

VALIDATING FORMATS, A PREREQUISITE
FOR PRESERVING DIGITAL OBJECTS

THE PREFORMA HANDBOOK

Published by PREFORMA
September 2017, ISBN 978-91-87491-24-5

Printed by Grafitalia di Sandro Gherardini
Zona industriale La Fila, 56037 Peccioli, Pisa, Italy

Graphic design by Promoter SRL
www.promoter.it

The volume has been produced in the frame of the PREFORMA project.

PREFORMA is a project funded by the European Commission under European Union's Seventh Framework Programme for Research and Technological development (FP7).

Start date:	1 January, 2014
Duration:	48 months (end date: 31 December 2017)
Partners:	15 Partners from 9 European counties and a growing network of affiliate partners
Website:	www.preforma-project.eu
Showcase:	www.digitalmeetsculture/preforma-project
GitHub:	https://github.com/preforma
Email:	info@preforma-project.eu
Project Coordinator:	Börje Justrell, National Archives of Sweden, borje.justrell@riksarkivet.se
Technical Coordinator:	Antonella Fresa, Promoter SRL, fresa@promoter.it
Innovation Manager:	Claudio Prandoni, Aedeka SRL, prandoni@aedeka.com



The materials in this booklet are licenses under a Creative Commons Attribution 4.0 license. This means the documents can be re-used for any purpose, built upon and adapted – even for commercial purposes – when proper attribution is given to the authors. We kindly ask you to use the form of attribution outlines at the end of the individual document/s used.

ACKNOWLEDGEMENTS

To preserve digital objects involves a variety of challenges, which today are recognised by most parts of society. These challenges include issues such as policy and legal matters, intellectual property rights, and the need for metadata to explain and describe digital objects. But behind these issues, there are also substantial challenges on a practical level that affect the possibility of successful preservation. One is the file format of digital objects and the sustainability of these formats over time, which is the focus of this handbook.

Knowledge in this special field has started to be explored, but is usually spread out over the digital culture heritage community, often in a way that mirrors each institution's specific interest and commitments. Through the PREFORMA project, representatives of different stakeholder groups have come together and shared their expertise. They include: national and local memory institutions and other cultural heritage organisations engaged in digital culture initiatives; developers contributing code for open source tools; standardisation bodies maintaining the technical specifications of the preservation formats covered in PREFORMA; and other projects in the digital cultural heritage domain. By doing so, they provided invaluable contributions to this Handbook and to the project. Many thanks for that.

The PREFORMA project is also in the favourable position of having gathered together highly skilled professional partners in different areas: archives (national and local including special ones on audio-visual and film), libraries, applied research, digital preservation, project management, testing, and open source.

We would like to express our gratitude to all our partners in the project and in particular to Michelle Leamy from Libraries Development, Local Government Management Agency, who kindly reviewed the English language of the many contributors. You have been outstanding! Both the project and this handbook has been real teamwork.

Special thanks go to the European Commission for supporting us and to our Project Officer Manuela Speiser who has so neatly guided us through the challenges of being one of the first European pre-procurement project in the cultural heritage field.

We also want to thank the reviewers. Their advice and recommendations have really improved the outcome of the PREFORMA project.

Börje Justrell
Project Coordinator

Antonella Fresa
Technical Coordinator

Claudio Prandoni
Innovation Manager

TABLE OF CONTENTS

Executive Summary	6
Foreword	7
1. Introduction	9
2. The PREFORMA project	12
3. Making digital objects accessible and usable over time	24
4. Steps to consider when preserving digital objects	37
5. Conformance checking: a key to sustainable digital archives	47
6. The PREFORMA tools	56
7. Taking control of conformity tests process of digital files: an action plan	108
8. Conclusions	117
Back to the future? Digital preservation needs of future users anticipated today	120
Abbreviations	133
Authors	136

EXECUTIVE SUMMARY

This publication is intended as a practical guidebook based on lessons learnt and the explanation of problems to be addressed when planning the long-term preservation of a cultural digital archive, with particular regard to the phase of verification of the conformance of file formats stored in the archive. In order to provide a useful tool, the structure of the PREFORMA Handbook is conceived as a combination of critical considerations about the steps that must be taken into account by the archive managers, together with the guidelines on how to use the PREFORMA tools for running the actual conformance checking.

The Handbook aims to provide an overview of the background of digital preservation, the problems that PREFORMA addresses with its tools, how DPMManager, MediaConch and veraPDF contribute to solve these problems and, at a complementary practical level, information about the installation of the tools and the usage of the various functionalities of the software.

The publication closes with an afterword that looks at the future of digital preservation, to trigger the discussion about the new challenges, posed for example by the emergence of new formats and new content, such as 3D digitisations, augmented reality (AR) and virtual reality (VR) scenarios, linked data and geo-referencing.

We expect the PREFORMA Handbook to be offered as a critical instrument to decision-makers in cultural heritage institutions, to support them in the analysis of problems and the identification of viable solutions, and as a technical reference to managers of digital archives and developers, to offer them guidance on how to use the PREFORMA tools.

The PREFORMA project is a pre-commercial procurement project supported by the European Commission in the ambit of the Seventh Framework Programme for Research and Technological Development.

FOREWORD

By **Lars Ilshammar** (Deputy National Librarian, National Library of Sweden)
and **Rolf Källman** (Director, Head of Operational Support, National Archives of Sweden)

There are certainly those who claim that memory institutions don't have or shouldn't have an ideological agenda, that they are basically empty containers that can be filled with content and narratives society needs or finds reason to submit.

The key words, Open Access, Open Source, Open Formats, Open Science and Open Data indicate that while memory institutions have a distinct ideology represented most obviously by their everyday practices, they have an ideology of openness embedded deep in their instincts and DNA.

Memory institutions are of course the custodians of their own collections, but also of the common knowledge that these collections represent. In an era dominated by the forces of information protectionism, information economism and information consumerism this leaves memory institutions as the main and sometimes the only forceful defenders of an open public sphere.

Today information and knowledge are transferred into commodities or commercial services in large quantities and on a global scale. The process is fast and has been on-going for some twenty or thirty years with little or no opposition from policymakers. The great irony of the information age is that access to the public sphere is becoming more and more restricted when technology facilitates a free flow of information and knowledge for all.

In this context memory institutions need to develop strategies of resistance for their own long-term survival and also for the encouragement and benefit of their users. Open source software plays an important role in this effort as both a means and an objective.

Open source software and the communities that develop them can adjust the balance between openness and the forces of closure in favour of a more open information society. Openness is positive and encourages creativity and dynamism in research and in the private sector.

Commercial use of information and commercial software is dependent on a free and constant flow of data, knowledge and ideas. The success story of the Internet itself is evidence of the power of free software. PREFORMA is an excellent example of the successful results an open source projects can achieve in a central complex and complicated area as digital preservation.

1. INTRODUCTION

1.1. AIM AND KEY STAKEHOLDERS OF THE HANDBOOK

The PREFORMA Handbook primarily targets four key stakeholder groups with the purpose to support them to plan ahead:

- **Policymakers** at various levels that decide upon digitisation programmes and may promote the use of the PREFORMA tools in the digitisation process
- **Memory institutions** and other organisations who are involved in (or planning for) digital culture initiatives
- **Developers** coding and maintaining software for reading and writing files
- **Standards organisations** maintaining the specifications of file formats used for long term preservation

The aim is also to assist preservationists in all kind of organisations in their planning and decision-making for preserving digital assets long-term.



1.2. DOCUMENT OVERVIEW AND STRUCTURE

The PREFORMA Handbook is one of the main publications produced by the PREFORMA project. It gives an overview of the project and the contextual framework of digital preservation; it also comprehensively presents the tools for checking digital file formats developed by PREFORMA, and provides an action plan for the process of validating the file formats to be ingested in digital archives.

The document consists, excluding the **Introduction**, of seven chapters:

The PREFORMA project – This chapter presents the PREFORMA project, its structure, partners, aims and objectives.

Making digital objects accessible and usable over time – This chapter sets the context of digital preservation by focusing on four areas: (1) the main challenges in digital preservation, (2) definitions and strategies for sustaining the use and access of digital objects, (3) the functional framework (e.g. the OAIS model) and methods for analysing how to maintain accessibility and usability over time, and (4) the use and importance of standardised formats in digital preservation.

Steps to consider when preserving digital objects – This chapter gives a high-level overview of important aspects to consider in digital preservation, which are in line with the OAIS functional model (ISO 14721).

Conformance checking: a key to sustainable digital archives – a key to sustainable digital archives – This chapter highlights the importance of conformance checking of digital file formats, but also what conformance checking is.

The PREFORMA tools – This chapter presents the PREFORMA checking tools and how to use them.

Taking control of conformity tests process of digital files: an action plan – This chapter outlines the most important actions to follow, in order to ensure control of the conformity tests process.

Conclusions – This chapter summarises the discussions in previous chapters and suggests a condensed set of recommendations for the long-term preservation of digital archives.

The Handbook also provides an **Afterword** that discusses future changes in the digital preservation landscape.

2. THE PREFORMA PROJECT

2.1. GENERAL OVERVIEW

PREFORMA is a project funded by the European Commission (EC) in the Seventh Framework Programme for the Research and Technological Development (FP7).

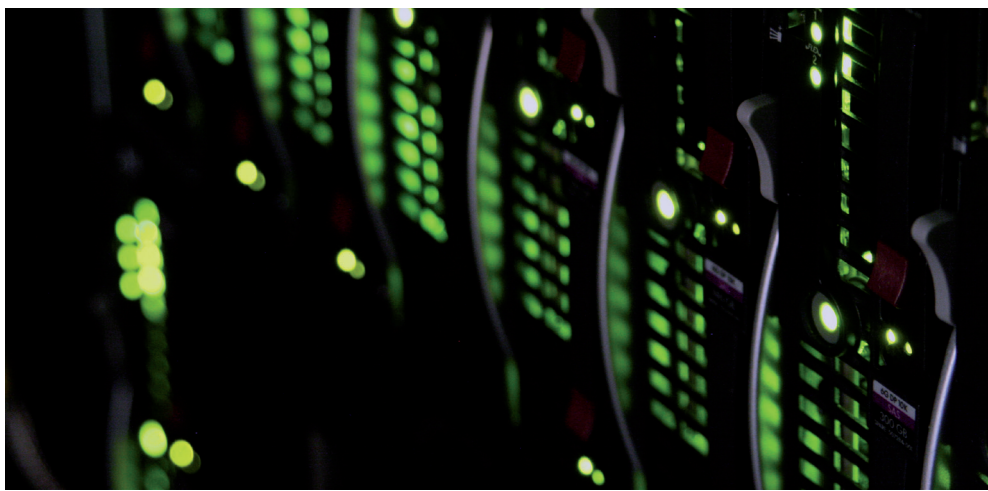
The project has been running from January 2014 for a duration of four years.

The scope of the project is to deliver a set of software tools that can allow memory institutions to check the conformance of the file formats implemented in their digital archives.

The tools are designed to allow their use either as a stand-alone online instrument, or as a service to be integrated in the legacy system of the memory institution via APIs.

Conformance checking is a key step in the preservation of digital objects. Even if solutions already exist for this purpose on the market, they are normally provided by the same provider of the software used to produce the electronic files. As a result, the production of these files is not controlled by the institutions that produces them or by the memory institutions who have the task to preserve them. This situation creates a condition of lock-in, which could be dangerous for the entire process.

13



Furthermore, the checking tools are generally based on proprietary solutions, where the access to the source code is not open, which is an additional weakness in the preservation process.

The media types covered by PREFORMA are documents, images and audio-visual contents and their respective formats, as illustrated in the table 1.

