next Committee Draft (CD) to the community.
- Submitted a request for an ISO WG review of the PDF Validation TWG's Resolution of Ambiguities document prior to its publication as a PDF Association Technical Note.

#### **4.4.2 STILL IMAGES: TIFF AND T/I/A**

As already explained, EasyInnova and the University of Basel started to work since the beginning of the prototyping phase on the definition of a technical specification for the use of TIFF in archival environments, called (TI/A <http://www.ti-a.org>), to be published as an ISO recommendation.

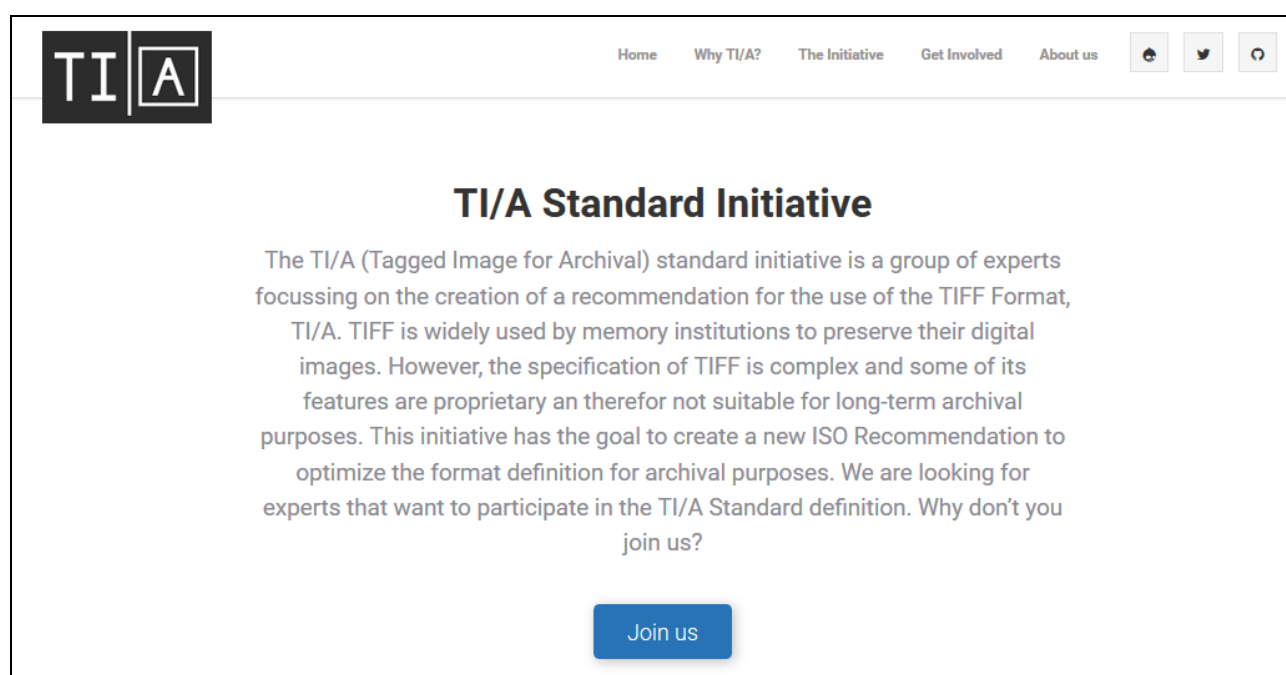
Initially the initiative was named TIFF/A. After a few conversations with Adobe, owner of the TIFF trademark, it was made clear that they had no intention to allow to use the TIFF acronym as part of the new initiative. This explains the change of the name to TI/A. With these premises, and taking into account that the final result will not be a specification of a new file format but a recommendation of the proper use of the existing TIFF format, Adobe agreed to support the TI/A initiative.

In order to create this new recommendation, the following actions have been performed:

- EasyInnova joined AENOR, the Spanish standardization body, and they are now part of the CNT50/SC1, the working group that is responsible for digital formats. The CNT50/SC1 members started informal discussions with other standardization bodies from other countries on our behalf in the ISO/TC46 annual meeting that took place in Beijing, China, on June 2015.

- University of Basel joined the Swiss standardization body, SVN, to lead the effort to create the new recommendation. Both AENOR and the SVN are participating countries in ISO/TC171 technical committee.
- EasyInnova created a website for the TI/A initiative, [www.ti-a.org](http://www.ti-a.org), an Intranet as a tool to allow TIFF experts to participate and to discuss around the recommendations, [intranet.ti-a.org](http://intranet.ti-a.org), and a Twitter account, [twitter.com/TI\\_A\\_Standard](https://twitter.com/TI_A_Standard), to allow any interested party to follow the development of the standard.
- EasyInnova and the University of Basel produced a first draft recommendation based on the feedback collected from experts participating in the TI/A Intranet and from memory institutions, to be used as the starting point to engage with GLAM institutions (Galleries, Libraries, Archives and Museums). The objective is to bring together a core group of experts to further develop the standard based on the draft, and a group of interested people to provide feedback on it. This draft version has been presented and discussed in Basel in November 2015 at the ISO/TC171 committee meeting.
- EasyInnova and the University of Basel analysed about 2 Million TIFF files of three large Archives in Switzerland to get a comprehensive picture of the real data stored by memory institutions and how they evolved in time, with particular regard to the application and selection of tags. The objective of this work was to decide the most appropriate rules to include in the technical recommendation so to mitigate the risk of unnecessary migrations due to technical advancements in the field of imaging.
- The final draft is available at the TI/A intranet and it was submitted to the ISO TC171 committee at the end of October and discussed during the working group meeting in Sydney at the end of November 2016.

Now the TI/A draft is following the ISO standardization process.



The screenshot shows the homepage of the TI/A Standard Initiative website. At the top left is the TI/A logo, consisting of the letters 'TI' and 'A' in white on a black background. To the right of the logo is a navigation menu with links: Home, Why TI/A?, The Initiative, Get Involved, and About us. Further right are three social media icons: Facebook, Twitter, and LinkedIn. The main heading is 'TI/A Standard Initiative' in a bold, dark blue font. Below the heading is a paragraph of text explaining the initiative's purpose: 'The TI/A (Tagged Image for Archival) standard initiative is a group of experts focussing on the creation of a recommendation for the use of the TIFF Format, TI/A. TIFF is widely used by memory institutions to preserve their digital images. However, the specification of TIFF is complex and some of its features are proprietary and therefore not suitable for long-term archival purposes. This initiative has the goal to create a new ISO Recommendation to optimize the format definition for archival purposes. We are looking for experts that want to participate in the TI/A Standard definition. Why don't you join us?'. At the bottom center is a blue button with the text 'Join us' in white.

*Fig.10. TI/A website - snapshot*

#### **4.4.3 AUDIOVISUAL FILES: MATROSKA AND FFV1**

The MediaArea team has actively participated in the ongoing refinement of the FFV1 and Matroska specification (focusing initially on the EBML specification, which is the basis of the Matroska format), bringing communities around the standardization efforts of these formats. In particular, MediaArea has been analyzing existing standards for both formats and presenting proposals for change or comment on the Matroska and FFV1 listservs, requesting feedback from the communities.

Within the EBML specification, MediaArea helped draft the concept of the EBML Schema, which defines an EBML Document (of which Matroska is one type) in a schema which can be used to validate a document similar to how an XML Schema can function. This effort helps facilitate and consolidate many implementation checker tests as the specification for Matroska now has a more machine-readable form.

Development of the FFV1 Specification has focused on migrating the format of the document from lyx to markdown to encourage collaboration in a more accessible format. By rewriting specifications into Markdown and posting on Github, more people are able to contribute to the specifications, including archivists without development or prior Github experience.

Via the Matroska listserv and Github page, MediaArea team members have been initiating conversation surrounding further standardization and needed clarification for the Matroska EBML specification and contributing new, changed, or restructured guidelines for the format. The Matroska specification page hosted by MatroskaOrg in GitHub now contains 136 commits by 7 code contributors (and many more commenting or giving feedback on Github or via CELLAR listserv communication). The foundational format of Matroska, EBML, has additionally received 296 commits from 7 contributors and support from the working group members.

MediaArea is also actively working with the IETF community to bring Matroska and FFV1 communities into the IETF review process for standardization. Both lead authors of Matroska and FFV1 publicly stated their support for the process on the IETF DISPATCH list.

MediaArea presented first on planned standardization efforts at IETF 93 in Prague in June 2015 as part of the DISPATCH group's agenda. There was significant interest from the IETF community both during and after the session in bringing this work to the IETF. Following the conference, MediaArea spoke with working group chairs and area directors about proceeding with work to start the IETF review process. The consensus from IETF members was that FFV1 and Matroska should be brought into the IETF standardization process through the formation of a new working group called CELLAR: Codec Encoding for LossLess Archiving and Realtime transmission.

On October 22, the IESG approved the CELLAR working group charter for external review with no objections or blocking issues. The inaugural CELLAR working group kickoff meeting was held at IETF96, hosted in July 2016 in Berlin. The meeting, which included presentations by Steve Lhomme and Jerome Martinez on the state of Matroska and FFV1's specification work, was productive and received helpful feedback from many members of the IETF community.

Once approved, work commenced via the CELLAR mailing list and live online discussion sessions organized by the co-chair Tessa Fallon.

