

PROTOTYPING PHASE 2 INTERMEDIATE REPORT

Project Acronym: PREFORMA

Grant Agreement number: 619568

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

DPF Manager

Revision: [final]

Authors:

Miquel Montaner (Easy Innova)
Xavi Tarrés (Easy Innova)
Víctor Muñoz (Easy Innova)

Dissemination Level		
P	Public	X

1 INTRODUCTION

During the PREFORMA Prototyping phase, suppliers are expected to provide software prototypes that fulfil the requirements of the PREFORMA project, to demonstrate the results of their development work, and to provide explanations and documentation (manuals) on how the developed software can effectively be used in archiving scenarios at memory institutions regardless of their size and the file type they make use of.

Following the same approach used last year, during the Second Prototyping Phase the plan for releases is as follows:

- Frequent releases: monthly;
- Intermediate releases: end of July 2016 and end of October 2016.

The intermediate release shall contain two parts:

- A functionally stable release, if possible even more organised release compared with the respective predecessor versions
- A report which
 - Describes
 - More in detail the respective release;
 - The time line along with the current position (on time, delayed, ahead)
 - How suppliers managed to provide the required functionality (so far);
 - What is still missing compared to the original specifications and which is the plan to implement it.
 - Provides basic information to be used by PREFORMA WP8 in their deliverables to be submitted to the EC, reporting the work done by both suppliers and PREFORMA consortium members during the prototyping phase.

2 PROTOTYPING PHASE 2 - INTERMEDIATE REPORT

1. Details

Type of Organisation: SME

Registered Name of Organisation: Easy Innova, S.L.

Registered Address: Emili Grahit, 91 (Parc Científic i Tecnològic de la UdG)

Town/ City: Girona

Postcode: 17003

Country: Spain

Report Author: Miquel Montaner

Telephone Number: +34 972 41 88 54

E-mail Address: miquel@easyinnova.com

Project Name: DPF Manager - Digital Preservation Formats Manager

Report Type: Prototyping Phase 1 – Final Report

Total Contract Price [euro]: 699.475 €

Start Date: 14/04/2015

End Date: 31/12/2016

Sub-contractors: University of Girona (Spain), University of Basel (Switzerland), ID Law Partners (Spain), Bas Van Leeuwen (Netherlands)

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

In the October 2015 release the DPF Manager was able to validate TIFF Baseline 6, TIFF/EP and TIFF/IT ISOs, generate basic reports both in machine and human readable formats, and run a small set of simple policy rules. It was available for Windows, Linux and Mac OSX operating systems.

In the redesign phase ended in February 2016 we defined a set of tasks to be done to improve several components of the tool, and most important, we decided a redesign of the internal architecture of the project. All these tasks have been successfully implemented and included in the end of July 2016 release.

The project architecture has been restructured in order to follow a modular event-driven approach. We used Java Spring and JacpFX frameworks for building the GUI interface and we adopted the model-view-controller pattern for implementing the individual components. For the communication between modules, the JacpFX messaging has been used, which works through messages that are sent between the modules, firing the events.

This new modular architecture allows DPF Manager to be able to run checks in parallel, which is a new feature that has supposed a big step forward. The overall performance of the DPF manager has been improved by 500%, thanks in part to this multi-threading feature and also to efficiency optimizations developed in several core components. A task manager has been created to control the multiple running tasks through a built-in SQLite database, and a new widget has been incorporated to the GUI showing the list of running tasks, which can be paused, resumed and stopped.

A maven artefact has been created to allow external projects to easily integrate the DPF Manager into their solutions. This object encapsulates the whole DPF Manager in a single JAR file that is publically accessible in the maven repository. It can be included in any other java project, and the main functions of the DPF Manager (e.g. checking a file) have been published to allow external calls.