Media type	File Format
Documents	PDF/A
Images	TIFF
Video coding	FFV1
Sound coding	LPCM
Wrapper	Matroska MKV

Table 1: Media types and file formats covered by the PREFORMA project

PREFORMA aims to cope with issues connected with conformance checking by delivering a comprehensive and open solution, composed by:

- Open source conformance checkers
- Instructions for establishing and documenting the policies of the concerned institutions in terms of metadata, preservation workflow, etc.
- Collections of re-usable test files, covering both examples of conformant and non-conformant cases
- Full documentation of the checkers
- An API to allow interoperability of the checker
- A web-based application that allows to configure and to execute the conformance checks

- A lively community of users and developers who have adopted and are interested in the use and further development of the PREFORMA tools

2.2. PROJECT PARTNERS

A consortium made of fifteen partners from all over Europe is responsible of the successful completion of the EC funded PREFORMA project.

This consortium is complemented by the suppliers that have developed the IT solution and by a wide and active network of associate partners.

The fifteen partners include complementary interests and a very rich and multidisciplinary expertise.

Nine partners represent memory institutions active in the preservation of digital archives of documents, images and audio-visual contents. They are: the National Archives of Sweden (Riksarkivet), the Netherlands Institute for Sound and Vision (Beeld en Geluid), the Royal Institute for Cultural Heritage (KIK-IRPA), the Greek Film Center, the Libraries Development department of the Local Government Management Agency of Republic of Ireland, the Prussian Cultural Heritage Foundation (SPK), the Records Management, Archives and Publications Service (abbreviated SGDAP in Catalan) of the Girona City Council, the Ministry of Culture of Estonia, the National Library of Sweden (Kungliga biblioteket).

Three acknowledged and internationally well-known research institutions supervise the scientific value of the project. They are: the Institute of Digital Media Technology of Fraunhofer in Ilmenau, the Department of Information Engineering (DEI) of the University of Padua, and the Informatics Research Centre of the University of Skövde.

Three SMEs complement the cultural heritage and academic knowledge. They are: Aedeka S.r.l., Packed vzw and Promoter S.r.l..

2.3. PROJECT CONCEPT

Transfers of electronic documents or other electronic media content for long term preservation are continuously increasing.

Electronic documents and media content are made of two parts: the metadata and the data content. Metadata is often stored in XML and specified in different schemas. XML is a stable and easily accessible exchange format, and the schema specifications, like METS, PREMIS, EAD, are controlled by the community of professional curators in digital preservation through different international boards and committees. On the other hand, data content is normally stored in specific file formats for documents, images, sound, video, etc., depending on the originating system. These files also contain a certain degree of metadata and are usually produced by software from different vendors. However, even if the transferred files with data content are in standard formats, the implementation of these standards cannot be guaranteed. As mentioned above, the main reason is that the software used for implementation of standards for producing electronic files is not controlled either by the institution that produces them, or by the memory institution that holds the archives. As a result, the memory institutions must perform conformance tests but the lack of control on the software used to perform these tests can produce sometimes different results. Therefore, many institutions are using several testing instruments on the same files in order to get a valid outcome.

If the conformance tests do not provide positive results, the collection files are normally returned to the producer

for corrections, and the transfer process starts all over again.

This situation can cause costs to rise out of control. Furthermore, data meant for preservation, passing through an uncontrolled generative process, can jeopardise the whole preservation exercise. For example, migration of data files can be more difficult to carry out if the authenticity and integrity of the files are not guaranteed.

The Conformance checkers produced by PREFORMA will ensure that data content is produced according to standards, tested for conformity, and (if needed) re-processed for corrections. And all this happens within a process that is under full control of the memory institutions who are appointed to preserve the long-term electronic documents and other electronic media content.

2.4. PROJECT AIM AND OBJECTIVES

17

PREFORMA provides a full set of tools that can support memory institutions to address the challenge of implementing good quality, standardised file formats for preserving data content in the long-term.

In this light, firstly, the PREFORMA project aims to give memory institutions full control of the conformity tests process of files to be ingested into archives. The conformity tests process guarantees that the content is produced according to standards and, if necessary, the content can be re-processed for corrections. The PREFORMA tools enable this process to happen under the full control of the institutions. In fact, it is a basic requirement that memory institutions are trustworthy in performing their preservation programmes.

Secondly, PREFORMA aims to establish a long-term sustainable ecosystem around the developed tools. This ecosystem involves interested stakeholders from a variety of groups, including researchers, developers and memory institutions.

These two overarching aims are articulated into four specific objectives:

- **Development and deployment of the open source software** consisting of a set of tools (described below in Chapter 6). The tools are modular and validated against standard specifications used by European memory institutions for preserving different kinds of data objects i.e. documents, books, images and audio-visual records. In order to demonstrate effectiveness and to allow for refinement, these tools have been developed using an iterative process with multiple releases and with a number of experiments with 'real' data sets (files) from memory institutions during each iteration.
- **Setting up of the PREFORMA network of common interest** made of representatives from memory institutions, researchers and developers. They have taken part in the assessment of software tools during the development and deployment phases providing also collections of test files. These representatives are the base for a sustainable network that continues beyond the EU funded period, aiming at encouraging future use and development of PREFORMA tools and services, possibly also via new joint procurements.
- Taking a full, **open approach**, combining open file formats and open source software. This approach provides the necessary basis for long-term sustainable workflows through the integration of the software components, which are expected to be deployed in memory institutions either at the pre-system stage, or integrated into existing (legacy) systems already used. (In this case, we are assuming

that an API is available in the legacy system that allows for integration of external components).

- **Dissemination** of project's results to the wider cultural heritage community, including researchers and developers beyond the consortium, in order to enable them to benefit from this research and from the work of the project. A wide range of dissemination and outreach activities have been carried out by PREFORMA, including: online promotion of the project through websites and through specialised magazines (e.g. digitalmeetsculture.net); activity on social networks; international conferences and workshops; and presentations at public events. Such a comprehensive communication campaign encouraged many new memory institutions to express their interests and requirements, enlarging the network of users and reinforcing the impact of the initiative. Dissemination targets two audiences: memory institutions, to foster the adoption of the open source PREFORMA checkers; and technology providers, to foster their participation in the open source software project.

The achievement of these objectives has been pursued by the project very seriously and documented on the project's website that is open for consultation at the following Internet address: www.preforma-project.eu

2.5. PRE-COMMERCIAL PROCUREMENT

PREFORMA is a pre-commercial procurement (PCP) project co-funded by the European Commission.

PCP is an approach for public procurers to buy research and development services in a novel way. It is becoming

increasingly common within the public sectors of the European Union.

By acting as early adopters of the services and technology procured, public procurers can drive innovation from the demand side. This can improve the quality and effectiveness of public services and help create opportunities for companies to take international leadership in new markets.

The PCP approach enables public procurers to pool their efforts and resources in addressing a demanding technology problem. This approach allows for the share of risks and benefits. The procurers can design, carry out prototyping, and test a limited volume of new products and services with the suppliers, all with the aim of creating optimum conditions for a large take-up of the results by a wider community by the end of the EC funding period. The phase of PCP can be followed by the public procurement of innovative solutions (PPI), as illustrated in the figure below.

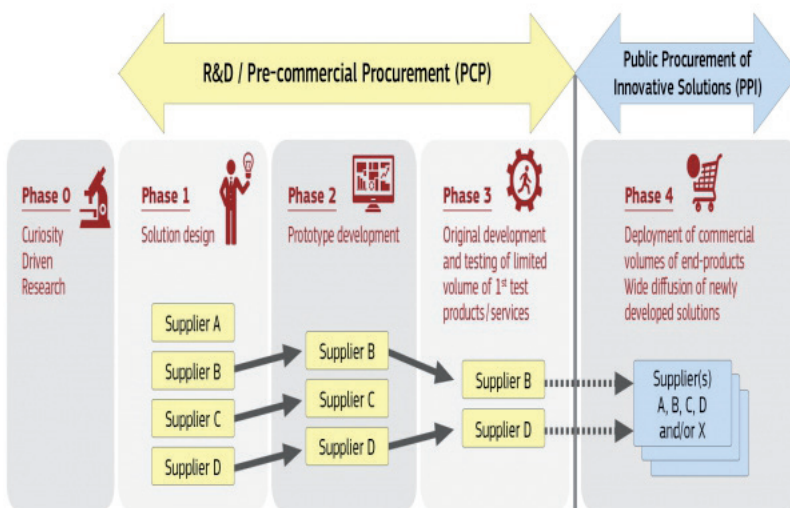


Figure 2: Innovation procurement phases. Source: <https://ec.europa.eu/digital-single-market/innovation-procurement>

As part of the initiatives to support innovation in Europe, the EC is reinforcing the policy framework for procurers to use PCP and PPI. To this regard, the following EC policy initiatives can be mentioned:

- Benchmarking national innovation policy frameworks. To this scope, a new study has been launched in 2017¹
- Public consultation on the interest of public procurers for innovation procurements of ICT based solutions, to inform the development of the Horizon 2020 work programme for 2018-2020²
- Promotion, training and provision of local assistance to public procurers that are interested in implementing innovation procurements

Dedicated funding instruments for PCP and PPI are implemented in Horizon 2020 and can be used across all areas of research and innovation.

The PCP instrument has demonstrated to be very appropriate for the support to the implementation of the PREFORMA tools.

21

2.6. OPEN SOURCE APPROACH

PREFORMA provides archival support for several open file formats. The choice of an open format is linked to the

1. Study on benchmarking the strategic use of public procurement for stimulating innovation in the digital economy across Europe (<https://ec.europa.eu/digital-single-market/en/news/study-benchmarking-strategic-use-public-procurement-stimulating-innovation-digital-economy>)

2. PREFORMA participated in the consultation. Summary of the findings of the consultation at http://ec.europa.eu/newsroom/document.cfm?doc_id=19671

scope of long-term preservation, where the concept of openness is very relevant. For each incoming file in any of these file formats, the PREFORMA tools provide technical support for checking conformance against the technical specification of the file format as published by the standardisation organisation, as well as conformance against the individual policies of each memory institution.

The conformance check of each file format is implemented as a separate open source software (OSS) licensed module that can be developed and deployed independently. The architecture for the implemented OSS component allows that one or several modules (even beyond the file formats addressed in this project) can be included when deploying the component. The software and each separate module is provided on the GitHub platform and licensed under a copyleft license that allows for use with software systems provided under different OSS and different proprietary software licenses.

22

The OSS component, with one or several modules, is integrated in a software system, which is deployed in a specific usage context at the memory institutions. A range of workflow pilots have been carried out during the project, at different memory institutions, to demonstrate that the software system can process files as expected in the usage context. The integration in legacy systems is ensured by the availability and appropriate functionality of APIs that allow for integration of externally developed OSS components. The deployment of OSS components at memory institutions can represent a benefit in two ways. Either, the institutions can adopt the new software using directly the OSS component and including it as a new step in their own institutional workflow, or the institutions can incorporate the OSS component in an upgraded version of their own existing archiving systems.

A large representative set of files normally handled by the memory institution has been used for validation of the software system in order to analyse effectiveness of the deployed OSS components. Two critical aspects of the

performance of each OSS component related to the different file formats were checked: that it accepts only files which are in conformance with the specification, and that it rejects only files which are incorrect according to the specification of the file format. Feedback to different stakeholder groups was provided based on an analysis of the outcome of the conformance test performed for each for each testing file. Furthermore, as part of the feedback, the OSS component provided details on precisely what was found as incorrect in each file. The overall purpose of the feedback has been to obtain a basis for continuous improvement of the effectiveness of the OSS component when used in the workflow.

The same process will continue after the end of the EU funding period for the sustainability and maintenance of the tools.