Surrounding the IETF meeting, MediaArea co-hosted a symposium in Berlin, called No Time to Wait, that focused on the standardization and use of FFV1 and Matroska in archives. The symposium was a great success and extremely productive at speeding along the standardization efforts of FFV1 and Matroska. Since the meeting, the MediaArea team has posted videos from the symposium's presentations and has been engaging with new participants in the CELLAR working group. The No Time to Wait symposium provided good opportunities to learn from and collaborate with related standardization efforts such as Google's work to standardize a Matroska branch, webm, and the Library of Congress's work in defining AS07.

In the CELLAR working group, MediaArea have aided in rewriting sections with clarified language and contributed recommendations from mining the Matroska datasets. They also adapted a draft of Google's draft of 360 degree and virtual reality elements for Matroska presentations to help aid in keeping Google's work in the WebM project (<http://www.webmproject.org/>) and CELLAR's work in Matroska well aligned.

In order to facilitate the implementation of these standards, MediaArea have sponsored and participated in patches to related open source projects, and they planned to support soon a training seminar to help teach interested standardization participants in the tools of the working group, such as Markdown and GitHub.

Finally, MediaArea continues in community work to prepare for a set of recommendations for the archival use of Matroska and FFV1 in a manner analogous to the Library of Congress's similar standardization effort in AS-07, which relies upon MXF and JPEG2000.

## 5. SUSTAINABILITY, FUTURE USE AND EXPLOITATION PLANS

### 5.1 SUSTAINABILITY OF THE PREFORMA RESULTS

When we talk about how to sustain the results of the PREFORMA project, we must distinguish among the following three main categories:

- Sustainability of the PREFORMA **network**, i.e. the community of project's partners, suppliers, associate partners who are working together at different levels in the project and who share the interest in the development of 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* (cfr. The PREFORMA Challenge).
- Sustainability of the PREFORMA **infrastructure and data**, i.e. the Open Source Portal, through which the software released by the suppliers, the documentation and other related material is made available to the community under the PREFORMA open source licenses, and the datasets made available for examination, training, evaluation and dissemination purposes by the memory institutions.
- Sustainability of the PREFORMA **tools**, i.e. the open source conformance checkers developed by the suppliers.

#### 5.1.1 SUSTAINABILITY OF THE PREFORMA NETWORK

Discussions already started to understand how to keep alive the PREFORMA network in the future and how to leverage on the expertise of its members to follow up the results of the project.

Three main strands have been identified so far to follow up the results of PREFORMA:

- First line – **research**: continue the research on formats, consider new PCP proposal to develop new services to make the archives more resilient (Work Programme 2019-2020)
- Second line – **implementation**: consider PPI proposal to deploy the services procured in PREFORMA and create an infrastructure which can serve different users with different needs (35% co-funded by the EC for deployment, installations, training, equipment, etc.)
- Third line – **networking**: create/reinforce the community of organisations and institutions interested to use PCP instruments in the cultural heritage field

In this light, a document has been drafted by the consortium and sent to the European Commission providing some ideas that could serve as an input for the finalisation of the H2020 Work Programme 2018-2020. The document included:

- Ideas for possible PCP projects to procure new technologies to improve the digital preservation workflow in digital archives:
  - continue with the same model as PREFORMA and expand to new file formats and file types, such as 3D models. Year: 2018. Budget: 3.5-4 M euro

- develop new tools for automatic correction of metadata, automatic record management, intelligent recommendations and decision making based on policies. Year: 2019. Budget: 5-6 M euro
- develop machine learning tools to assess and improve the quality of services and for benchmarking and performance measuring. Year: 2020. Budget: 6-7 M euro
- Ideas for possible PPI projects, following up the results of PREFORMA:
  - combine models and build an intelligent workflow around the developed tools, in the line of what PRESTO did in the AV domain but as a PPI project and involving also the other domains (documents and images). Year: 2018. Budget: 3.5-4 M euro
- Ideas for a CSA on PCP/PPI actions on digital preservation, as there is no space for preservation in the current research topics. Year: 2018. Budget: 1 M euro

During the last year of the project, a greater effort will be allocated by the PREFORMA partners to clarify the contributions and roles played by the consortium as well as the suppliers in this sustainability effort and to find an agreement how to formalise them.

### 5.1.2 SUSTAINABILITY OF THE PREFORMA INFRASTRUCTURE AND DATA

PREFORMA developed two tools to store and make available to the community the software and the data to be used to test it.

- The **Open Source Portal**, embedded in the project's website, which provides an overview and references to each open source project that is currently working in the prototyping phase. It acts as an entry point for all interested suppliers and memory institutions allowing easy navigation to all externally hosted resources. This section contains an overview webpage for each open source project, featuring information, material, and links to specific resources used in each project and hosted on the project's open platform (including documentation and coding practices, test files, outreach and standardisation activities, collaboration and development tools, etc.). From the overview page it is possible to reach another dedicated webpage, from which it is possible to download executables, source code and complete build environment for each conformance checker, divided by release. The source code and executables are provided under the two specific open source licenses "MPLv2 or later" and "GPLv3 or later". The build environment is provided under an open source license, i.e. a license approved by the Open Source Initiative ([www.opensource.org](http://www.opensource.org)).
- The **PREFORMA Vault**, based on Amazon S3 service, which is the central logical point where all content providers submit their files to be used for examination, training, evaluation and dissemination purposes. The purpose of the Vault is to have an effective and practical instrument for handling all submissions of incoming and outgoing files from Providers to Suppliers ensuring the best possible distribution of training and evaluation files and keeping track of the content that has been used for the training files. In this way, we will be certain that the evaluation files correspond to the training files and that they have not been released in advance to the Suppliers. The Data management plan for training, testing and demonstration files in the PREFORMA project describes the functionality of the Vault and



outlines the framework that governs the provision and management of the files to be used to test the prototypes.



The screenshot shows the PREFORMA Open Source Portal website. The header includes the PREFORMA logo, the European Union flag, and the text: "This project has received funding from the European Union's Seventh Framework Programme under grant agreement no 619568". The navigation menu includes: HOME, PROJECT, PARTNERS, TENDER, EVENTS, OPEN SOURCE PORTAL, COMMUNITY, DOWNLOAD, and CONTACTS. The main content area is titled "OPEN SOURCE PORTAL" and contains three project entries:

- PROJECT N.1. veraPDF: INDUSTRY-SUPPORTED PDF/A VALIDATION**  
by Open Preservation Foundation, PDF Association, Digital Preservation Coalition, Dual Lab, KEEP SOLUTIONS  
A unique collaboration, the veraPDF Consortium brings together an end user community and a software industry rooted in the principle of interoperability based on ISO standardized technology... [access project page >>](#)
- PROJECT N.2. DPF MANAGER: DIGITAL PRESERVATION FORMATS MANAGER**  
by Easy Innova  
DPF Manager is an open source modular TIFF conformance checker that is extremely easy to use, to integrate with existing and new projects, and to deploy in a multitude of different scenarios... [access project page >>](#)
- PROJECT N.3. MEDIACONCH - CONFORMANCE CHECKING FOR AUDIOVISUAL FILES**  
by MediaArea.net  
MediaConch is an extensible, open source software project consisting of an implementation checker, policy checker, reporter and fixer that targets preservation-level audiovisual files for use in memory institutions... [access project page >>](#)

On the right side, there is a section titled "PREFORMA OPEN SOURCE PROJECTS" listing:

- veraPDF
- DPF MANAGER
- MEDIACONCH
- >> View all the successful proposals that participated to the design phase
- PREFORMA VAULT (Access restricted)

Below this is a section titled "OTHER RELATED TOOLS" listing:

- ARCHIVEMATICA
- EXACTLY
- JPYLYZER
- KOST-VAL
- MEDIA FILE CHECKER
- XENA

*Fig.11. PREFORMA Open Source Portal - snapshot*

Discussions already started in the PREFORMA consortium to agree how to ensure long-term availability of the PREFORMA software and associated digital assets after the end of the project.

The idea is to find an agreement between Riksarkivet (the project leader) and a public sector organisation (e.g. SUNET) to ensure the technical maintenance of the PREFORMA infrastructure.

SUNET is the Swedish University Computer Network. It is governed by a board appointed by the Swedish Research Council and it provides Swedish universities and colleges with access to well-developed and effective national and international data communication and related services that meet their needs, whatever their geographical location.

This possibility will be further investigated in the last year of the project.

### 5.1.3 SUSTAINABILITY OF THE PREFORMA TOOLS

The sustainability of the results of PREFORMA relies most of all on the commitment of the **suppliers** that are working in the prototyping phase to continue to maintain and improve the tools after the end of the funding period and on their ability to establish a sound open source community of users and developers.

All released software provided in development version, stable version, and deployed version are kept available at each open source project website during and after the PREFORMA project. At this point in time it is expected that each open source project has successfully developed a vibrant and sustainable open source community and it is therefore in the interest of all stakeholders that there is continued development in each project. This is also required by the specific open source licenses chosen, which goes beyond the time frame for the PREFORMA project.

The **veraPDF consortium** is not a commercial entity bringing products to market. Nevertheless, veraPDF will underpin a variety of commercial offerings from third-parties including both PDF industry and repository vendors offering additional services and integration into widely adopted open-source repository systems.

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.