The implementation checker has also been rewritten, and now, instead of being hardcoded in the source, it is read from an XML file that contains all the rules to validate the ISO, and a rules engine executes it to perform the validation. This XML-based implementation checker allows an easier maintenance and extendibility of the checker. Also we have extended the baseline rules with the possibility to define warning rules alerting the user of illogical values that, although might not be explicitly mentioned in the ISO specification, would be non-sense. The policy checker of the DPF Manager has also received numerous enhancements, and now incorporates an extensive set of policy rules, which have been added from the feedback received from various memory institutions (such as blank page detection, extra channels checking, DPI and X-Y resolution coherency, etc.).

The configuration file of the DPF Manager, which contains the ISOs to be checked, the policy rules, metadata fixes and report formats, have been also rewritten in XML format to facilitate maintenance and to prepare it for interoperability purposes between conformance checkers. Finally, the generated reports have been improved and extended including more detailed information regarding ICC Profile, XMP, IPTC and EXIF.

Client-Server architecture has been implemented to allow DPF Manager to be executed in

server mode in one computer and client-mode in another computer, so that the client can send a request to the server to make a job and get the results back when it's done. This new mode has been also used to build an [online validator](#), which can be accessed publically in the DPF Manager website and allows any user to straightforwardly check their TIFF files online.

The console window that appeared behind the GUI has been eliminated, as suggested by the PREFORMA reviewers, and now the log messages appear in a console widget integrated into the GUI.

Multi-language feature has also been developed and now the DPF Manager interfaces (both GUI and CLI) are available in English and Spanish, with the possibility to be translated to any language. The language files are stored in a resources directory that automatically loads all the translations, thus it is very easy for contributors to add new translations into the project.

Periodical checks feature has been also implemented, so the user can define tasks to be run in different periodicities (daily, weekly, monthly) to check all the files in a given disk location and generate a report from a custom policy rules configuration.

Basic interoperability between conformance checkers has been deployed through a plug-ins folder that contains external checkers and can be called using the command line when needed, returning an XML report that is automatically appended to the output.

Our next steps include a big improvement on the XML report. After our research on standards for reporting, and also from the feedback received from Memory Institutions, we want to enhance the DPF Manager report including not only the validation results of a TIFF file, but also to show information of its structure and metadata in a language that archivists could understand. We finally decided to use the [METS format](#) for achieving this. The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed with XML. This standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation. METS standard can be used in the role of an Archival Information Package (AIP), Submission Information Package (SIP) or Dissemination Information Package (SIP) within the Open Archival Information System (OAIS) reference model, which ensures the DPF Manager integrability with Memory Institutions applying the OAIS model in their data object lifecycle. We also include in the METS report [PREMIS](#), [NISO](#) and [Dublin Core](#) data dictionaries, which all form a robust, standard, sustainable, integrable and interoperable report.

We also plan to use this METS report as the basis of the other report formats (HTML, JSON and PDF), by using XSLT transformations that will generate the other reports from an XML report.

Finally, we plan to improve the policy checker configuration, to allow users to generate custom rules by adding the capability of editing the single rules in Schematron format directly, or even editing the whole Schematron policy file.

2. Testing

We use several sets of images for different testing purposes. For the validation of the different ISOs (Baseline, TIFF/EP and TIFF/IT) we have formed sets of valid and invalid images for each standard. Some of the test cases have been manually created (edited with hexadecimal text editors) to simulate several possible errors in the TIFF structure and data.

The invalid test cases cover every possible error in the TIFF header, the IFD entries, tag values, and also within the image data, given the TIFF specification that we have revised in depth in order to create case examples for each possible violation.

Also, we have created new validation rules that, although not explicitly stated in the TIFF specification, would produce non-sense TIFF information, like tags out of place, illogical values, or invalid field definitions. In these cases, we have created warning rules (instead of errors) to inform the user that the data in the TIFF file is not coherent. We plan to create more of this kind of rules in order to inform for any inconsistency in the metadata of the TIFF file.