3. MAKING DIGITAL OBJECTS ACCESSIBLE AND USABLE OVER TIME

3.1. MAIN CHALLENGES IN DIGITAL PRESERVATION

Today, the significance of preserving digital information is well understood and accepted by society. Most commercial and public institutions, as well as many individual users of modern IT-technology, have learned their lessons; hardware and media obsolescence, lack of support for older computer formats, human error, and malicious software can all lead to loss of digital objects. If several of these factors are present, the higher the probability that such loss will occur.

The time frame of digital preservation is normally longer than electoral cycles, and it is questionable whether policymakers, funders, etc. realise the costs incurred when they fail to act. Therefore, a risk management approach should be advised as it allows digital preservation to be explained as an insurance and thus, makes it more understandable. UNESCO has also emphasised that digital documentary heritage is of critical importance for humanity as it has become the primary means of knowledge creation and expression.¹

25

Digital preservation is defined by the DigitalPreservationCoalition as a “series of managed activities necessary to ensure continued access to digital materials for as long as necessary”². In this section, the focus will be on what these activities are and the challenges they will bring on.

1. See <http://www.unesco.org/new/en/communication-and-information/memory-of-the-world/recommendation-concerning-the-preservation-of-access-to-documentary-heritage-in-the-digital-era/>

2. <http://dpconline.org/handbook/glossary#D>

The over-all characteristic of digital objects is their machine-dependency. Unlike traditional analogue objects (books, photographs, paper records etc.), where the user has direct access to the content, a digital object always needs a software environment to render it. Furthermore, it can only be accessed through a computer.

The key challenge in digital preservation is, therefore, the rapidly accelerating development in technology which will result in physical storage media, data formats, hardware, and software all becoming unavailable over time. This process, generally referred to as technology obsolescence, is regarded as the most significant threat to the continued access to and use of digital resources. The speed of changes in technology also means that the timeframe during which actions must be taken to safeguard future use of digital data is very short. Hard drives only last around five to seven years; a web page is forever changing; software is continuously upgraded; and there are not many machines left that read old storage media like floppy discs. If not earmarked for active preservation treatment at an early stage, digital objects will very likely be lost or unusable.

But digital preservation is not only concerned with sustaining single digital objects; to be used meaningfully long-term, digital objects need to be preserved in a context which makes them understandable and usable. Consequently, a set of non-technical requirements needs to be included when managing digital holdings and collections. This implies a need for policy decisions and the implementation of a sustainable preservation strategy, as well as practical solutions and tools to manage digital data.

Another challenge for digital preservationists to consider is the rapidly growing amount of digital information worldwide. The number of transfers of digital objects to memory institutions is expected to increase, which will induce higher costs for them.

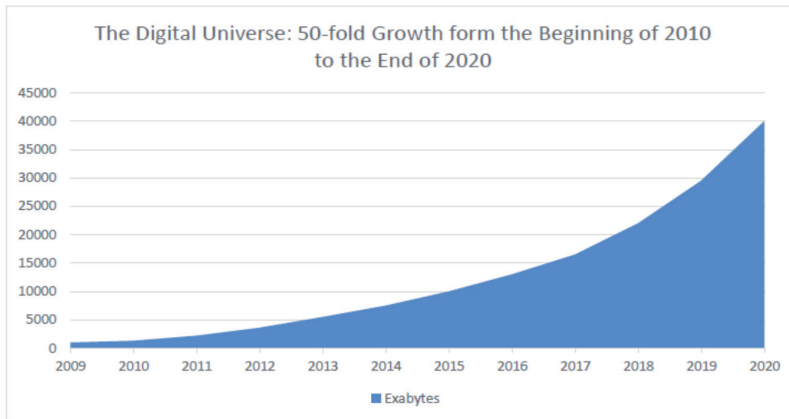


Figure 3: The expanding digital universe. Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

The economic challenges of digital preservation are already substantial. Preservation programs require significant upfront investment, and along with ongoing costs for data ingestion, data management, data storage, and staffing, the situation can easily become unmanageable, especially for those institutions who are receiving digital objects but lacking specific processes and systems for treating them long-term. One of the strategic issues to highlight when trying to raise funds for preservation programs is that, while requiring significant resources, the benefits of the programs will be obvious to future generations.

There are also challenges with digital preservation that relate to the differences between digital and paper-based material. Not only is the form of the objects different: there is a different way of working following the introduction of digital objects. This has forced memory institutions to integrate new concepts, methods and tools for digital preservation to be carried out in parallel with traditional analogue preservation. But many memory institutions in Europe are still in a limbo-like situation: they have neither implemented internal strategies and operational solutions for long-

term digital preservation, nor safeguarded their digital holdings and collections according to international standards for digital archiving. The reasons may differ, but the results could be devastating for future use of their digital objects.

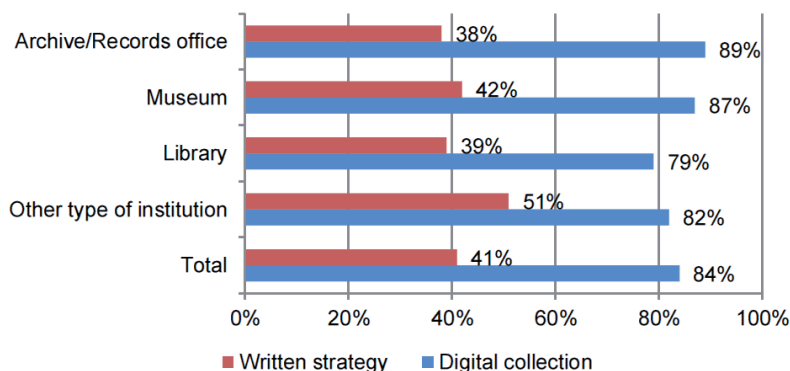


Figure 4: Institutions in Europe with a digital collection and a written strategy for handling it. Source: *The challenges of digital preservation as a public mission*, key-note speech by Marco de Niet, Digital Heritage Netherlands, at the PREFORMA Innovation Workshop in Padua, 7 March 2017.

A new trend, partly behind the figures for “Other type of institution” in Fig: 4 above, is that privately managed digital information services have started to challenge the public institutions to reassess their mandates, strategies and information services.

3.2. STRATEGIES

From what is said earlier in this chapter, it is obvious that the preservation of digital objects requires a radically different approach compared with preserving

analogue materials. This is further confirmed by the fact that the greatest asset of digital information: the ease with which it can be copied or transferred, is paralleled by the ease with which the information can be corrupted or deleted. The nature of the digital technology presupposes a life-cycle perspective in handling the maintenance of digital resources long-term. A continual programme of active management is needed - from the design and creation stage of the system and onwards.

Two to three decades ago, the focus was mainly on finding the 'ideal' digital media for long-term storage. Since then it has moved on to weighing the advantages and risks of different digital preservation strategies, and to define practical solutions based on standards that may use several strategies concurrently. But a good model for digital preservation is not only to use a well-defined strategy or a combination of different strategies, it also requires a well-defined long-term vision outlining what needs to be preserved. In order to be successful, the strategy often requires to be implemented step-by-step.

Today, there are several strategies available for sustaining the future use of digital objects. The main ones are³:

- **The techno-centric strategy**, which aims to preserve original hardware and software in a usable state into the future. It involves regular storage media renewal to make sure that the physical digital objects are not corrupted.
- **The incremental change approach**, which relies on two rather different strategies: either migration of digital objects into new formats or preserving the formats of the digital objects and using emulation to be able to use them.

3. The text below is mainly based on the paper Digital Preservation Services: State of the Art analyses by Raivo Ruusalepp and Milena Dobrova (2012)

- **The migration strategy** normally uses standardised file formats, which are repeatedly converted to keep up with present technical generation.
- **The emulation strategy** preserves the original file formats and uses emulation at alternative levels to overcome technical generation changes. Either the original software, the original operating system or the original technical platform are emulated into the new technical environment. In the latter cases, it is combined with preserved original software.
- **The analytical strategy**, which is currently based on techniques used in computer forensics. The underlying logic for this strategy is to apply specialised methods for recovery of objects which are in demand in the future instead of 'mass preservation'. The basic idea is that this does not seem realistic, having in mind the volume of digital information involved.
- Another strategy seeks for methods that change the formats of the digital objects in a way that allows the objects themselves to invoke preservation actions. Such objects are sometimes called **durable digital** objects.

The first three strategies require rigorous organisation of processes; the fourth one is still under development. All these strategies outline the principles of preservation. In practice, they are implemented within archival lifecycles that integrate various tools and/or services. These lifecycles can be specific to organisations, depending on organisational mandate, the types of object they hold, and their target users.

Of the strategies mentioned here, the migration strategy has for a long time been the dominant one. Combined with the OAIS model - see below - it is used by most institutions working with digital preservation. Standardised file formats are normally used for the digital objects to be preserved and to avoid technical obsolescence, they are



converted to new standardised file formats at the point of technical generation changes. These conversions are expected to be carried out without information loss. In the foreseeable future, the migration strategy will probably remain as the most common strategy, at least for in-house preservation. However, taking a longer perspective, increased use of distributed preservation services like outsourced e-Infrastructures may change this situation.

Regardless of the chosen strategy or combination of strategies, memory institutions often make a distinction between the master version of digital data and at least one surrogate dissemination version. The master version should contain as much intellectual, visual or audio content as possible, be saved in a standard (non-proprietary) file format, and preferably be duplicated across multiple locations. Dissemination versions of data may be re-sized, compressed, and saved in whichever format that is suitable for public exposure.

Sometimes, this distinction between different versions of data can be a part of a so-called flight-ahead strategy called proliferation. The leading idea behind proliferation is that the more copies of the file available, the better its chance of survival. Digital files can easily travel around the world via networks, and proliferation is a strategy to deliberately spread files over different repositories and backup services. Even if this is an efficient way to combat the risk of

losing a file, there is a potential risk of losing track of the authentic file.

Using proliferation as a strategy can also give an incorrect feeling that “everything can be found elsewhere”, which may cause digital preservation programmes to be regarded as “low hanging fruits” easy to pick off in discussions on budget cuts.

3.3. THE OAIS MODEL AND THE ANALYSIS OF PRESERVATION LAYERS

32

Although the underlying process in digital preservation could be described as universal, the diversity of both digital objects and types of memory institutions that are responsible for the preservation of digital resources creates variations in the level of tools used in practice. The central standard in the preservation domain is ISO 14721 Space data and information transfer systems – Open archival information system – Reference model, widely known as the OAIS model. It is a functional framework that presents the main components and the basic data flows within a digital preservation system. It defines six functional entities that synthesise the most essential activities within a digital archive: ingestion, preservation planning, archival storage, data management, administration, and access. Recently, some major European libraries have proposed to combine these six stages into a smaller number of use-cases that preservation systems address.

The OAIS model looks at data stored in the digital archive as fluid objects that can (co-)exist as three types of information packages:

- Submission (SIP), used to transfer data from the producer to the archive
- Archival (AIP), used for the archival storage and preservation
- Dissemination (DIP) used within the access function when consumers request archived materials.

As a reference model, the OAIS standard does not imply a specific design or formal method of implementation. Instead, it is left to users to develop their own implementation by analysing existing business processes and matching them to OAIS functions⁴.

The OAIS standard also consists of an information model. This information model describes the information packages consisting of the actual data object and the different types of metadata which may be needed to preserve the object. The metadata is structured and described and may concern process, content and context. But part of the metadata is technical and in gathering these metadata the PREFORMA tools may be of great use.

33

4. The text about the OIS model is taken from the handbook "DCH-RP: A Roadmap for Preservation of Digital Cultural Heritage", produced by the EU financed project DCH-RP.