In addition, part of the mission of the **Open Preservation Foundation**, which is a partner of veraPDF consortium, is the sustainability of project results in the digital preservation field. Open-source projects are not typically governed by a central organisation or group. They tend to evolve in a fairly uncontrolled way driven by the wishes of individual contributors. In order for an open-source product to retain a certain level of quality, however, a core team must exist that takes upon itself the responsibility of testing and validating the contributions of the community. This activity is quintessential to ensure quality and to make sure that the product remains open and faithful to its original objectives. The OPF plays a major role in leading this task, and co-ordinating the sustainability of a product which has value to multiple communities. Its substantial experience in maintaining open source software backed up by the knowledge and influence of its members will ensure that the product will remain viable and true to its original intents while at the same time being able to evolve to meet new requirements from its user base. Since the end of the Planets project, they have sustained and made available the results of a whole set of projects. Over this time, they 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.

**Easy Innova** will also guarantee the sustainability of the DPF Manager project as this is part of their core business as SME.

Their exploitation plan is based on offering services like Cloud-based SaaS, on premise deployments, technical support and maintenance contracts, consultancy services and training courses to developers, integrators and end-users. Memory institutions would need the technical support and knowledge of a specialised company to integrate, maintain and evolve their long-term preservation systems. In addition, taking into account that DPF Manager will be the first tool to validate the new TI/A ISO recommendations, this should give more value to the software compared to other possible competitors, e.g. JHOVE.

Finally, **MediaArea** is already exploring opportunities to extend and improve the functionalities of MediaConch and seeking sponsorships to add additional formats to the validation checker.

Functionally, MediaConch can already be expanded to support file formats beyond Matroska, FFV1, and LPCM (and PDF/TIFF via integration of work from the other suppliers). There is potential for MediaConch to become the conformance checking software for *\*any\** audiovisual formats after the completion of this project, not just limited to Matroska and FFV1.

MediaConch's local policy creation feature can already be extended too out to any format that MediaInfo supports. For this reason, MediaConch is delivered with some basic support of other formats used by archives e.g MOV, AVI or MXF, as a demonstration that MediaConch is versatile.

Moving forward, there are several directions MediaArea hope to continue exploring in order to continue to develop, promote, and apply MediaConch and related work in archival, broadcast, and other media applications. This includes:

- including support for additional formats in MediaConch, either extending the implementation checker or via integration of plugins;
- developing aspects of MediaConch out as a workflow management tool so MediaConch can add transcoding and others steps to its existing functions;
- exploring the use of the policy checker as a compatibility checker, i.e. exploring the need for machine-readable media policies and how the policy features of MediaConch could be better used by systems, hardware, and software that depend on particular media compatibilities;
- converting written archival audiovisual policies into MediaConch policies as a method for testing compliance to help encourage digitization vendors that work with preservation communities to work more directly toward similarities in policies;
- developing new reporting templates and demonstrating the ability to customize, refine, and change reports to targeted usage;
- continuing to advocate for file format strategies that regionalize fixity to support potential corrections when needed.

The following chapter presents more in details the initial ideas and plans of the suppliers related to the commercial exploitation and route to market of the PREFORMA tools.

## 5.2 ROUTE TO MARKET OF THE PREFORMA TOOLS

The results presented in this section is based on the preliminary business analysis and models developed by the three suppliers – verPDF, Easy Innova and MediaArea – during the design and prototyping phase.

In particular, the market analysis and segmentation sections have been taken from the Business Plan submitted by Easy Innova as part of their proposal and further refined at the end of the design phase.

### 5.2.1 MARKET ANALYSIS

According to the Planets whitepaper (An Emerging Market: Establishing Demand for digital Preservation Tools and Services. July 2010) the digital preservation market is still in its infancy. Ibis World Research estimates a growth rate of c. 10 % – 12 % per year, until 2019, within the European Union alone. As such, the relevance of digital preservation is extending far beyond memory institutions into all business sectors as the financial sector, healthcare organisations, pharma and oil and photographic industries. For many, though, digital preservation is seen as an onerous obligation not an opportunity to realise the long-term value of existing resources.

Memory institutions are culturally predisposed to retaining information and see inherent longterm value in it. By contrast, other sectors are having this imposed on them from outside, by legislation or business pressures, and regard it as ‘another drain on the bottom line’. This is manifested in the lack of digital preservation policies, or their incomplete nature, and the shortterm project-based nature of budgets, which vanish in times of economic uncertainty. It is also seen in the confusion about what digital preservation entails.

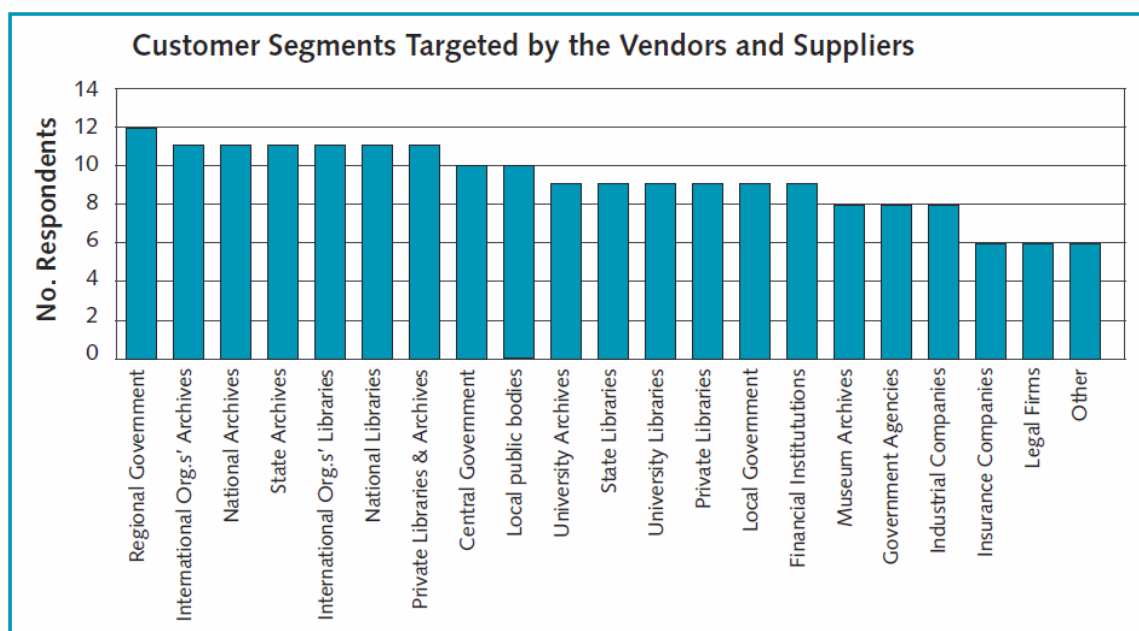
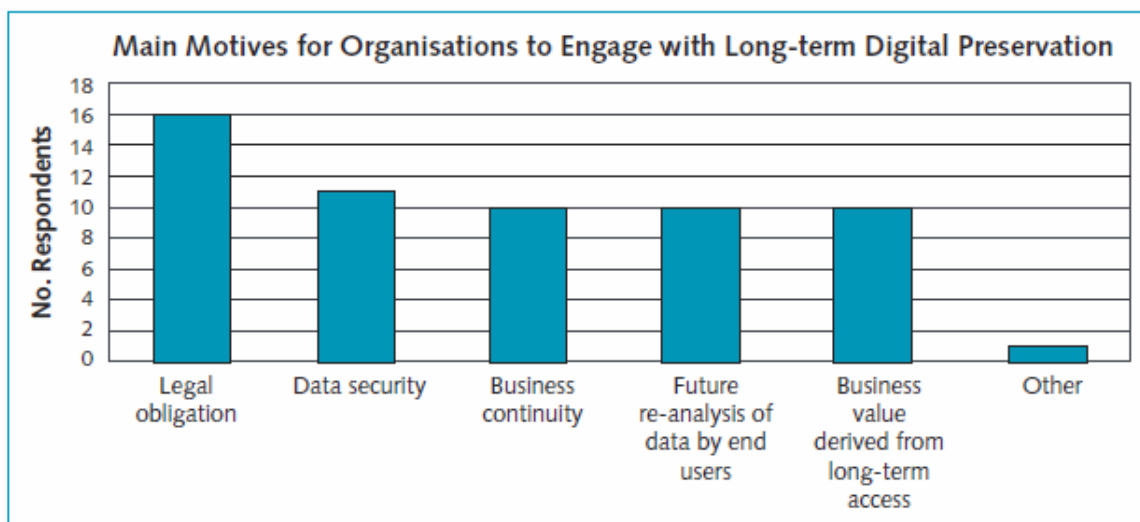


Fig. 1. Customer segments targeted by the planets supplier and vendor briefings' participants



*Fig. 2. Main motives for organisations to engage in long-term digital preservation*

Solutions already need to deal with a wide variety of content and this will only increase in the future. It is therefore no surprise that scalability to high rates of ingest and volumes of content are important now and all research seems to indicate that potential clients care deeply about this. Although preservation is not yet deemed to be necessary for as long as 50 years, the ability to trust a preservation solution to keep an object intact and accessible is regarded as important.

The need for preservation standards is recognised, as is the need for their rationalisation. While there is a strong preference for migration, there is emerging recognition that emulation has a place for particular types of content.