Unit tests have been created to run checks in all these files and compare the validation result with the expected output. These tests are run every time a build is made through the maven test plugin. This plugin runs all the tests and only creates the release (executables and installers) if they are all correct.

We also use the Travis continuous integration tool that runs all these unit tests every time a modification is committed to the source code repository, so it assures that the new features that are incrementally added to the project do not break anything that was working before. The Travis tool has been configured in order to run the tests in three virtual machines with different settings, so it checks the DPF Manager in CLI mode, GUI mode and client-server mode. Only when the tests are successful in all three modes a commit is considered correct. We also plan to create new virtual machines to perform tests in different operative systems such as Mac OSX.

The DPF Manager has been downloaded and used by some early adopters from different countries, which agreed to send us feedback from their tests. We received thousands of reports from their usage, which provided us with useful information on the usage of our tool, the most common type of TIFF files that are checked and the errors found. Also it has allowed us to discover some new private tags that we were not considering before.

Moreover, the integration of the DPF Manager in the Europeana Space project was successful and provided us some more feedback generally very positive. DPF Manager has been also extensively tested by Packed for the Tapiés Foundation with tests of several thousands of files, which have made us aware about the importance of parallelization.

As future plans we are working on creating a more extensive test set of TIFF classes, including some new handcrafted files and creating a new website of TIFF test cases. We want to build an evaluation platform to compare the DPF Manager with other tools and see the differences in the validation of the results and the information reporting.

3. Dissemination and community building

During the last period, we reinforced our strategy to create uncertainty to image archivists around their TIFF assets saying that despite their files be correctly generated, this does not guarantee that they are ready for digital preservation. The specification of TIFF is complex and some of its features are proprietary and therefore not suitable for long-term archival purposes. We have a lot of archivists and memory institutions waiting for the result of the TI/A initiative and therefore for the DPF Manager, the first tool that will validate the Tagged Image for Archival recommendations for long-term preservation.

The TI/A community has been build up around three online channels (and many offline communications), the TI/A website www.ti-a.org, the TI/A twitter account twitter.com/TI_A_Standard and the TI/A Intranet for the involved experts intranet.ti-a.org.

- The TI/A website has received 5.500 visits of 3.900 unique visitors in 15 months.
- The twitter account has almost 350 followers and we have published around 770 tweets related to the TI/A initiative and digital image preservation in general.
- In the TI/A Intranet we have 72 experts registered from 16 different countries.

To raise awareness amongst the scientific community, we have also published a white paper about the TI/A initiative, <http://ti-a.org/TI-A%20Whitepaper.pdf> (updated on March 2016) and we have presented the TI/A and DPF Manager initiatives in the following scientific conferences:

- De la Rosa, Peplluís, “Preservación digital como servicio (SaaS)”, **Congres d’arxivística de Catalunya**, 28th to 30th May 2015, Lleida (Spain)
- Fornaro, Peter, Rosenthaler, Lukas, “*Archiving Digital Image Data and Motion Picture: Concepts and Solutions*”, **International Conference on Advanced Imaging**, 17th to 19th June 2015, Tokyo (Japan)
- Fornaro, Peter, Rosenthaler, Lukas, “*Long-Term Preservation and Archival File Formats: Concepts and Solutions*”, **IS&T Archiving 2016**, 20th to 22th April 2016, Washington (USA)
- Fornaro, Peter, Rosenthaler, Lukas, “*File Formats for Archiving: Stability and Persistence Issues*”, **Digital Humanities Conference 2016**, 12th to 18th July 2016, Krakou (Poland)

As a result, the most important image researchers are closely following our initiative and waiting for the final technical specification that will be submitted to ISO, which will be also presented to the most important conferences.

We have to stress that to carry out this initiative we count with 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.