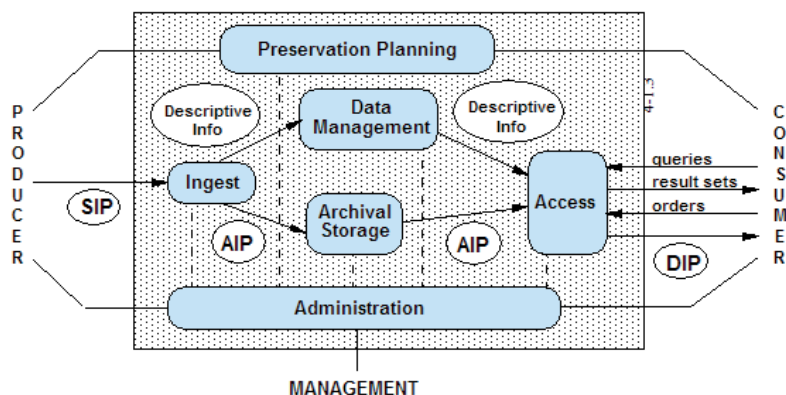


Figure 5: The OAIS model

In maintaining the accessibility and usability of digital objects over time, an often-used method for analysing them is built on the presumption that every digital object consists of three layers: a physical, a logical and a conceptual layer. All three layers and their relations must be considered and understood to get proper preservation actions. These actions are often identified and referred to as "bit preservation", "logical preservation", and "semantic preservation".

Bit presentation covers several basic actions ensuring the integrity of the 0: s and 1: s (the sequence code) over time and serves as the ground pillar for any other preservation actions.

Logical preservation focuses on the representation of the digital object, and activities in this field have the aim to ensure the quality of being able to retain the object and maintain accessibility over time. File format is of course one major issue here. Over the years, much effort has been spent on setting up requirements and recommendations for file format sustainability and PREFORMA aims to make a significant contribution to logical preservation

Semantic preservation includes activities focusing on long-term understanding of the content, but also on capturing contextual information about the domain/environment in which the digital object was created. Here, part of the OAIS metadata is of considerable importance⁵.

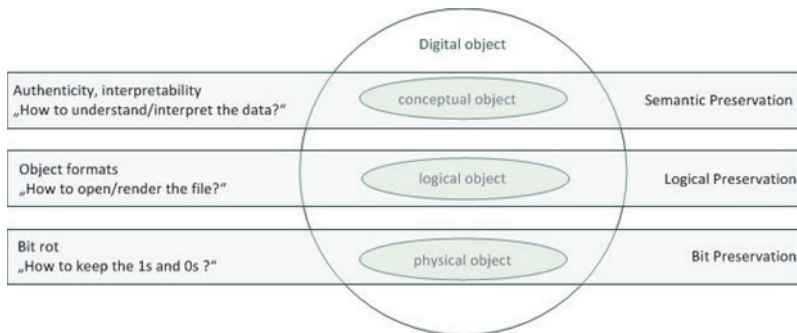


Figure 6: The layers of a digital object

35

3.4. THE USE OF STANDARDISED FORMATS IN PRESERVATION

As pointed out in the previous section, the use of digital object formats that are reliable for archiving is an important part in digital long-term preservation. In this case, the reliability means minimum changes to the information and maximum foreseeability of what will happen when performing future migrations and conversions of software and hardware. The use of reliable formats is not just necessary for keeping unreliable formats out of

5. See EU project DURAARK, deliverable D6.6.1 Current state of 3D object digital preservation and gap-analysis report, <http://www.duraark.eu/deliverables>

the storage facilities, they also reduce the number of formats for the digital curators to deal with.

But even reliable formats need to be migrated or converted into future formats to avoid becoming obsolete sooner or later, and having less file formats to deal with means, of course, less trouble. The key point is that to be defined as reliable, a file format must pass a process of development, acceptance and implementation based on the consensus of concerned parties in society; in other words, a file format must pass a standardisation process. The outcome, i.e. the standard, contains specifications of the file format that explain the 'key' to the translation of colours and other properties in the bit-code, for example letters and sound.

To assure that a file could be treated in the same way by every implementation (now and in the future) that follows a standard, the file format needs to be tested against its standard specification. The possibility to foresee what will happen is both a matter of security and a basic condition for transferring files over time and space, which is crucial for the preservation of the file, but also for the trustworthiness of data in an e-government context.

Data objects are normally stored in specific file formats for documents, images, sound, video etc. and these files are usually produced by software from different vendors. The problem is that even if these files are in standard formats, and should be reliable, the correct implementation of standards cannot be guaranteed. The main reason is that the software used to produce the files is not in control either by the institutions that produces them or by the memory institutions that must preserve them long-term. If the tools used to generate the files do not follow exactly the standard specifications, these files can show different properties, although they have identical formats, and therefore act differently when converted or migrated in the future. This can be devastating for the long-term preservation of the files.

4. STEPS TO CONSIDER WHEN PRESERVING DIGITAL OBJECTS

The process of preserving digital objects long-term includes several steps, each of them containing various action points. What to accomplish in these steps varies, but some of the steps are more fundamental to the entire process than others and therefore, need to be considered in more depth. These are:

- Selecting the data to preserve
- Keeping the data alive
- Retaining the meaning of data
- Maintaining trust in data
- Keeping the context and dependencies of data
- Establishing and maintaining good governance of the data

38

4.1. SELECTING THE DATA TO PRESERVE

What needs to be preserved, and what can be preserved? The question of selection arises given the huge amount of digital resources that are produced and waiting to be preserved. Memory institutions may have a public commitment to preserve (including legal deposits and public archives duties), and privately financed bodies may have specific commitments based on the aims and ambitions of their owners. But in the end digital preservation will always be a matter of value, contemporary or future, and what is important to some owners of the data and their end-users is not necessarily important to others.

The expanding rate of data creation is followed by an expanding need for data storage, which places challenging pressures on selection policies and other organisational

decision-making. Simple preservation processes that function at a basic level may not necessarily scale easily to handle large volumes of data files. Although, the technology to work at scale is improving, some digital repositories still face significant challenges in developing and maintaining scalable architectures and procedures to handle growing quantities of data. Data selection, often followed by data appraisal and disposal, are important components in any digital preservation process.

In the cultural heritage field, criteria for selection and disposal of digital objects are often in place, but differ between domains and professions. Diverse types of data and digital objects also require diverse types of preservation methods and activities. As an example, analogue data carrier converted into digital ones by digitisation and “born-digital” objects are often treated differently, mostly because digitised objects may be needed during a shorter period and sometimes are considered only as digital copies of physical objects.

Another aspect of the selection of data to preserve, is that different countries may have different rules about preserving digital data. For example, it is not always possible to preserve valuable data outside a country or on a server owned or controlled by a commercial service provider.

39

4.2. KEEPING THE DATA ALIVE

A digital file is built up by a series of zeros and ones, called binary digits. The normally used symbol for binary digit is bit (recommended by the IEC 80000-13:2008 standard). A group of eight binary digits (bits) is commonly called one byte.

To ensure the survival of a digital file, the stream of bits needs to be captured and retained over time, without

any loss or damage. Consequently, the preservation system in use must have the ability to store the file in such a way that it can be retained with full accessibility and usability, but also with its authenticity guaranteed.

A systematic process for bit preservation is, therefore, a basic requirement. This process could for example contain the following activities:

- Monitoring and refreshing of storage media
- Creating redundancy by designing systems that duplicate components to provide alternatives (like replicating or backing up files) in case of component failure
- Introducing strategies to replace obsolete technology with recent technology and aiming at diversity in used technologies to avoid lock-in dependencies
- Avoiding storage at a single geographical location
- Generating checksums to be frequently recalculated to identify any loss and ensure the integrity of the bits

To orchestrate these activities, the preservation system in use must have an ability to provide curation services like:

- Schedule-based integrity checking
- De-referencing and deleting
- Migration of (and the possibility to move) preserved files to new versions of software and/or hardware
- The possibility to export data
- Conversion and transformation of data

4.3. RETAINING THE MEANING OF DATA

To reconstruct the information, encoded as a stream of a bits within a digital file, requires software that is designed to render, manipulate, analyse or otherwise interact with the format in which the data is encoded. But over time, there will be new versions of the file format and the format itself will also run the risk of becoming obsolete in the sense that the software applications it interacts with may no longer be supported.

Therefore, it is important for the digital preservation curators to understand the technology on which digital materials are dependent, as it enables them to take appropriate actions to safeguard their preservation process. A preservation plan is the normal instrument in which these actions are specified; this might be the migration of digital files from format to format, the emulation of obsolete software, or the use of alternative software applications to interpret the data. Each of these options, including others not mentioned here, have both advantages and disadvantages and need to be evaluated carefully, keeping in mind the cost aspect and expected end-user needs.

41

4.4. MAINTAINING TRUST IN DATA

For an end-user to maintain trust in the results of a digital preservation process, it requires careful consideration of the entire life cycle of the preserved material, including who or what has interacted with it over time. The key objectives are to keep and maintain the authenticity and integrity of the digital objects. The information management systems administering

the preservation process must be able to link to the essential contextual information, describing the business procedures of the organisation that has created the digital objects.

The application of data integrity techniques and the maintenance of audit trails can provide users with confidence that a digital object has remained unchanged since deposit in an archive (with necessary preservation actions as the only exception). However, authenticity to an end-user may depend more on the broader trustworthiness of the preserving organisation. Transparency in the organisation's processes for the maintenance of high-quality preservation, which shows that these processes are based on current best practice and validated by appropriate audit and certification, may be crucial in this context.

42 Issues concerning authenticity and integrity of digital resources are also important in sectors not focusing on digital preservation. For example, academics need to feel confident that references they cite will stay the same over time; courts of law need to be assured that material can withstand legal evidential requirements; government departments may have legally enforceable requirements regarding authenticity; and so on. This issue overlaps with both legal and organisational issues and it may be best resolved within each individual sector, rather than through generic procedures.

4.5. KEEPING THE CONTEXT AND DEPENDENCIES OF DATA

In understanding the meaning of a digital object, a future user may be dependent on additional information, which was obvious when the object was created or used, but less clear later. Therefore, to have sound procedure in place for identifying and capturing relevant contextual

information can be crucial for the achievement of successful preservation results.

Digital objects are created in a manner that is increasingly complex and complicated. What looks like simple stand-alone objects could in fact, to be fully understood, be deeply dependent on related information, software or data sources at various locations on the Internet.

To understand digital objects, their context and dependencies on other sources is a vital component.

4.6. ESTABLISHING AND MAINTAINING GOOD GOVERNANCE OF THE DATA

43

Not only technological issues can be challenging, there are also numerous challenges relating to governance issues. These include how digital preservation is organised and delivered, and how those responsibilities change over time, as well as the life cycle of digital objects. Yet every organisational context will be different. There is no “one-size-fits-all” approach for digital preservation; each organisation must decide upon its own governance model and tailor it to the its unique conditions and requirements. However, some general issues need normally to be tackled:

PRIORITISE PRESERVATION ACTIVITIES

How to prioritise digital preservation activities and applying them in a timely manner can be crucial, not just to avoid losses, but also to ensure the best use of resources. To intervene early in the life cycle of a digital object can determine whether or not it will

survive into the future. Small investments upfront when choosing trusted file formats, capturing critical documentation or the description of key relationships in metadata can deliver considerable savings further down the line.

On the other hand, choosing the appropriate actions is a delicate matter, which preferably should be decided on a case-by-case basis. Early interventions to avoid technological obsolescence may provide greater confidence in long-term sustainability, but are difficult to orchestrate and can result in resources being wasted. Therefore, a just-in-time approach is preferred as it minimises unnecessary activities. The disadvantage is that this kind of actions normally requires a knowledge specialist who is not always in place.