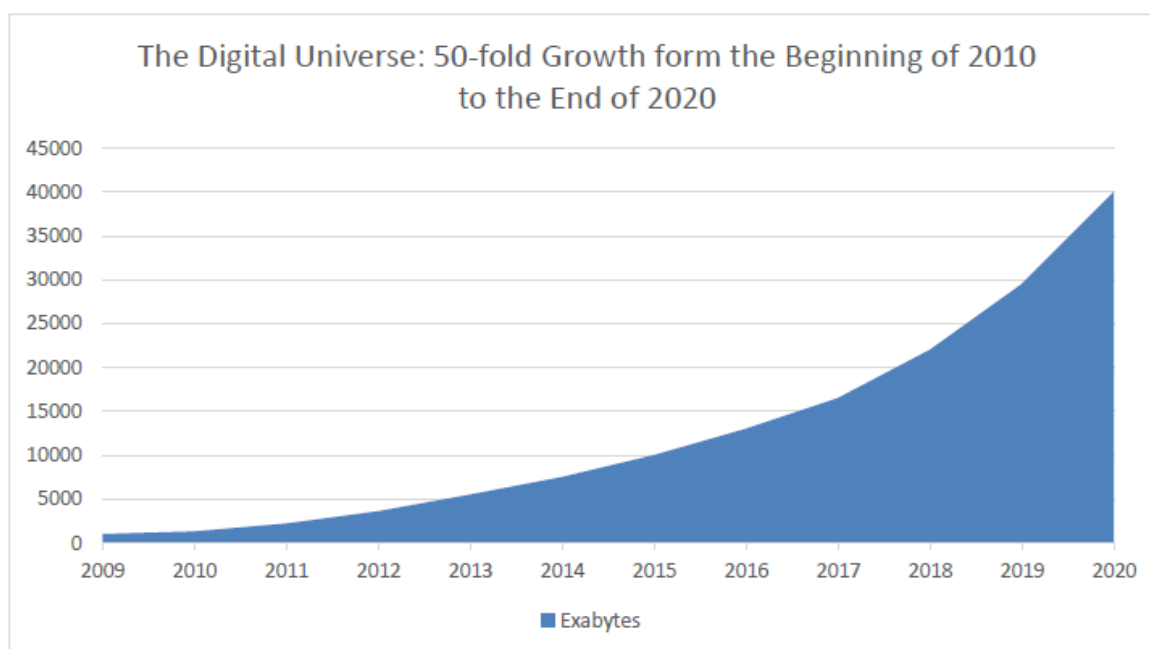
Future engagement requires work to be done to raise awareness about the importance of preserving digital information and to articulate a business case for it. At the most basic level there needs to be a clear definition of what digital preservation is. Following on from that, the costs and benefits of digital preservation need to be set out and the cost of no action needs to be clearly demonstrated.

Further guidance on how to implement digital preservation is required and there is a demand for information and training. There is an on-going need for the development of tools and services, in particular to cater for the needs of smaller organisations with fewer resources (SMBs – small and medium businesses account for over 60 % of the European Economy, according to the European Commission). Finally, workable approaches to emulation need to be developed and so predispose end-users to use it where appropriate.

On the other hand, IDC (the International Data Corporation) reported the exponential growth in digital information and stated that the number of "files," or containers that encapsulate the information in the digital universe, is growing even faster than the information itself as more and more embedded systems pump their bits into the digital cosmos. They forecasted a growth by a factor of 8 in the 2011-2016 periods and just a slight growth of IT staff available to manage them. Since 2005, the investment by enterprises in the digital universe has increased 50% to \$4 trillion.

That represents money spent on hardware, software, services, and staff to create, manage, and store and derive revenues from the digital universe.

IDC estimates that by 2020, as much as 33% of the digital universe will contain information that might be valuable if analyzed, compared with 25% today. This untapped value could be found in patterns in social media usage, correlations in scientific data from discrete studies, medical information intersected with sociological data, faces in security footage, and so on. However, even with a generous estimate, the amount of information in the digital universe that is "tagged" accounts for only about 3% of the digital universe in 2012, and that which is analyzed is half a percent of the digital universe. Herein is the promise of "Big Data" technology — the extraction of value from the large untapped pools of data in the digital universe.



*Fig. 3. Source: IDC's Digital Universe Study, sponsored by EMC, December 2012*

However, the CIOs, data scientists, digital entrepreneurs, etc. already know the value that can be found in this ever-expanding collection of digital bits. Hence, there is excitement about Big Data technologies, automatic tagging algorithms, real-time analytics, social media data mining, and myriad new storage technologies.



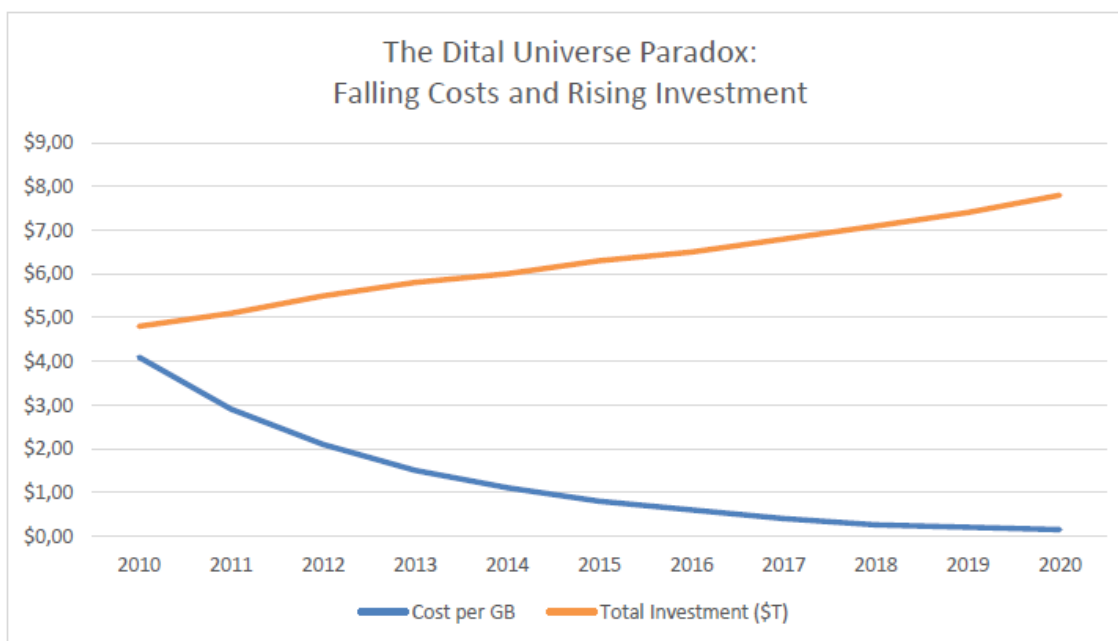


Fig. 4. Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Finally, an ESG (Enterprise Strategy Group) Report 2014 reveals the current amount of storage capacity in terms of archiving for enterprises as well as its expected annually growth. ESG also states that a growing number of people are beginning to realize that while archives and backups are different, they are complementary with both having a reasonable place of consideration within one's overall data protection strategy.

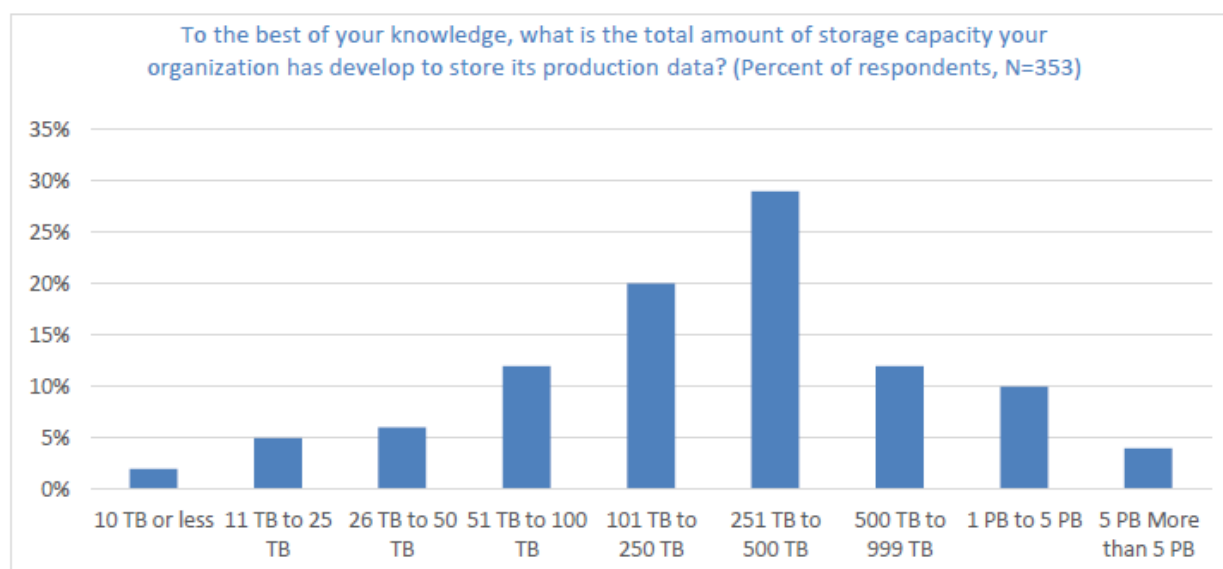
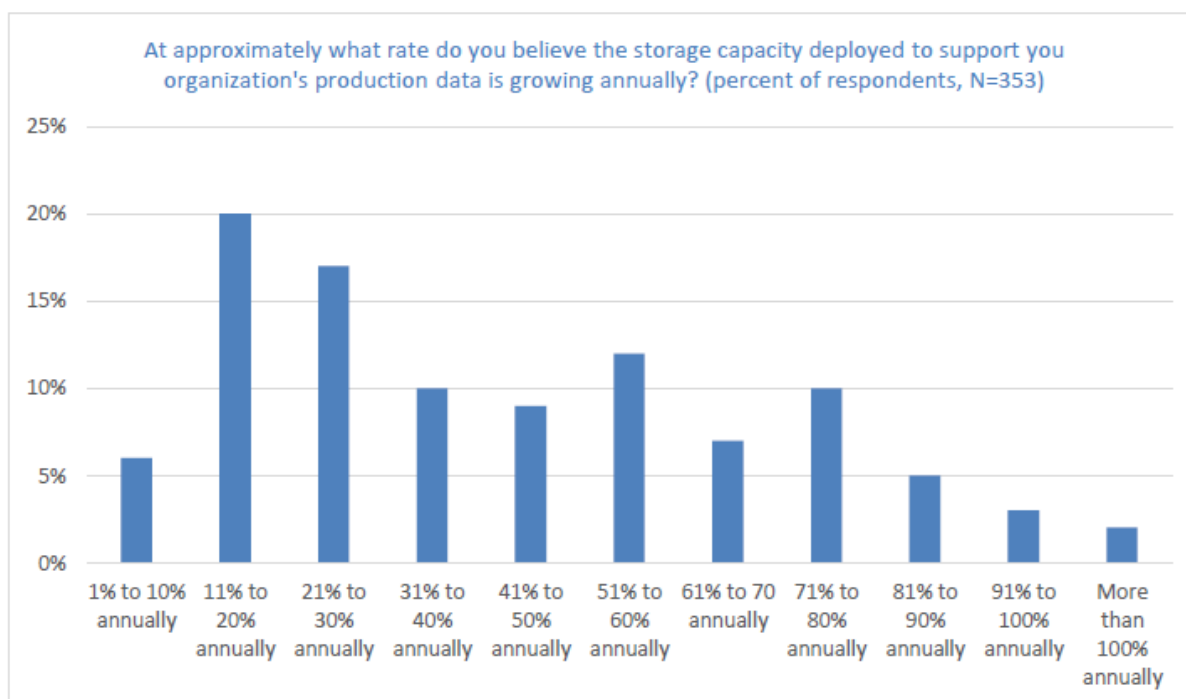