Besides to all the dissemination performed for the TI/A initiative, the advanced current status of the DPF Manager development allows us also to disseminate the features and advantages of this tool. During the last months, our communication campaign has been oriented to hook early adopters and, what it is more important, to get feedback from them. We obtain feedback from the automatic reports that we receive from the DPF Manager but also from direct communication with them. In this sense, the Open Workshop in Stockholm helped a lot on this.

Some of the early adopters are National Archives of Sweden, Aquaforest Limited, 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 and University of Pittsburgh.

From October 2015, the DPF Manager website received 7.445 visits of 3.320 unique visitors and the twitter account has over 130 followers from different countries in Europe, mostly archivists and developers belonging to memory institutions. Our @dpfmanager tweets have a total amount of 64.223 impressions from October 2015 to June 2016.

We have also released 7 specific newsletters (1 in October 2015, 1 in November 2015, 2 in April 2016, 2 in May 2016 and 1 in July 2016) containing only news around DPF Manager, and we have also disseminated the Open Source Workshop in a newsletter from [Blue Room Innovation](#). We also published several blog posts in the [DPF Manager Blog](#) mainly related to the new software releases, which have also been communicated via Twitter.

After holidays we will go on with our dissemination campaigns, first of all communicating all the new important features of the DPF Manager July release (explained in section 1) and then promoting the coming experience workshop in Berlin.

4. Open Source approach

Following the recommendations of the PREFORMA consortium received in the evaluation of the first prototyping phase, we started to use all the functionalities that our open source repository GitHub provides.

In our repository we use the well-known git flow methodology. Our master branch always contains the current code of our releases and the quick fixes created. All new development is done in the develop branch and each specific feature is a temporal branch pending on develop. When we submit a release we merge the develop branch with the master branch tagged as a release.

All the developments of the DPF manager are organized in [milestones](#), giving the information when a new feature or a bug fix will be released. All the enhancements and bugs are reported and linked to the corresponding milestone in the [issue tracker](#) section. So far, 149 issues have been closed in our repository.

All the DPF Manager releases are fully documented in the [GitHub](#) with all new features and bugs fixed.

All the bugs submitted in the GitHub are treated as high priority and always answered as soon as possible. Then, after evaluate the importance, we assign a milestone release and we correct the bug immediately in the master branch as a quick fix. See an example [here](#). In addition, when it is needed, we create a new test to prevent this bug again.

All the code is self-documented using the java doc standard. The java doc comments not only improve the code readability and facilitate the contribution; they are also used to generate the [API documentation](#) section in our website.

Now, with the most part of the redesign already developed, we would like to improve the documentation with more details about the architecture and how to extend the application creating new modules or conformance checkers. We also plan to create a new FAQ section in our web page using the forum.

All the DPF Manager information regarding the contribution is currently only available on the DPF Manager web site. We noticed that some traffic of our repository comes directly from GitHub. So, we are going to introduce all the information inside the GitHub Wiki where contributors could also edit and introduce new content.

All the code submitted is licensed with the required MPL v2+ and GPL v3+ licenses and all the test images are distributed using the CC Attribution-ShareAlike 4.0 International license. In order to guarantee that any contribution in our repository accomplish with our licence requirements, we are planning to introduce a Maven artefact able to check for any licence conflict and to generate a report with all the licences of our dependencies.

5. Standardisation efforts

As a reminder, we are working on the definition of a technical specification as a set of recommendations for memory institutions to preserve their TIFF files (called TI/A <http://www.ti-a.org>) to be published as an ISO recommendation.

During this period, we progressed in several fronts. First of all, on April 18th we had a meeting with Adobe in their headquarters in San Jose (USA) to follow the discussion we started in November during the ISO/TC171 committee meeting in Basel. We explained them that we followed their advice to work on an ISO recommendation instead of a new file format. We were very welcome at Adobe and they understood that we also have to take care on the existing assets in museums and archives. They will certainly help us with the plans and the discussions with ISO.