IN-HOUSE OR OUTSOURCING

44

The decision whether to do all or part of the digital preservation via a third-party or in-house, or as a combination of the two, is often a complex one. Digital preservation may be undertaken in-house if there is sufficient staffing and infrastructure available, but outsourcing some activities or support can also be cost-effective.

Of critical importance, when choosing an outsourcing model, is to have and retain sufficient knowledge to be able to prepare effective specifications and monitor performance. Outsourced work must be easy to verify and quality check, and this is best provided through a considered design of the specification, and through the reporting providing by the third party. Cost will clearly be a key consideration when deciding whether to contract out digital preservation activities, but there are also other factors to consider such as legal issues on privacy or confidentiality, which may influence whether outsourcing is appropriate or not. The advantages and disadvantages of each option will need to be balanced considering the individual organisation's mission and responsibilities.

HANDLE ORGANISATIONAL CHANGE

The modern digital world is characterised not only by rapid technological development, but also by organisational changes. Organisational bodies re-organise, merge, and/or change ambitions and owners. Digital preservation is a long-term activity, and the likelihood of it being affected by organisational changes will increase over time. This will happen not only through changes to its parent organisation, but also through changes to its major depositors, users and other stakeholders. Therefore, organisational change is a major risk to be managed when maintaining good governance.

BALANCING SECURITY AND ACCESS

There is a strong natural link between preservation and access; the latter normally being the main objective for the first. Repositories need to ensure that their digital objects are kept safe and secure, but must also provide access to a variety of users. The aspect of access often provides valuable input when an organisation is designing its preservation facilities. Some digital objects selected for long-term preservation may contain confidential and sensitive information that must be protected and not accessed by non-authorized users. There may also be legal or regulatory obligations on a repository that can affect the possibilities for access. An important part in establishing and maintaining good governance is to find a balance between security and ease of access.

45

LEGAL COMPLIANCE

Legal issues could be complicated when dealing with digital preservation. Digital objects are generated by a wide group of creators and incorporate more diverse formats and intellectual property rights (IPR) than objects in analogue form. Often, the delays and gaps in law relate to regulating technological changes and digital

preservation needs, which need special observations. Some of the key legal issues that affect collecting, preserving, and providing access to digital objects are:

- legal requirements concerning management, preservation, and access placed on the repository and its parent organisation by donors and funders or via legislation by Government
- legal obligations relating to third party rights in, or over, the digital objects
- legal elements of relationship between a repository and any third-party provider or providers (e.g. terms of service contracts and service level agreements).

5. CONFORMANCE CHECKING: A KEY TO SUSTAINABLE DIGITAL ARCHIVES

The problems observed by the memory institutions in the existing tools for producing and validating files of a specific format have different causes. There are cases when the specification of the standard format is not used correctly; in other cases, the specification is not precise enough. In a longer perspective, the solution is to recommend the SDOs to review the standard specifications to gain more precise definitions; in the short run, it is necessary to enable the memory institutions to check the conformance of the files to be preserved.

5.1. THE NEED FOR CONSISTENT PROPERTIES

48

Any set-up that systematically shares items over a certain period of time needs to be able to (1) ingest items, (2) provide a minimum level of description and (3) store them in such a way so that they can be found and redistributed. But set-ups differ significantly depending on how long it is necessary to keep these digital assets retrievable and reusable. The longer this period is, the more these set-ups need to address the potential technical obsolescence of digital assets.

The projected timeline of memory institutions is traditionally without a defined end. Whereas the horizon for information technologies is usually limited to approximately five years, the timeline for archives and libraries is far beyond the lifetime of any technology. Planning for a limitless future can seem an implausible undertaking. Nevertheless, memory institutions are held accountable for their long-term preservation tasks and they are forced to make decisions for the future with the technological means of today. This challenge is not unique for memory institutions. Universities, courts of law and the medical profession are all increasingly producing all sorts of information that ought to remain

accessible beyond the lifetime of the media types used to capture them.

To some extent, basic life cycle management and digital asset management procedures considerably help making digital assets future-proof. But the real challenge lies in extending or even transcending the life cycle of the technology underlying these digital assets. A straightforward strategy consists of simply replacing obsolete technology with contemporary technology without losing the intelligible content of the digital asset, usually referred to as 'migration'¹. It may seem a simple strategy, but in practice, it comes with some difficult decisions. After all, it involves changing and replacing bits and bytes while guaranteeing that the end-result looks the same. So, when are digital assets obsolete enough to take the risk and migrate these assets? And how do you verify whether the migrated digital object is sufficiently 'the same'?

Moreover, many organisations with a long-term vision, particularly smaller cultural heritage organisations or institutions with other primary concerns, have little control over the files they guard and rely heavily on market-conform technology and/or manufacturers. Although in the production chain, archival demands (e.g. providing a basic level of metadata) have moved to the fore, in the IT and media industries this shift has not happened in the same way. Cultural heritage organisations have few choices, and generally little experience, in voicing their demands when it comes to setting requirements for long-term sustainability of file formats. However, any migration attempt, which aims to be successful, requires precise knowledge of the technical properties of the digital assets to ensure the preservation of the intelligible content. Every organisation should make sure it maximises its confidence in their digital assets by knowing their format inside out.

49

1. See more about different strategies for digital preservation in section 3.2 above

PREFORMA aims to support organisations gain control over the technical properties of their digital assets. The project starts from the presumption that limiting the number of file types and versions of one file type makes migrating these files a less hazardous task. The more homogenous the collection, the simpler the preservation task becomes. Homogeneous collections allow for bulk monitoring and bulk solutions.

Conformance checking of the format of the files in a digital archive is a key step in the process of preservation of the content of the archive. The OAIS reference model (see section 3.3 above), uses the term ‘validation’ in two diverse ways: (1) regularly checking internal procedures in the OAIS to ensure that all functions work properly and (2) regularly checking the integrity of files processed by the OAIS using checksums. Validation, as referred to in this handbook, denotes the process of checking if the technical properties of a digital file conform with the specifications of the corresponding file format. This particular meaning of validation is only indirectly referred to in the OAIS specification, not coincidentally in the context of file migration. The objective of validation, as conformance checking, is basically to harmonise the syntax and semantics of the file structure, based on a file format specification. Normalised files enable mass transcoding with consistent results and enable the identification of proper tools to perform the transcoding.

50



Conformance checking is particularly relevant for memory institutions because they ingest files from a variety of sources. Even though these files have the same file format, they can be very different. These differences between files tend to originate from differences in the software the files are made with. When developers write software that reads or writes files using a particular file format, they typically refer to a human readable specification that defines how the file should be structured. But human readable specifications inevitably contain flaws and ambiguities, and different interpretations lead to files with technically dissimilar files. Of course, developers are aware of this. However, for a developer, the issue whether a file is valid or not depends on the file being processed successfully by their system, rather than whether it complies to a particular human readable specification. Their job is to write software that facilitates a particular workflow and avoid dissimilarities between files, which could distort the proper functioning of the system.

So typically, production systems, and to great extent their users, are pragmatic about files coming from various sources and systems; as long as the file is processed by the system and looks convenient for the end-user, there is no issue. But organisations with a long-term view on reusing their digital assets do have an issue dealing with these dissimilarities. Transcoding files from one format to another, without losing intelligible properties of the content, requires unambiguous information about the technical properties of the file.

There are several reasonable explanations why file format specifications contain flaws and ambiguities. The most common cause is a lack of time among developers to document the format. Documentation always happens afterwards and usually there is only time to provide a general outline of the format. Systematic documentation of all design choices made in the design process is rare. Developers who reuse the specification for other purposes usually must deal with missing features, which they then develop themselves. At best, such additions flow back to the original specification or a daughter

format splits off. But most often such additions remain undocumented, which causes software to produce files with properties that are not described in the specification. Particularly in the case of proprietary formats, access to the detailed specification is licensed to earn back the initial investment in the format and to exert a form of quality control software developed by third parties. But it also allows to control the end-of-life of the file format (so-called planned obsolescence), which pushes the market to adopt newer file formats.

But even when file formats have been well specified, it may be challenging for curators of digital assets to find out how a file is put together. Analysing files in an asset management system obviously is an automated process and requires tools that translate the human-readable specification in machine-readable validation rules that can be processed. Though, for the clear majority of the files maintained in systems, these tools do not exist or they are not publicly available.

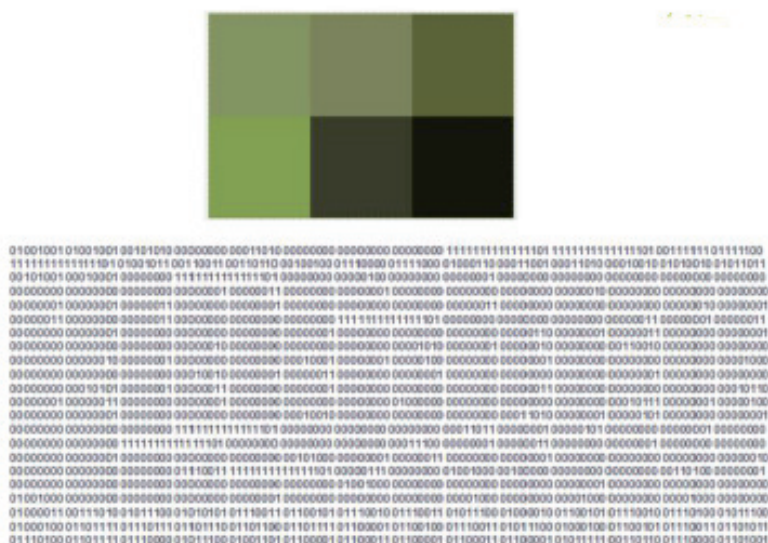
52

All these technical thresholds come on top of the factual observation that memory institutions lack the in-house resources and expertise to assess the quality of a specification or the proper implementation of a file, let alone the expertise to assess if a validation tool performs well. This level of understanding of file validation goes well beyond today's practice in cultural heritage institutions, which basically consists of basic file identification.

The basic conclusion is that to preserve digital object requires institutions to develop a strategy, make decisions and deploy tools to implement these decisions. The strategy will generally be a combination of approaches that take into account the specificity of the archive (e.g. the type of documents, the life duration of the documents, ownerships and rights, and many other characteristics) as part of a long-term vision, within a step-by-step programme.

Any preservation strategy that relies on migration to overcome file format obsolescence (see section 3.2

For example, we can say a lot about colours and resolutions of images, but we hardly understand how they are encoded in digits. The following figure 7 represents an image encoded in a TIFF file and it demonstrates that the comprehension of the sequence of bits is rather difficult for a non-specialist.



Furthermore, we can find the properties tab in the browser, we can read how these properties are declared, but it becomes then difficult to find out if this information is also 'true' i.e. if the data in the file corresponds to

the properties declared in the metadata associated with the content of the file.

Naturally, the specification of a file format explains how colours, resolutions and other properties are transformed into a sequence of bits, but the implementation of the transformation rules described in the specifications into the actual software application that generates the file is subject to the interpretation of the developers. Then, different interpretations can produce different results, which eventually can generate ambiguities when the file is read by another software application, and may even cause the document to become unreadable.

Consistent properties are important for digital archivists. It is essential to know how the files are generated, in order to:

- Ensure authenticity of the content
- Improve efficiency in monitoring and maintaining large collections of digital documents
- Be ready for migration and emulation when needed, at large scale, by applying the same procedures to the widest number of files at the same time

The PREFORMA tools are designed primarily to empower memory institutions to gain full control over the technical properties of digital content intended for long-term preservation.