Fig. 5. Source: Enterprise Strategy Group, 2014.



*Fig. 6. Source: Enterprise Strategy Group, 2014.*

All the above suggest that digital preservation is a need for certain businesses but in the near future it will become of key relevance as the amount of digital information begins to increase. This creates a prone horizon to earn money with preservation tools like the ones developed in PREFORMA.

### 5.2.2 MARKET SEGMENTATION

Although the bits of the digital universe may travel at Internet speeds around the globe, it is possible to assign a place of origin to them and chart the map of the digital universe.

In the early days, the digital universe was a developed world phenomenon, with 48% of the digital universe in 2005 springing forth from just the United States and Western Europe.

Emerging markets accounted for less than 20%. However, the share of the digital universe attributable to emerging markets is up to 36% in 2012 and will be 62% by 2020. By then, China alone will generate 21% of the bit stream entering the digital universe.

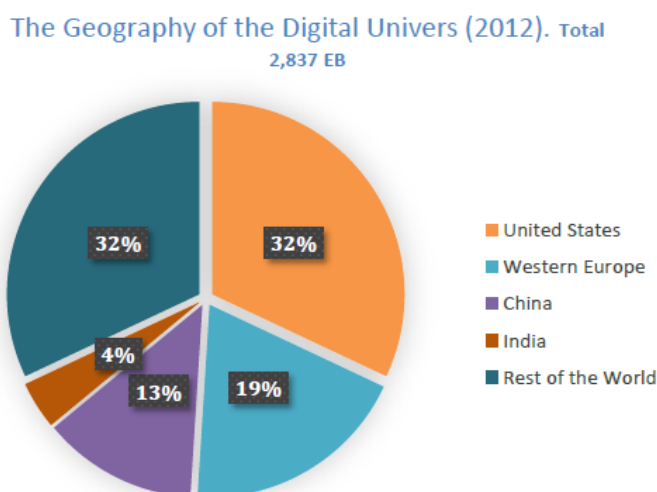


Fig. 7. Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

The PREFORMA conformance checking tools are designed primarily for memory and heritage institutions that need to preserve content such as museums, libraries and multimedia archives, as well as for other data keeping institutions that are planning digital preservation strategies, e.g. publishers, medical institutions, universities, municipalities, etc.

Focusing in Europe as an example of the main market segment for the PREFORMA suppliers, it is shown in the next figure that within the 106,000 institutions found in the EU27 area, around 6,000 relevant institutions are culturally predisposed to retaining information and see inherent long-term value in Digital Preservation.

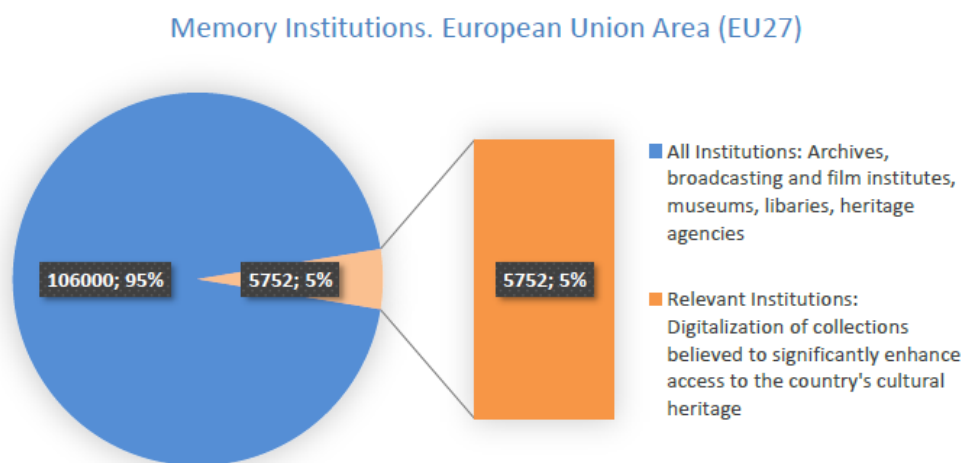


Fig. 8. Source: NUMERIC's FP7 Project. Study deliverable № 8. 2009.

The following diagram and table show the number of estimated relevant institutions by cultural domain and country.

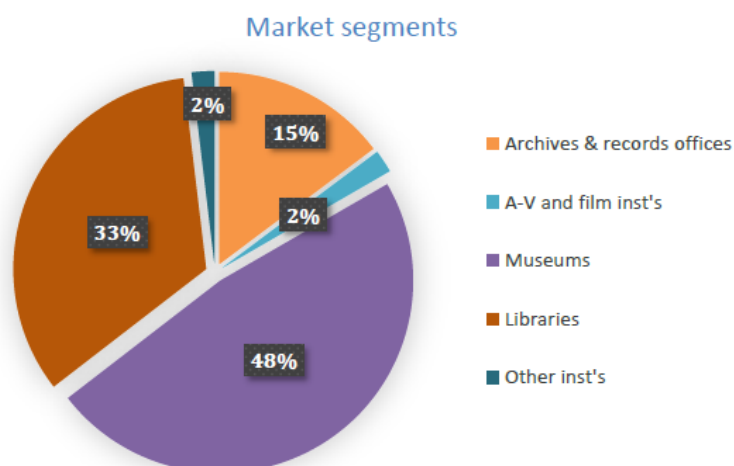


Fig. 9. Source: NUMERIC's FP7 Project. Study deliverable № 8. 2009.

Country/ Type of Institution	Archives and records offices	A-V & film inst's	Museums	Libraries	Other inst's	All domains
Austria	15	7	33	14	2	71
Belgium	24	9	54	15	18	120
Bulgaria	3	4	15	14	1	37
Cyprus	1	6	37	60	4	108
Czech Republic	29	1	162	41	4	237
Denmark	1	3	8	2	-	14
Estonia	1	2	21	5	-	29
Finland	17	2	60	15	1	95
France	154	27	101	156	39	477
Germany	17	3	528	249	4	801
Greece	11	8	15	6	8	48
Hungary	85	4	184	156	1	430
Ireland	3	3	31	49	-	86
Italy	100	3	424	48	3	578
Lithuania	2	1	6	6	-	15
Luxembourg	2	1	5	1	3	12
Latvia	2	3	110	27	-	142
Malta	1	-	2	2	-	5
Netherlands	50	5	73	32	6	166
Poland	8	3	18	88	2	119
Portugal	10	2	50	22	2	86
Romania	20	1	230	312	2	565
Slovakia	47	2	63	14	3	129
Slovenia	7	2	102	68	1	180
Spain ‡	55	2	104	101	3	265
Sweden	56	4	91	60	1	212
United Kingdom	119	1	227	360	1	708
EU27	840	109	2754	1923	109	5735

Fig. 10. Source: NUMERIC's FP7 Project. Study deliverable № 8. 2009.

Service providers and IT Systems integrators also play a very important role in the provision and deployment of digital preservation solutions. Therefore, they might represent a second market segment to be addressed and it is of key relevance to design specific actions to reach them, which will result in turn in a wider customer base thanks to these intermediaries.