We already have produced a first draft of the recommendation based on the discussions with several experts inside the TI/A Intranet and our network within the community of memory institutions. Besides this technical approach we want to have a detailed picture of the assets already existing in memory institutions so that the recommendation does not imply unnecessary migration. We therefore have asked for access to “hot” image data of large institutions to do a deeply analysis of the structure of TIF files already archived by memory institutions. Like this we want to answer questions like: What are the most common used tags and values for these tags and which private tags are mostly used. The result of this analysis will give us a very strong basis and valuable information to finally decide the best rules to include in the technical recommendation.

In order to perform this analysis, we have organized access to about 2 Million TIFF files of three large Archives in Switzerland. We have set up an infrastructure in those archives because we are not allowed to take the data out of the building nor are we allowed to work on the hot systems. We have a set of Linux systems and larger NAS storage setup in situ. Then, we have developed a software that systematically is scanning each file for all (theoretical possible) 65.535 tags and write each tag found together with its value into a log file. This process is very time-consuming so we need some weeks to finish with this execution; we are still running it. Then, we are going to analyse these log files to extract all the information we need to proceed with the definition of the final technical specifications. The result is a feature histogram that is giving a good overview of the data out there. This analysis will be also very valuable to help us in the preparation of the testing phase.

Our aim is to finish the final specification by the end of September to share it with the TI/A expert's community and validate its content. Our next appointment is in Sydney the week of November 28th at the ISO/TC171 committee meeting, where we are going to submit the final version of the technical specifications of the TIFF format for Archival as ISO recommendation.

6. Impact assessment, sustainability, future use and exploitation

We have an important handicap regarding sustainability, future use and exploitation: the low technical knowledge about TIFF format leads memory institutions to be not aware about the lack of conformance of their TIFF assets to the standards and about the most probably unsuitability of their tagged image files for long-term archival. Therefore, our first activity has been to create awareness. The TI/A initiative precisely aims to build a set of recommendations for memory institutions certified by the International Organization for Standardization (ISO) with the aim of guarantee the long-term preservability of their TIFF files. This ISO Recommendation indeed would be our ultimate weapon to create awareness among archivists in the same way that PDF/A did some years ago.

Taking into account that DPF Manager will be the first tool to validate these ISO Recommendations, we ensure sustainability and future use of our software.

Memory institutions are increasingly understanding that they have to validate their TIFF files conformance to the standards and to the coming TI/A ISO Recommendations; therefore, some of them are already early adopters of the DPF Manager. Getting as many early adopters as possible is crucial for our strategy since, besides validating the new developments and becoming the first's clients of our services, they would be our apostles.

Some of the early adopters are National Archives of Sweden, Aquaforest Limited, 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 and University of Pittsburgh. We have to say that not all the users that download our tool are registered, thus we have a lot of unknown early adopters.

These early adopters allowed us to receive 2.526 reports of 5.603 files analysed (4.761 with correct baseline, 449 with correct TIFF/EP and 1.346 with correct TIFF/IT) where we could discover 18 private tags and 16 typical errors in the baseline.

It is important to mention that DPF Manager has been used intensively in two cases:

- Packed validated the conformance of around 40.000 TIFF files of scanned paintings for *La Fundació Tàpies* (<http://www.fundaciotapies.org/>).
- The University of Basel is analysing 2 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.

For us it is also very important the collaboration with associations or entities protecting the interests of memory institutions, as in the case of [KOST-CECO](#) in Switzerland, which is managing the digital preservation issues of their 30 members.

We are also collaborating with [Europeana Technical Space project](#), which is using our DPF Manager to validate the TIFF files uploaded by memory institutions. The reports generated from these analyses are really helpful for us to improve our tool.

The long term success of the DPF Manager depends on establishing a successful community around the open source project and also on developing a set of commercial services to ensure the project doesn't end up like JHove that died as soon as the public funding stopped. In the

documentation we submitted at the end of Phase 1 we included a brief business plan where we outlined that our 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. Although we have simplified the use of the conformance checker compared to our main competitor JHove, memory institutions would need the technical support and knowledge of a specialised company to integrate, maintain and evolve their long-term preservation systems. In order to offer our services in the near future, we already registered the domain www.dpfmanager.com.