5.2. WHAT IS CONFORMANCE CHECKING

Conformance checking is the process that controls if the properties of a file are conformant with the standard specifications.

Conformance checkers are the tools that implement this control. They investigate the digital document and extrapolate its technical properties, in order to establish if these properties are compliant with the specification and, consequently, if the digital file is well-formed or corrupted.

Because of its nature, conformance checking is based on an initial selection of the file formats to be considered, namely to define which kind of content and which kind of standard are considered in the checking procedures.

The digital archives held by memory institutions contain documents, images and audio-visual content. The PREFORMA conformance checkers can process files representing all these different content types.

With regards to the standard specifications, two considerations have been taken into account.

The first consideration is that different stakeholders are involved in the life cycle of digital files: on one hand, there is the industry that provides the tools to produce the digital files and on the other hand, there are the memory institutions that are responsible for preserving the digital files.

55

Furthermore, PREFORMA opted for a consistent, open approach, which led to the choice of open standards, for three types of content (see section 2.1 above), namely: PDF/A for documents, TIFF for images and a combination of FFV1, LPCM and Matroska MKV for the audio-visual content.

6. THE PREFORMA TOOLS

6.1. OVERVIEW OF THE TOOLS

PREFORMA research and development activities aim to empower memory institutions to gain full control over the technical properties of preservation files. This is achieved through the development of three open-source toolsets for conformance checking of digital files (electronic documents, still images and audiovisual files), intended for long-term preservation in memory institutions.

All the conformance checkers developed in the project are presented in more detail in the upcoming sections. They include the following basic components:

- **Implementation Checker:** validates compliance with the specification in all respects.
- **Policy Checker:** enforces custom institutional policies beyond the scope of the specification itself.
- **Reporter:** produces customisable reports formatted for both human readability and automated parsing.
- **Metadata Fixer:** carries out any corrections to file metadata, if necessary, to achieve conformance with the specification.

The conformance checkers interact with other systems through a “shell”, which allows for using multiple checkers at the same time. This enables the integration the integration of conformance checkers from different suppliers into one application.

The development of the tools started from the definition of four use cases that aim to facilitate the interaction between the supplier, academic research and memory institution. These use cases are based on the OAIS Reference Model (ISO 14721:2012) and represent conformance

checking procedures at different moments in the life cycle of a preservation file:

- Conformance Checking at **Creation Time**: Producers pro-actively check if the technical properties of a file meet the acceptance criteria of a digital archive.
- Conformance Checking at **Transfer time**: Archivists check the technical properties of the files to be ingested, assessing whether they meet their acceptance criteria and conform to the relevant standard specification for that file format.
- Conformance Checking at **Digitisation time**: Archivists check the technical properties of digital representations of collection items, internally or externally produced, assessing whether they meet the requirements specified in the digitisation tender.
- Conformance Checking at **Migration time**: Archivists check the technical properties of files that are repackaged or transcoded, following the rules defined in the preservation strategy of the digital archive.

The main benefits offered by the PREFORMA tools are:

- **Preservation Standardisation**: Digital files may deteriorate over time, if stored improperly and without sufficient back-up copies. In such cases, through the PREFORMA tools, archivists can ensure that their files are interoperable and conform to established preservation standards by checking that they adhere to format specifications. This checking will allow for revitalisation of potentially corrupt files, prompting an archivist to correct errors within the file when necessary or create new preservation-level files when feasible.
- **Migration to other formats**, such as accessible copies for the web or switching to another preservation format in the future, is also

facilitated by the checks performed by the PREFORMA tools. Possible interoperability issues or broken files can be detected and corrected as early as possible.

- **OAIS Adherence:** The conformance checking of files helps archivists adhere to the OAIS reference model. Files can be checked for conformance upon initial digitisation, upon ingestion into a digital repository, upon migration to different locations, and during regular quality control and check-ups, as defined by the implementing memory institution. The Reporter component integrated in the PREFORMA tools may be used to gather the needed technical metadata described in the OAIS model and further specified in standards like PREMIS, MIX, etc.
- **QA Expansion:** Quality control can be better monitored through the PREFORMA tools by algorithmic detection of conformance errors as well as the supplemental institution-based policy conformance checks. Since files are checked in a systematic way, preservationists can know definitively whether the file is working or how the file has changed since the last time it was reviewed (whether that is from previous quality analysis or during digitisation, ingestion or migration). The result increases the usability of files, while also maintaining constant contact between the producer and consumer of content, once again adhering to OAIS standards.
- **Streamlining & Customising Routine File Check-up Process:** Conformance checking, with the PREFORMA tools, allows institutions to perform routine file check-ups in a streamlined process. Such check-up standards can be altered according to shifts in standards regarding preservation-level quality, whether established internally or externally. The PREFORMA tools allow for potential users to establish their own policies, choosing files based on personal conformance preferences, which can be altered to fit specific, situational needs.



60

The conformance checkers are all available both for local use (through a GUI or a CLI) and as a web-based application. They allow also the deployment in different infrastructures and environments: as a standalone executable; as a client-server application to be deployed in network-based solutions; or as a plug-in/library to be integrated in third party systems (legacy or future ones) via API.

Ultimately, the three tools are provided to all interested organisations as an Open Source web portal hosting the developed code (<http://www.preforma-project.eu/open-source-portal.html>). 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 ecosystem. In particular, all software is released under the GPLv3+ and MPLv2+ open licenses and all digital assets are released under the Creative Commons license CC-BY v4.0.

The following sections present an introduction to the three tools, explaining what they are and how to use them. It must be noted that the instructions on how to use the GUI and the CLI are based on the prototype releases available at the time of printing this booklet (Summer 2017). To get access to the most up-to-date information, readers are invited to consult the PREFORMA website, where source code and documentation is available through the PREFORMA Open Source web portal.

6.2. VERAPDF: INDUSTRY-SUPPORTED PDF/A VALIDATION

INTRODUCTION: PDF AND PDF/A

As “born digital” documents eclipse the scanned materials in contributions to memory institutions, those organisations are besieged with a wide variety of file-types, with more appearing on an irregular basis. Validating file-format integrity is an important step in preserving digital content for the long term. It helps memory institutions identify potential issues in their collection at an early point in their digital preservation workflow. If a file is valid, this means that it can be rendered by a valid PDF reader in the future.

One simple fact has emerged from the recent history: PDF is the dominant electronic document format for deliverable documents worldwide, and is thus one of the dominant file-formats received by memory institutions. The core value proposition of PDF is simple: reliability when shared. PDF files must be fully portable – entirely self-contained – as well as flexible and capable. Accordingly, PDF technology is undeniably complex.

Since 1993, when the technology was first introduced by Adobe Systems, PDF’s specification was published and available for royalty-free use. But the PDF specification is not a recipe for a “good” PDF, but more a cataloguing of functionality. Over time, the flexibility and rich capability of PDF technology facilitated diverse applications meeting the needs of a wide variety of market segments. Without a validator, these specifications are occasionally interpreted differently by different vendors, sometimes causing customer confusion.

PDF/A, the archival specification for documents in Portable Document Format, is an ISO standard that establishes the technical basis for determining that a given PDF document (irrespective of scanned or electronic source) possesses archival qualities, i.e. that it is properly coded, and includes the resources necessary to long-term usability.

PDF/A was introduced in 2005 as a means of providing software-developers with a standardised means of:

- Creating archival PDF documents, either from existing PDF documents or source-files.
- Verifying the actual conformance of files claiming to meet archival requirements for PDF.

Validating PDF documents for long-term archival using ISO 19005 (the PDF/A standard), is a challenging task for a number of reasons:

- PDF is very different from XML syntax, where schema validation has been an established technology for a decade.
- The PDF format is highly flexible and extremely complex.
- Apart from the 1,000-page ISO standard defining PDF, there is no other formal description of the format.

Since PDF's reliability is at the core of the format's value proposition, standardised creation and processing of PDF data structures is of value to any PDF developer. That's where veraPDF comes in.

WHAT IS VERAPDF ABOUT?

veraPDF is an industry-supported open source validator and authoritative test-suite for PDF/A, the ISO standard for long-term preservation of PDF documents. veraPDF has been developed with collaboration between industry software developers, led by the PDF Association, and

memory institutions, represented by the Open Preservation Foundation and the Digital Preservation Coalition.

veraPDF provides memory institutions with a broad range of authoritative quality-management capabilities that include, but are not limited, to PDF/A conformance. In conjunction with an authoritative test corpus, veraPDF represents a formalisation of validation rules to ensure precision. Over the 18 months of development, the PDF Association PDF Validation Technical Working Group (TWG) met monthly to review questions arising from the test-corpora and software development.

veraPDF's architecture is that of an abstract validation model consisting of an object-oriented hierarchy of object types to be validated. Each object type contains a predefined inheritable set of simple properties, as well as named links to lists of objects of other types. A validation profile lists all requirements for each object type, or validation rules in formal terminology. Each rule is a certain Boolean expression built from the object properties, elementary arithmetic, and Boolean operations.

63

This approach is designed to be as generic as possible, and is, in fact, not specific to PDF at all, thus gaining the following advantages:

- veraPDF fits the purposes of PDF validation and, in particular, is aligned with the internal PDF syntax. As such, it may be readily extended to validate other PDF subset standards such as PDF/E (Engineering), PDF/X (Printing) or PDF/UA (Universal Accessibility), or the PDF specification itself.
- The veraPDF model is readily employed for validating other file formats in digital content such as ICC profiles, images and fonts.

veraPDF is written entirely in Java, and includes desktop GUI and command-line interfaces (CLI).

veraPDF assists memory institutions archive digital content at several levels by providing:

- For institutions receiving PDF files, authoritative information on file quality in terms of specific failures to meet PDF/A's specifications.
- For institutions receiving PDF/A files, positive identification of PDF/A conformance as well as documents that conform to PDF/A's specifications in all but metadata.
- For institutions receiving PDF/A-3 files, identification of embedded "associated files", and those files' declared relationship(s) to the container PDF/A-3 document.
- For institutions recommending use of PDF/A to their contributors, a means of advising contributors on their software procurement and archival-preparation workflows.
- For institutions using acceptance or assessment policies other than PDF/A, a means of identifying conforming and non-conforming documents based on PDF features.
- For contributors to archival institutions, a means of authoritatively assessing in-use and proposed software, identifying problems with outputs from existing workflows.

Institutions can integrate veraPDF into their regular input-processing workflows, and use it to identify files that fail (or pass) PDF/A. This capability, as well as veraPDF's rich machine and human-readable reports, allow for true knowledge, and therefore management, of received documents.

By leveraging veraPDF's ability to identify characteristics of PDF file outside the scope of PDF/A (such as, for example, restrictions on image-type, specific fonts or the presence of annotations), institutions can utilise

their own policies outside of PDF/A to manage their collections.

By utilising veraPDF's capability to optionally add or remove PDF/A metadata to otherwise-conforming PDF files, based on validator results, institutions can enhance their holdings (when justified) as meeting ISO standards for long-term preservation.

Through veraPDF's ability to detect embedded files and extract associated metadata, memory institutions can begin to leverage new archival applications potentially including source files, associated data files, email mailbox files and other non-PDF electronic material intended for archive.

Memory institutions will find that with veraPDF their ability to handle and manage PDF documents is greatly enhanced, creating new options for managing high volumes of electronic documents.

In addition, veraPDF can reduce the amount of exceptions in existing and new workflows, eliminating a potential source of costs.

65

INSTALLING VERAPDF

The veraPDF installer package (available at <http://www.preforma-project.eu/verapdf-download.html>) is a zip file that contains:

- The Java installer and application as a single jar file, `verapdf-izpack-installer-<version>.jar`.
- A Windows batch file that runs the installer, `vera-install.bat`.
- A bash script that executes the installer on Linux or Mac OS machines, `vera-install.sh`.