Finally, it is important to take into account the software companies and freelance developers community, which are always looking for business opportunities. The PREFORMA tools will offer them the possibility to increase visibility for their conversion products and tools through the possibility to integrate their applications or develop specific modules.

### 5.2.3 COMPETITION

The market presented in the previous paragraph is best described as an oligopoly. Within this situation, there are several agents of supply, i.e. the competitors, and several agents for demand, i.e. the clients. Consumers within the market assess that there are not significant differences between competitors in terms of price, deliverability, and flexibility, i.e. how the solutions meet their specific needs. As such, they are willing to pay for additional services, should those services be provided in a streamlined manner, and without any kind of disruption.

The fact that the market is oligopolistic means that each competitor pays attention to the development patterns of its rival companies. As a consequence, it adapts its own strategy based on the actions and strategies of similar companies.

File format validation tools are used to determine the level of compliance of a digital object to the file format specification. This is very important to ensure that these files will be compatible with future viewers and therefore to guarantee its preservation in the long-term.

Among the reference tools currently available for file format validation, we can list:

- **JHOVE** (<http://jhove.sourceforge.net/>) [OS], which provides functions to perform format-specific identification, validation, and characterization of digital objects.
- **JHOVE2** (<https://bitbucket.org/jhove2/main/wiki/Home>) [OS]. With respect to JHOVE, JHOVE2 project generalizes the concept of format characterization to include identification, validation, feature extraction, and policy-based assessment. The target of this characterization is not a simple digital file, but a (potentially) complex digital object that may be instantiated in multiple files.
- **KOST-Val** (<http://kost-ceco.ch/cms/index.php?id=250,436,0,0,1,0>) [OS], an open source java-based validator for different file formats (TIFF, SIARD, PDF/A, JP2, JPEG) and Submission Information Package (SIP) developed by CECO (Centre de coordination pour l'archivage à long terme de documents électroniques), a community of state archives from Switzerland. The tool integrates JHOVE for TIFF validation.
- **Archivematica digital preservation system** (<https://www.archivematica.org/>), a compilation of open source tools to set up a digital repository. It provides in one effort a complete workflow for ingesting, normalising, storing and publishing digital assets. However, Archivematica currently does not include a conformance checker and it focuses on file identification rather than validation.

- **Jpylyzer** (<http://openplanets.github.io/jpylyzer/>), a validator and properties extractor for JPEG 2000 Part 1 (JP2) images which was designed to address validation of image and encoding properties against both the JP2 format specifications and an institute-specific profile.
- **Media File Checker**, a software agent developed in the framework of the Europeana Creative project as part of the Content Re-use Framework. It performs automated checking and gathering of technical properties linked to Europeana data Model (EDM) entities (edm:WebResources).
- **Xena** (XML Electronic Normalising for Archives, <http://xena.sourceforge.net/>), a normalising tool developed by the National Archives of Australia as part of their Digital Preservation Software Platform. Xena detects the file format of a digital object and transforms digital files into open formats for long term preservation.
- **Exactly** (<https://www.avpreserve.com/tools/exactly/>), a simple and easy to use application for remotely and safely transferring any born-digital material from a sender to a recipient.

As the market addressed by PREFORMA and by its suppliers is constantly growing, additional competitors can always enter it. However, entry barriers are considerable, due to the amount of capital required for the successful design and implementation of similar platforms.

## 5.2.4 COSTS AND REVENUES ANALYSIS

### Cost Structure

The main costs to sustain the commercialisation of the PREFORMA tools are those related to the payment of the technical and commercial staff involved in the maintenance, deployment and customisation of the platform, the “on premise” installations, the marketing campaign and the direct sales through sales agents.

Apart from the people involved in the technical development and maintenance both of the core functionality and of any additional module requested by the clients, taking into account that the main target segment is composed by museums, libraries, archives and other institutions with preservation needs, it is important to foresee a dedicated personal assistance to the customers, with the possibility to offer customisation services, consultancy and training. Community development also requires time and dedicated resources.

In addition, a cloud provider will be needed. One of the most critical aspects to be considered will be the quality, stability, scalability and performance of service together with an extremely robust disaster-recovery mechanism. Data storage capacity and data transfer volumes will be also very important. The cloud-based approach would enable a smooth or aggressive growth in the infrastructure requirements according to the real needs.

Finally, it is necessary to take into account the commercial and marketing actions needed to raise awareness about the new products and services and promote their uptake.

### Revenue Streams

As already explained in the Introduction to the deliverable, all the PREFORMA software and the associated digital assets need to be released under specific open source licenses. This open-source approach ensures 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/communities, to integrate and deploy them either in new software systems or in already existing (legacy) systems that are already used at memory institutions.

In addition, this approach will encourage participation and contribution of developers and organisations in the further development and improvement of the tools.

However, this does not prevent service and technology providers, like the PREFORMA suppliers, to build commercial offers around the open source tools developed, offering support and consultancy services to the memory institutions, for future the adaptation, improvements, integration and deployment of the PREFORMA tools into their legacy systems and into their networked environments.

The possible revenue streams identified so far can be grouped in the following areas:

#### *Selling optional proprietary extensions*

Development of extensions to the open source software to be sold using a closed license. 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;
- advanced metadata fixers and software for automated correction of files that have not been validated and where the reason might be more complex than a metadata error.

#### *Combinations with other software offerings*

Integration of the open source software developed in PREFORMA as a complement to other commercial software products, 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.

This includes also the possibility to develop plug-ins to other open source solutions within the field, e.g. plug-ins to collection/records management and document handling systems such as Alfresco (<https://www.alfresco.com/>), MuseumPlus (<http://www.zetcom.com/en/products/>) or TMS (<http://www.gallerysystems.com/products-and-services/tms/>).

#### *Hosting area for validation services*

A Cloud-based/SaaS hosting and maintenance service can be offered to small organizations and institutions that have a need for a hosting area for validation. Different packages can be offered accordingly to the volume of files to be processed.

#### *On premise deployments*

On-premise installation and deployment of the PREFORMA software is another possibility that can be offered to those clients that prefer to keep the full control of their IT infrastructure, but need help setting the platform up.

#### *Marketplace for third-party developers and sellers*

The tools developed in PREFORMA have been designed as a modular and open platform over which 3rd parties can develop their own modules. This approach will foster the development of a third party modules/plugin market that can enrich and make more dynamic the original tools. A centralised marketplace can be offered to developers interested in creating and commercializing modules developed on top of the PREFORMA tools, applying a fee to each sale for the administration of the marketplace and for the validation of the correct functionality of the new modules.

#### *Certification for service providers*

A certification service can be offered to those companies/providers who want to offer the service of installing configuring and deploying the software in their countries. As certified providers, they will get training, technical support and publicity.

#### *Customisation, maintenance and technical support services*

Professional services such as customisation, maintenance and support can be offered to those clients who want to outsource the technical maintenance and support. This includes:

- Use requirement analysis: includes the analysis of the requirements from the client and the suggestion for the best architecture based on a variety of open source and commercial tools;
- Installation and configuration: includes the correct installation, configuration and deployment of the whole system in the production site, this being either on premise at the client location or in the cloud;
- Maintenance and support: encompasses the diagnosis and resolution of problems, user support, and changes to the system's settings in order to cope with modifications in its execution environment. It also includes bug fixing;
- Feature development and customisation: consists of the analysis and development of new software features in order to cope with additional client requirements, e.g. development of modules to interoperate with existing systems in the client organisation;
- Implementation of customer-specific workflows;
- Set up of the continuous integration system and test automation framework;
- Manual quality assurance, including the test plan creation and management;
- Design of user interfaces (desktop, web, mobile) and custom report templates.

#### *Training services*

Different types of training courses can be defined to cover the needs of different user profiles. Training courses can be offered to developers, service providers, integrators, managers, decision takers and end-users. The content can vary from how to configure and use the validator software

and reports, to how to customise it, until how to identify the file formats that are most suitable for digital preservation.

#### *Other consultancy services*

Other kind of consultancy services can be designed according to the user needs, e.g. regarding how to work with the OAIS framework in relation to the validator software. This includes understanding on how to produce files in such a way that they meet the interpreted ISO-standards in different cases, such as by creation, transfer, digitization and migration. This can be extended to consultancy services regarding how to implement system environment for the entire digitization and preservation process and develop or update digital preservation strategies.

## 6. NEXT STEPS AND CONCLUSION

Taking into account that the prototyping phase has just ended and that a functionally complete and stable version of the PREFORMA prototypes has been just released, the potential of the solution developed by the suppliers demonstrated to be of great interest for the community and considerable impact has been achieved already on memory institutions, open source companies and standardisation activities.

Initial ideas have been drafted for the exploitation of the results of the project and to bring to the market the software prototypes and they will be further developed in the coming months.

During the last year, the PREFORMA Project Coordinator Riksarkivet, on behalf of the whole Consortium, is committed to invest more resources to wider the emphasis on sustainability, impact assessment and innovation delivery, by actively engaging with the community, encouraging new users to download, test and use the tools, and setting up a plan for the long-term sustainability of the project's results.

One of the main lessons learnt during the third year of the project was in fact the importance of getting memory institutions more involved, which is a key element for the uptake of the PREFORMA tools and to establish a self-sustainable community. The success factor seems to be to take a step back and start discussing - and also train people in - what a conformance check actually is and why file format validation is so important in long-term digital preservation.