At the beginning of 2016 we applied for an SME Instrument phase 1 in order to take the exploitation plans further. SME Instrument phase 1 is a grant for companies like us who need to develop feasibility studies of innovative products and services. The funding is to be used for the creation of detailed market and business plans, covering travelling expenses to meet and interview potential customers and partners, and any other activity that helps companies analyse the commercial potential of an innovative product/service. For us it is very important to have a very realistic and comprehensive business plan that will guide our strategy going forward. Our main goals are:

- To elaborate a market study to understand our customers (libraries, archives, memory institutions) and survey end users to estimate more precisely the demand.
- To define the best commercial strategies to reach the target customers
- To identify potential partners
- To make an exhaustive financial plan that will define a realistic roadmap for the exploitation stage.
- To develop a complete business plan.

Unfortunately our proposal was rejected, although we were very close, we received a total score of 12.63 with a threshold of 13. We want to improve and re-submit our proposal, thus your support is very welcome!

7. Gap analysis and next steps

As described in the final report of the re-design phase and in the technical specification of the DPF Manager, the first prototype was focused on the TIFF Conformance Checker. Then, during the re-design phase, we have internally re-designed the application from bottom to top focusing on the development of the Shell component.

In the first period of the second prototyping phase we have developed the redesign of the shell architecture (Tasks 1.1.3.1), we have provided a basic interoperability between conformance checkers (Task 1.3.1.2), we have completed the planned interfaces (Task 1.3.1.6) and we developed the scheduler functionality (Task 1.3.1.8). Regarding the TIFF conformance checker, we have changed the implementation checker module (Task 1.3.2.1) and we designed the new report (Task 1.3.2.3). All finished tasks are marked in green in the Gantt diagram below.

The only task planned for this period that has not been yet finished is the task 1.3.2.4 (marked in red in the Gantt diagram below). We completed the development of the reader for the IPTC and XMP information but we didn't finish yet the writer, which delayed the entire task. However, we are sure that we can overcome this inconvenience to have all the functionalities ready on time by the end of the second prototyping phase without affecting the global plan.

Now, we are working on the improvement of the TIFF/IT and TIFF/EP implementation, adding warning messages (Task 1.3.2.1.3 and 1.3.2.1.4), we are developing the new XML report (Task 1.3.2.3.2) as well as implementing new policy-rules (Task 1.3.2.2.1). We have also started defining TIFF classes for the evaluation phase (Task 1.3.2.5.1) and we are improving the DPF Manager documentation (Task 1.3.5). All the tasks under development are marked in orange in the Gantt diagram below.

The next steps in the development of the DPF Manager that must be ready by the end of this phase (marked in blue in the Gantt diagram) are the following ones:

- Complete the interoperability between conformance checkers (Task 1.3.1.2).
- Finish the XML configuration file, using and XSD standard, to be properly validated before being used by the program. (Task 1.3.1.4)
- The implementation of the TI/A standard validation, using the draft that will be submitted to the ISO working group. (Task 1.3.2.1)
- Finish the new report with the remaining formats JSON and PDF (Task 1.3.2.3).
- Improve the metadata fixer, being able to not only detect and correct incoherence's inside the metadata but also discover file transformations, not reported, in order to reconstruct the file provenance.

The final task is about evaluation and test (Task 1.3.2.5). As we explained in the end of redesign phase report, we want to create a reference public repository of test TIFF files to be used as a benchmark of TIFF reader/writer tools. We aim to prove that DPF Manager implementation can validate the ISOs implementations more accurately than the other tools. Therefore, we plan to build an evaluation platform not only for the DPF Manager but also for the other solutions and publish the results of this evaluation in order to disseminate the DPF

Manager.