The installer jar file includes the application binary files and supplementary resources, including:

- Validation Model description.
- Test PDF Documents.
- The veraPDF Validation Profiles.

The installer simply unpacks components from the installer package to the local computer.

More information about the installation process is available at <http://docs.verapdf.org/install/>.

VALIDATING PDF/A FILES WITH THE GUI

The veraPDF GUI provides the features of veraPDF PDF/A Java Library in a desktop windows GUI. Users can configure their own PDF/A validation and policy checking jobs by selecting which:

- Combination of tasks to perform.
- PDF Documents to analyse.
- PDF/A part and conformance level to test for.
- Task specific settings to choose.

The software carries out the configured task and reports the results in both XML and HTML formats. The XML report is intended for consumption by automated processes, while the HTML report is designed for human readability.

The application installation folder contains the script that shall be executed to launch veraPDF Desktop GUI application. The script name depends on the platform:

- On Mac OSX and Unix systems: `verapdf-gui`.
- On Windows systems: `verapdf-gui.bat`.

When the application is started the following screen is displayed:

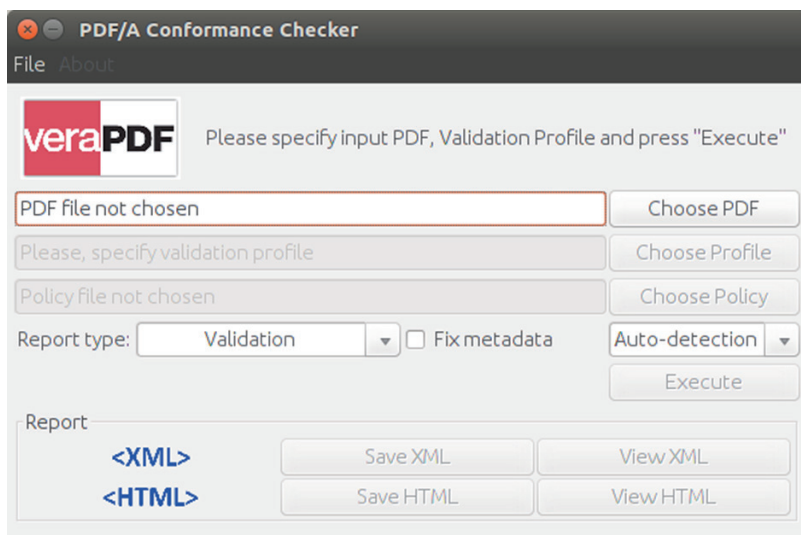


Figure 8: veraPDF GUI, starting screen

Choose PDF opens the file dialog, allowing the user to select a PDF file for validation or feature reporting. The Execute button will not be enabled until a file is selected.

The Choose Profile button and Profile dropdown is used to control validation processing. The user can:

- Select any of the built in PDF/A Validation Profiles.
- Allow the software to select a Validation Profile by analysing the PDF/A document metadata.
- Choose to load a custom Validation Profile from their filesystem.

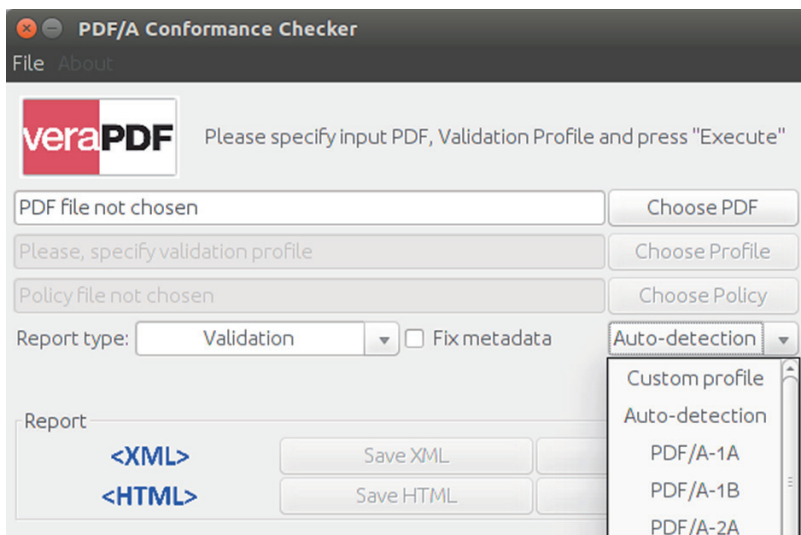


Figure 9: veraPDF GUI, choosing a Validation Profile

The options are illustrated in the following table 2:

Option	Description
Auto-detection	The veraPDF software will detect the PDF/A flavour when parsing the file and use the appropriate profile
PDF/A-1a	Use the PDF/A-1a validation profile, i.e. assume that the file is a PDF/A-1a
PDF/A-1b	Use the PDF/A-1b validation profile, i.e. assume that the file is a PDF/A-1b
PDF/A-2a	Use the PDF/A-2a validation profile, i.e. assume that the file is a PDF/A-2a
PDF/A-2b	Use the PDF/A-2b validation profile, i.e. assume that the file is a PDF/A-2b

Option	Description
PDF/A-2u	Use the PDF/A-2u validation profile, i.e. assume that the file is a PDF/A-2u
PDF/A-3a	Use the PDF/A-3a validation profile, i.e. assume that the file is a PDF/A-3a
PDF/A-3b	Use the PDF/A-3b validation profile, i.e. assume that the file is a PDF/A-3b
PDF/A-3u	Use the PDF/A-3u validation profile, i.e. assume that the file is a PDF/A-3u
Custom Profile	Enables the Choose Profile button allowing the user to load an external validation profile

Table 2: veraPDF Validation Profiles selection

Report Type, Policy, and Fix Metadata options allow the user to select the processing functionality and the information included in the generated report.

69



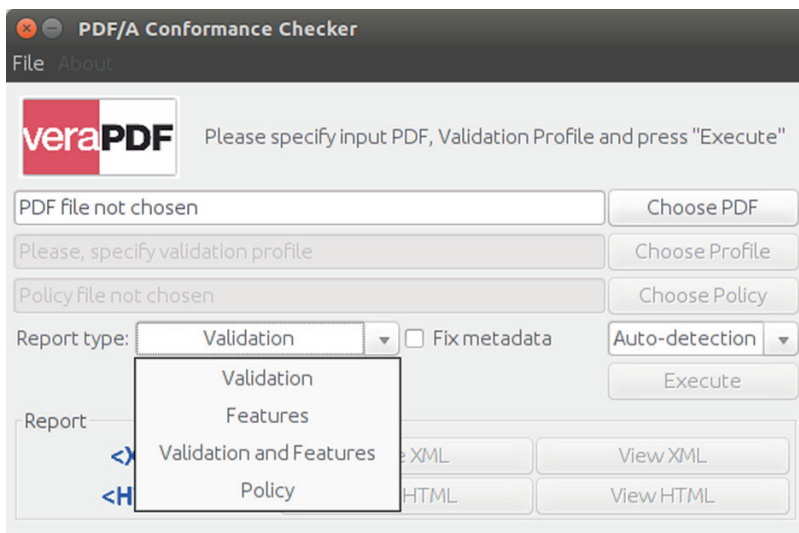


Figure 10: veraPDF GUI, choosing a Report Type

The available options are illustrated in the following table 3:

Option	Description
Validation [default]	Only perform PDF/A Validation and report the results. The feature reporting is disabled
Features	Only carry out feature reporting. Don't try to validate the PDF file
Validation & Features	Both PDF/A Validation and features reporting are carried out and the results reported
Policy	Perform a Policy check alongside validation and feature extraction

Table 3: veraPDF Report Type selection

If Policy is selected, the Choose Policy button is activated, and the user can use this to load a policy Schematron file. The Fix Metadata checkbox determines whether the software will attempt to amend the PDF document metadata to ensure it is compliant with the PDF/A specification.

The additional Settings dialog allows the user to configure the advanced settings.

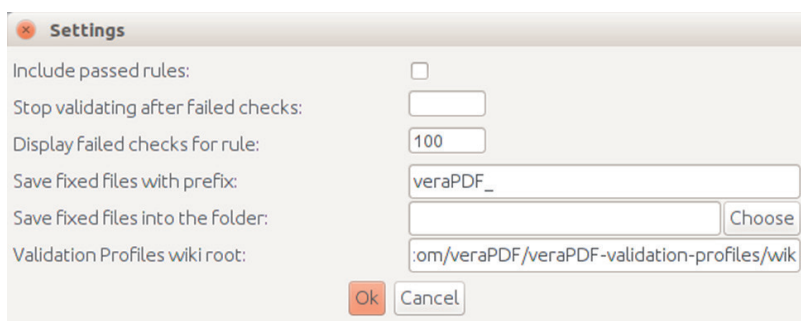


Figure 11: veraPDF GUI, advanced settings

The advanced settings are described in the table 4 below.

Setting	Description
Include passed rules	If checked, the passed validation rules are included into the resulting PDF/A Validation Report in addition to the failed rules. This setting is unchecked by default to reduce the size of the resulting report
Stop validating after failed checks	Specifies the maximum number of failed checks to be performed for all rules in the Validation Profile. Validation is halted once the number of failed checks is reached to speed up validation

Setting	Description
Display failed checks for rule	Specifies the maximum number of failed checks to be reported for each rule from the Validation Profile. The specified value is used as the limit for the number of failed checks that shall be included into the resulting report to reduce the size of the report
Save fixed files with prefix	Specifies the prefix that is added to the name of the original PDF document when saving it after automatic metadata fixing was performed. This setting is used only when 'Fix metadata' option is enabled
Save fixed files into the folder	Specifies the output folder for saving the PDF Documents after automatic metadata fixing was performed. Again, this setting is relevant only when 'Fix metadata' option is enabled
Validation profile wiki root	Specifies the base URL of the veraPDF Validation Profiles wiki. This provides contextual information about validation issues

Table 4: veraPDF advanced settings

A Validation Profile describes the tests that shall be performed during the validation. These tests are represented by rules that define a certain restriction on the PDF Document features. When validation is performed, the restrictions from the rules are checked for the relevant objects from PDF Document. A check may either fail or pass. In case of large documents, the number of passed and failed checks may be large, so the settings described above allow for reducing the number of redundant checks and thus, optimising validation time and the size of the resulting report.

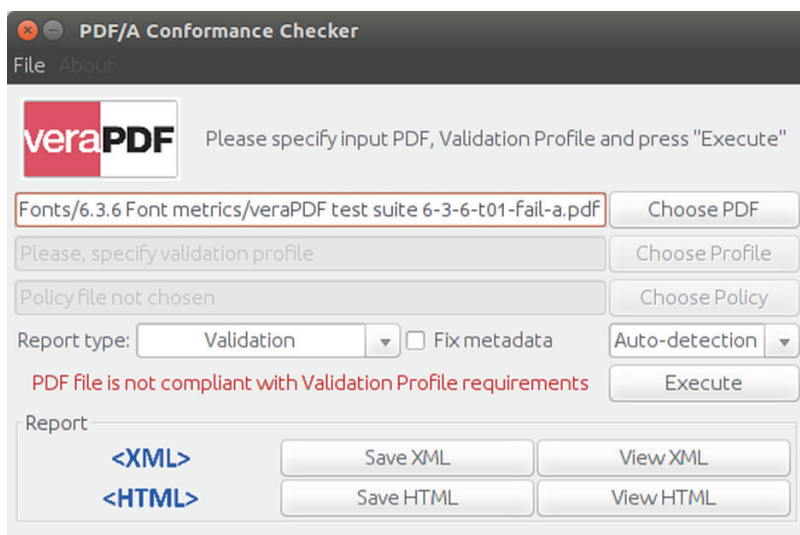


Figure 12: veraPDF GUI, result of the validation

When all the required settings are specified, the validation may be started by pressing the Execute button. During the processing, the progress bar is displayed. After the validation is finished, a resulting statement is shown and the reporting buttons are enabled. It is possible to save or view the resulting XML or HTML reports by pressing the corresponding button. The meaning of the reporting buttons is illustrated in the following table 5:

Option	Description
Save XML	Saves the processing results as a machine-readable XML report
Save HTML	Saves the processing results as a human friendly HTML report
View XML	Opens the machine-readable report in the default XML viewer
View HTML	Opens the human friendly report in the default HTML viewer

Table 5: Explanation of the veraPDF reporting buttons

The XML report contains a PDF/A Validation Report and a PDF Features Report, depending on the chosen processing options. Currently the HTML report only includes the results for the PDF/A Validation information.