To this aim, the following activities have been planned:

- Preparation of a **survey** to evaluate the perception of the outcomes by the community and the benefits achieved so far, identify possible improvements and plan for the sustainability of the PREFORMA outcomes.
- Cooperation with LIBER (Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries), the main association of research libraries in Europe, to organise a set of **interviews** with its members.
- Organisation of an additional **workshop in Padua**, with a special session focusing on impact assessment and sustainability.
- Organisation of a series of **hands-on workshops and training sessions** in several European countries to explain to the participants what does conformance checking mean, how to create their own policy profiles and how to download, install, configure and use the conformance checker to analyse their files.
- Preparation of a **business plan** for each of the three open source projects to show how the tools can be maintained and brought into the market.

The impact assessment and exploitation survey was launched in November 2016 and it was primarily targeted to those members of the PREFORMA community who actively participated in the project, either as project partners or as external partners contributing to the project. At the end of December, we collected 33 answers from institutions and organisations all over Europe, 13 of

which have been collected during the Experience Workshop in Berlin, where participants were asked to fill in a simpler version of the questionnaire.

The survey is structured among the following three main axes:

- Benefits that PREFORMA can provide to the institution/organisation
- Perception of the PREFORMA outcomes and suggestions for their improvement
- Ideas for the sustainability and exploitation of the PREFORMA results

Among the benefits that have been highlighted so far by those who filled in the questionnaire, the most important ones have been:

- Involvement in the development and testing of the software prototypes
- Better knowledge of standards and file formats for long-term preservation
- Increased awareness of quality control procedures, especially how important is file format validation for long-term preservation
- Experience in working in a joint PCP
- Contact with institutions that have similar challenges and with companies which developed the tools

In terms of expected impact, what has been highlighted is:

- Improvement of the institution's digital preservation strategy and workflow
- Ensure well-formed and uniform access copies and preservation masters in a collection, that would result in fewer files that must be re-encoded (saving time), and prevent faulty materials from being preserved
- Widening the scope of workflow tools from market-based to open source solutions
- Increase knowledge and get a more safe and homogenous hands-on treatment and consistent thinking about long term preservation and the challenges involved
- Ensure files will be preserved in a standardised format and can be shared between institutions
- Ease the amount of work done by manual checking what hopefully now can be done automatically
- Increase reputation and relevance in the digital cultural heritage community by being part of the PREFORMA network

The majority of the people who filled in the survey already tested the PREFORMA tools, either by integrating them in their legacy system or by using them as a standalone service launched by the web browser. Interest was equally spread among the three file formats and a number of suggestions have been provided to improve them, e.g.:

- Cover additional formats, e.g. 3D data, PDF/X, PNG, Quicktime MOV, MXF, RAW, JPEG2000
- Improve documentation, user experience (ease of use for non-technical users) and possibilities for automation

- Improve metadata fixing capabilities and provide clearer indications what to do when a file does not conform
- Aiding novice users to understand what the tools do and how they fit and can be integrated in their workflows and systems
- Make the tools fully interoperable

Finally, in terms of plans for the future, the majority of the people think that it is worth to keep alive the PREFORMA network and wish to participate in the activities of the open source community after the end of the project. The activities that have been identified as the most important ones to be carried out by of the network in the future are:

- Looking at other possible file formats that might be considered as preservation file formats
- Improving the existing tools and working on interoperability with other systems
- Following the constantly changing needs of the memory institutions and the continuous evolution and improvement of file format specifications
- Maintaining available tools, material and documentation
- Acting as a network or hub of expertise and professionals facing similar challenges
- Seeking to develop further innovative and open applications for the sector
- Continuing to share the knowledge and the research being done throughout the network.
- Organising trainings/courses, technical and commercial presentations
- Continuing to work towards the standardisation of the preservation file formats and the improvement of the standard specifications

The full analysis of the results of the survey will be included in the updated version of this deliverable which is due by the end of the project.

In parallel to the impact assessment and exploitation survey described above, another questionnaire has been prepared in cooperation with LIBER. LIBER (<http://libereurope.eu/>) is the main network for research libraries in Europe. Founded in 1971, the association (a Foundation or Stichting under Dutch law from 2009) has grown steadily to include more than 400 national, university and other libraries from over 40 countries. LIBER works to represent the interests of European research libraries, their universities and their researchers in several key areas, e.g. Copyright and Open Access. It collaborates on European-funded projects and through events such as the Annual Conference it creates opportunities for library professionals to meet and learn from each other. The idea behind the cooperation with LIBER is to involve research libraries in the project, making them aware of what PREFORMA is doing and encouraging them to make use of the PREFORMA tools in their environment. Since the members of the LIBER association are not as familiar with the project as the member of the PREFORMA community, we adopted a different approach compared to the previous survey, including more general questions aimed at understanding:

- the way memory institutions acquire, store and make available their digital files,
- whether long term preservation is part of their mission,



- if they use specific tools for validating their digital files and checking file-format conformance against standard specifications or other policies established by the institution,
- whether they are using or planning to use the results of PREFORMA in their legacy environment and/or in their digital archiving and preservation workflows.

The results of this analysis will be presented in a special session during the Innovation Workshop organised by PREFORMA in Padua on 7 March 2017. Aim of the workshop is to highlight the importance of standardisation and file format validation for the long term preservation of digital cultural content, present the innovative conformance checkers developed in the project, and involve memory institutions outside the PREFORMA consortium in testing, using and further developing the software. The event will include live demonstrations of the software developed by the three suppliers (the veraPDF consortium, Easy Innova, MediaArea) and a session led by Melanie Imming, EU projects manager at LIBER, on impact and benefits of the use of the PREFORMA tools. This workshop is aimed at anyone interested in digital preservation and cultural heritage: memory institutions or other cultural heritage organisations involved in (or planning) digital preservation initiatives and willing to integrate the PREFORMA software in their infrastructure, the open-source community of researchers and developers interested in contributing code to the PREFORMA tools, the community of enterprises interested in developing services around the PREFORMA tools, the standardization bodies looking for feedback on how to improve and advance the specifications of the standard preservation file formats.

Finally, a series of hands-on sessions and tutorials have been planned to further promote the tools, improve user engagement and collect feedback from people testing the software in their own environment.

These workshops/seminars invite archivists/conservators/librarians to bring their files and analyse them with the PREFORMA tools. At the end of the workshop, they should understand the files and have a policy profile that allows them to check if their files are compliant with the acceptance criteria for their digital repository.

The first of such events will be organised in Padua in combination with the Innovation Workshop. Other sessions will be organised between April and July 2017. The list of potential venues include Amsterdam (focusing on AV), Riga (PDF and TIFF), Stockholm (PDF), Girona (TIFF), Brussels (TIFF), Berlin (TIFF), and Tallinn (all formats).

## **ANNEX 1: QUESTIONNAIRE ON IMPACT ASSESSMENT AND EXPLOITATION OF PREFORMA RESULTS**

<b>SECTION 1: IDENTIFICATION OF THE RESPONDING PARTY</b>
--

NAME OF THE INSTITUTION/ORGANISATION: .....

CONTACT PERSON: .....

E-MAIL: .....

DATE OF COMPILATION: .....

<b>SECTION 2: BENEFITS THAT PREFORMA CAN PROVIDE TO YOUR INSTITUTION/ORGANISATION</b>
---

In which sector does your institution/organisation operate?

(free text)

How did you learn about the PREFORMA project?

(free text)

Which are the needs that brought your institution/organisation to join the PREFORMA project/community?

(free text)

How do you expect the PREFORMA's results will impact on the workflow of your institution/organisation?

(free text)

Which are the main benefits that you gained so far from the experience within the PREFORMA project/community and its network of partners?

(free text)

Did PREFORMA contribute to increase the reputation and relevance of your institutions/organisation in your target community, and in particular in the community of (digital) cultural heritage? Please explain why.

(free text)

### SECTION 3: THE PREFORMA OUTCOMES

Which is your opinion on the quality of the information provided in the project's delivery channels, namely: PREFORMA website and blog, software released and information available in the Open Source Portal and in the suppliers' repositories, and demonstrations delivered at the project's public events?

Please, express your opinion with a number from 1 (very poor) to 5 (very good).

	Overall score 1=very poor 5=very good
PREFORMA website and blog ( <a href="http://www.preforma-project.eu">www.preforma-project.eu</a> )	
Open Source Portal ( <a href="http://www.preforma-project.eu/open-source-portal.html">www.preforma-project.eu/open-source-portal.html</a> )	
veraPDF website & GitHub ( <a href="http://verapdf.org/">http://verapdf.org/</a> , <a href="https://github.com/verapdf">https://github.com/verapdf</a> )	
DPF Manager website & GitHub ( <a href="http://dpfmanager.org/">http://dpfmanager.org/</a> , <a href="https://github.com/EasyinnovaSL/DPFManager">https://github.com/EasyinnovaSL/DPFManager</a> )	
MediaConch website & GitHub ( <a href="https://mediaarea.net/MediaConch/">https://mediaarea.net/MediaConch/</a> , <a href="https://github.com/MediaArea/MediaConch_SourceCode/">https://github.com/MediaArea/MediaConch_SourceCode/</a> )	
Presentations and demos at public events ( <a href="http://www.preforma-project.eu/events.html">www.preforma-project.eu/events.html</a> )	

Please, indicate which of the three media types covered in PREFORMA you are interested in.

- ☐ Electronic documents
- ☐ Still images
- ☐ AV files

Is there any other file format for which you would be interested in having a conformance checker, besides those already covered in PREFORMA (PDF/A, TIFF, Matroska/FFv1)?

(free text)

Which is your assessment of the quality and value of the three open source projects?

Please, express your opinion with a number from 1 (very poor) to 5 (very good).

	Innovation potential	Performance / Precision	Usability	Support / Documentation	Scalability / Extensibility	Interoperability	OS work practices	Commercial feasibility
veraPDF								
DPF Manager								
MediaConch								

Which are in your opinion the main strengths and weaknesses of the three open source projects?

	Strengths (free text)	Weaknesses (free text)
veraPDF		
DPF Manager		
MediaConch		

Did you participate in the evaluation of the software? How many users did you recruited? Please, describe the categories of users that you involved.

(free text)

Are the software developed meeting your specific requirements/needs? If not, what is still missing?

(free text)

Are you using or planning to use the results of the open source projects in your legacy environment and/or in your digital archiving and preservation initiatives?

- ☐ Integration of PREFORMA in the legacy system
- ☐ Use of PREFORMA as a standalone service launched by the web browser
- ☐ Other uses. Please, specify .....

Did you already make any experiment or proof of concept? Please describe it.

(free text)

#### SECTION 4: PLANS FOR THE SUSTAINABILITY OF THE PREFORMA OUTCOMES

Do you think that there is a value in keeping alive the PREFORMA network beyond the end of the project? Please explain why.

(free text)

Do you think that the PREFORMA network will or can provide benefits to your organisation/institution, after the end of the EC funded period? How?

(free text)

Will your organisation/institution participate in the activities of the open source community after the end of the EC funded project? How?

(free text)

Perception of PREFORMA outcomes so far and ranking your interest in future developments. On the basis of the information/material available, please indicate which are the most interesting outcomes produced by the project, i.e. those that are expected to deliver a stronger impact on your business/activities. For each outcome that you plan to exploit, please describe which are the plans to leverage on and sustain it in your organisation/institution.

Please, express your opinion with a number from 1 (not relevant) to 5 (very relevant) and with a free text.

	Impact 1=not relevant 5=very relevant	Type of exploitation (e.g. re-use for your corporate activities, re-packaging for the creation of new products, offering a service layer on the open products, etc.) (free text)
<b>SOFTWARE</b>		
veraPDF checker		
DPF Manager		
MediaConch		
Shell / Interoperability API		
<b>OTHER RESULTS</b>		
PREFORMA Challenge Brief		
Standardisation activity		
Test files and test cases		
Open source practices		
Web portal		
PREFORMA community		
Organisation of relevant events / workshops / initiatives / WGs		
Experience in joint PCP		
... Other (please specify)...		

Which is in your opinion the best approach for the exploitation of the results achieved in PREFORMA?

- ☐ First line – research: continue the research on formats, consider new PCP proposal to develop new services to make the archives more resilient (Work Programme 2019-2020)
- ☐ Second line – implementation: consider PPI proposal to deploy the services procured in PREFORMA and create an infrastructure which can serve different users with different needs (35% co-funded by the EC for deployment, installations, training, equipment, etc.)
- ☐ Third line – networking: create/reinforce the community of organisations and institutions interested to use PCP instruments in the cultural heritage field
- ☐ Other uses. Please, specify .....

Do you have any other comment / suggestion / idea that you feel it is worth to take into account when planning the sustainability and future exploitation of the PREFORMA outcomes?

(free text)

<b>SECTION 5: LESSONS LEARNT (RESERVED TO THE PREFORMA PARTNERS)</b>
--

Please list the significant lessons that you learned during the various phases of the project and their possible impact on the exploitation plan.

Phase 1. Preparation of the tender (e.g. legal issues, market analysis, collection of requirements, description of the challenge brief, timing of the tender procedure, evaluation of the proposals, etc.)

(free text)

Phase 2. Design phase (e.g. evaluation of the results submitted by the suppliers, completeness of the documentation, clarity of the presentations, communication with the suppliers, etc.)

(free text)

Phase 3. Prototyping phase (e.g. monitoring, planning of the releases, testing and evaluation of the prototypes, feedback, training, etc.)

(free text)

Phase 4. Testing phase (e.g. definition of the test classes, preparation of the corpora of test files, definition of the evaluation framework and methodology, etc.)

(free text)



Phase 5. Public events (e.g. format, duration, quality of the speeches, locations, etc.)

(free text)

Phase 6. Website and online communication (e.g. quality of the information provided, updates, interactivity with the audiences, targeted messages, etc.)

(free text)

## **ANNEX 2: IMPACT ASSESSMENT INTERVIEW MODEL**

1. In which sector does your institution/organisation operate?

2. Describe in general terms how your organisation acquires/receives digital objects.

- Does your organisation 'receive' objects or does it actively 'seek' objects?
- Which media type(s) do you acquire?
- What is the extent of the data sets you acquire (complexity, size, numbers, ...)?
- Does your organisation acquire use rights for the object?
- Which metadata does your institution record?
- How quick are objects ingested/processed after acquisition?

3. Describe in general terms how your organisation archives digital objects.

- How does your organisation store digital objects?
- Does your organisation make a backup of the digital objects (and metadata?)?
- Does your organisation regularly checks presence and integrity of digital objects?
- Does your organisation manage rights and obligations pertaining to the digital object?

4. How does your organisation make digital objects available?

- Through which online and offline channels do you make digital objects available for users?
- Which types of re-use of digital objects do you provide/allow?
- What formats do you use for access copies?

5. Does your organisation have a policy regarding the collection of digital objects?

- Does your organisation have a formal policy about the type of digital objects it acquires?

- Has this policy been written down? Does your organisation disseminate this policy among stakeholders?
- Does this policy allow for rejecting digital objects that do not fit the collection profile, i.c. conformance requirements?

6. Describe a typical acquisition/ingest procedure at your organisation.

- Identify which steps in the procedure are performed manually, i.e. decisions are made by a human agent.
- Identify which steps in the procedure are performed automatically, i.e. decisions are made by a software agent.
- Identify which of the following actions are part of the ingest procedure: Virus control - File digestion - File characterization - File validation - Metadata extraction - Persistent Identification – Cataloging – IPR management - File transformation - Backup

7. Describe the commitment of your organisation as to the storage and long term preservation of digital objects.

- Has your organisation formally determined requirements, significant properties and formats for the digital objects it preserves?
- Is long term preservation part of the mission of your organisation?
- Are responsibilities and budget for long term preservation sufficiently allocated?
- How does your organisation manage IPR violations, e.g. copyright violations, privacy issues, possession of illegal materials?

8. Describe the human agents involved in the ingest procedure.

- Identify name, role and profile of each human agent.
- Are all tasks concerning the ingest of digital objects designated to a particular person?
- Has this person the required technical skills to perform its task?

9. Describe the software agents (tools, web services) involved in the ingest procedure.

- Identify name, role and specifications of each software agent.
- How can external tools interface with software agents?
- Can your organisation open all files ingested?
- What tools does your organisation use for storage of digital objects?
- Which software agents are missing in the ingest procedure and would improve/complement the ingest of digital objects at your organisation?

10. Does your organisation use specific tools for conformance checking?

11. Describe the rules digital objects must conform to. (minimum conformance level)

- Do you have to conform by legislation with particular standard(s) or regulation(s)?
- Do you have to conform for technical reasons with particular standard(s) or regulation(s)?
- Do you have to conform with particular standard(s) or regulation(s) for any other reason?

12. Describe any restrictions as to what your organisation is allowed to check. (maximum conformance level)

- Are there any restrictions as to accessing the content of the digital object?
- Are there any restrictions as to reading or writing the bitstream of the digital object?

13. Describe what your institution actually checks when processing the digital objects. (real conformance level)

- For which rules do you have the tools and skills to do conformance checking?
- Where is the real conformance level located, in relation to the minimum and maximum levels?

14. Describe what your institution aims for as to checking the conformance of digital objects. (conformance level aimed for)

- What conformance level does your organisation aim for?
- What are the technical thresholds for reaching this level?

- What are the legal thresholds for reaching this level?
- What are the organisational thresholds for reaching this level?

15. Did you ever heard about the PREFORMA project? How?

16. PREFORMA develops three open source conformance checkers for electronic documents (PDF/A), still images (TIFF) and AV files (Matroska & FFv1). Is there any other media file or file format for which you would be interested in having a conformance checker?

17. Are you using or planning to use the results of PREFORMA in your legacy environment and/or in your digital archiving and preservation initiatives?

- ☐ Integration of PREFORMA in the legacy system
- ☐ Use of PREFORMA as a standalone service launched by the web browser
- ☐ Other uses. Please, specify .....

18. Which is your opinion on the quality of the information provided in the project's delivery channels, namely: PREFORMA website ([www.preforma-project.eu](http://www.preforma-project.eu)), the Open Source Portal ([www.preforma-project.eu/open-source-portal.html](http://www.preforma-project.eu/open-source-portal.html)) and the suppliers' websites, (veraPDF <http://verapdf.org/>, DPF Manager <http://dpfmanager.org/>, MediaConch <https://mediaarea.net/MediaConch/>)?

19. How do you expect the PREFORMA's results will impact on the workflow of your institution/organisation?

20. Do you think that there is a value in keeping alive the PREFORMA network beyond the end of the project?

- Do you think that the PREFORMA network will or can provide benefits to your organisation/institution, after the end of the EC funded period?
- Will your organisation/institution participate in the activities of the open source community after the end of the EC funded project?