Because memory institutions have quite specific and sometimes strict requirements, the veraPDF software supports two features designed to help them perform effective quality assurance for the material they preserve:

- **Feature Extraction**, which uses the built-in parser to extract information about a PDF document and its contents.
- **Policy Checking**, which performs user-defined checks on the features of a PDF document allowing users to enforce local policy.

The Feature Extractor can be configured from the **Features Config** menu. To use it, the **Information Dictionary** item needs to be checked.

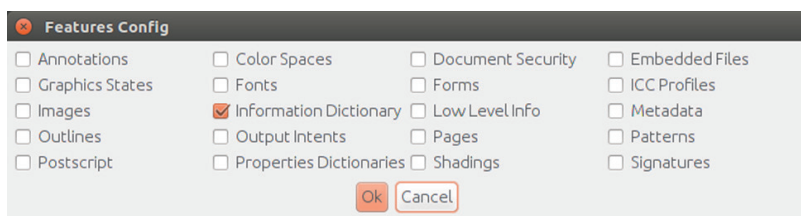


Figure 13: veraPDF GUI, feature extraction options

Through the Policy Checker, veraPDF can be used to perform additional checks beyond those mandated in the PDF/A specifications (e.g. to disallow a particular font by name or to ensure a named font is present in the document). To use the Policy Checker, the user need to select the Policy option from the report dropdown menu, which will enable the Choose Policy button through which

it is possible to load the policy file. Users can define custom checks using the XML Schematron syntax.

The veraPDF Policy Checker does not parse PDF documents directly. Instead, it processes the machine-readable report output generated by the PDF/A Validator and Feature Extractor. This means that the Policy Checker depends upon having the correct information in the report.

Further information about feature extraction and policy checking can be found on the veraPDF website.

VERAPDF CLI

The veraPDF command line interface is the best way of processing batches of PDF/A files. It is designed for integrating with scripted workflows, or for shell invocation from programs.

Once veraPDF has been installed, it is possible to get the software to output its built in CLI usage message, by typing `verapdf.bat -h` or `verapdf --help`. The main options are the following:

- `veraPDF -x, --extract`: Extracts and reports PDF features.
- `veraPDF --fixmetadata`: Performs metadata fixes.
- `veraPDF -f, --flavour`: Chooses built-in Validation Profile flavour.
- `veraPDF --format`: Chooses output format.
- `veraPDF -l, --list`: Lists built-in Validation Profiles.
- `veraPDF -o, --off`: Turns off PDF/A validation
- `veraPDF --policyfile`: Select a policy schematron or XSL file.

- veraPDF -p, --profile: Loads a Validation Profile from given path and exits if loading fails.
- veraPDF -r, --recurse: Recurses through directories. Only files with .pdf extensions are processed.
- veraPDF --savefolder: Sets output directory for any fixed files.
- veraPDF --success, --passed: Logs successful validation checks.
- veraPDF -v, --verbose: Adds failed test information to text output.



6.3. DPF MANAGER: LONG-TERM PRESERVATION OF DIGITAL IMAGES

INTRODUCTION: THE TIFF FORMAT AND ITS SPECIFICATIONS

TIFF is considered one of the best archival formats for preserving digital still images. Although TIFF is a proprietary format owned and maintained by Adobe Systems Software, the file format is fully documented and freely available. The TIFF Baseline 6.0 technical specification allows users to easily understand how the information is encoded as bytes, and also provides tools that can generate, edit and validate the technical integrity of the file.

TIFF is a flexible and adaptable file format. Those capabilities have contributed to the publication of several technical notes with extensions to the format, and several specifications have been based on TIFF 6.0, including TIFF/EP (ISO 12234-2), TIFF/IT (ISO 12639), TIFF-F (RFC 2306) and TIFF-FX (RFC 3949).

77

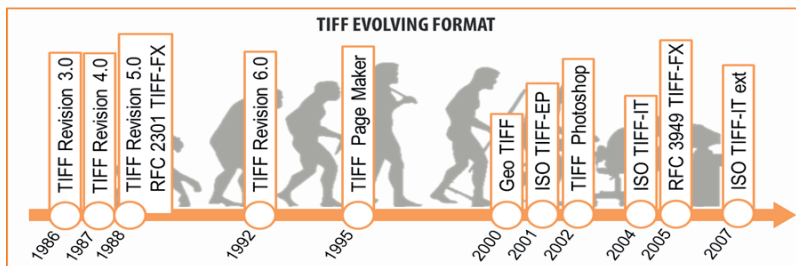


Figure 14: TIFF evolving format

Some of these extensions proposed new TIFF structures, tags, compression algorithms and colour schemas. However, not all the specifications fulfil the TIFF Baseline and, therefore, they may contradict previous recommendations.

Nowadays, the most preferable acceptance criteria by memory institutions is Uncompressed Baseline v6.0 IBM (little-endian byte order) RGB TIFF, although other photometric interpretations, like CMYK, Grayscale, Bilevel or compression schema , LZW, pack bits , CCITT Group 4 , CCITT Group 3 are also accepted.

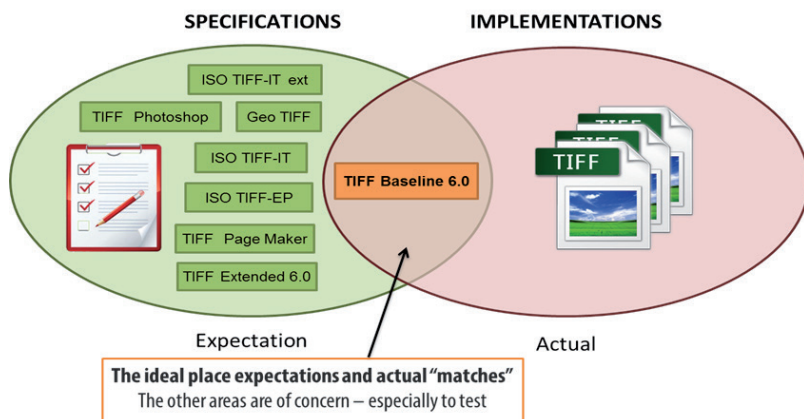


Figure 15: TIFF specifications and implementations

Using this acceptance criteria, we could wrongly think that we are correctly preserving our TIFF files.

Firstly, there are features that are compliant with the Baseline specification but negatively affect the preservability of a file: the usage of planar configuration, specific device colour profiles, extra channels, uncommon TIFF internal structures and so on.

Secondly, some parts of the specification can be interpreted in several ways, and the choices made by actual TIFF

implementation might not be the most appropriate in terms of long-term preservability.

Finally, there are features that are not fully covered by the current implementation. The TIFF format is self-documented and able to handle multiple metadata containers: XMP, EXIF, IPTC. The specification describes how to include the metadata, but it does not describe how to deal with specific situations, e.g.:

- There are duplicated tags defined in multiple containers.
- IPTC-IIM has evolved over time and has been replaced by the IPTC core schema in the XMP container.
- The new XMP container includes namespaces for storing EXIF and TIFF metadata creating duplicated metadata inside the TIFF.

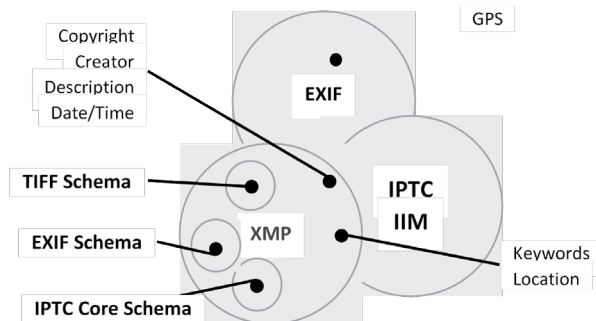


Figure 16: TIFF metadata containers

In order to solve the issues related to the management of multiple metadata containers in the same file, the Metadata Working Group, formed in 2006 by Adobe Systems, Apple, Canon, Microsoft and Nokia, created the specification Guidelines for Handling Image Metadata.

The main problem for memory institutions is that none of the current TIFF specifications are designed for archival purposes. TI/A (Tagged Image for Archival) is a new initiative that has been launched to fill this gap and produce a standard recommendation that clearly identifies which features of a TIFF file are appropriate for digital preservation and which ones negatively affect its preservability.

The first draft of the TI/A recommendation has been recently sent to the ISO committee and the standardisation process has started.

DPF Manager is the most advanced TIFF conformance checker for digital preservation and it is the only existing application able to validate TIFF files using the TI/A official draft specification.

WHAT IS DPF MANAGER ABOUT?

DPF Manager is a multi-platform application designed to allow end users and developers to gain full control over the technical properties and structure of TIFF images intended for long-term preservation. It displays, in an intuitive and easy-to-understand way, the information (metadata) contained in the files, and reports specific failures that can be traced to the TIFF specifications. The tool also allows the user to define and validate custom policies like quality standards for the images and includes a metadata fixer to solve common errors in the digital images, making them compliant with the selected criteria.

DPF Manager aims to reduce the time and effort required to revise file structure, metadata and the institution's acceptance criteria, as well as to give advice about image preservability.

The main features of DPF Manager are:

- TIFF file identification, including the TIFF specifications the file complies with.

- Validation of the conformance to a specific normative, either defined by ISO standards or based on locally-defined policy rules.
- Reporting metadata inconsistencies between different metadata containers.
- Fixing a TIFF file to make it compliant with the selected criteria, preserving at the same time the Image Representation.
- User and machine-readable reports in different formats, including data object structure, metadata and validation results.
- Easy integration with the Digital Asset Management (DAM) software of memory institutions using the OAIS model.
- Reporting information in METS, PRIMES and NISO standards.

DPF Manager is suitable for different scenarios, working as a standalone, client-server or as a command-line application.

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

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.

INSTALLING DPF MANAGER

DPF Manager Installer comes in two versions: one that includes the Java Virtual Machine, and another one that

does not include it ("lite" version). If Java 8 is already installed in the computer, the "lite" version can be used.

Installers (available at <http://www.preforma-project.eu/dpfmanager-download.html>) are provided for the most common operating systems, i.e. Windows, Linux and Mac. 32 and 64-bit versions are available too.

DPF Manager can be run in command line mode (CLI) and through a Graphical user interface (GUI).

In Windows there are two executables, one for the CLI, called "dpf-manager-console.exe" and another one for the GUI, called "DPF Manager.exe".

In Linux, there is a single executable for both interfaces, called "dpf-manager". To open the GUI, execute it without parameters. To run the CLI, execute it with parameters (see "dpf-manager --help").

For MacOS, there is a single package in the Applications folder called "DPF Manager" that runs the GUI by default. The CLI can be run by launching the executable "DPF Manager.app/Contents/MacOS/DPF Manager" through the terminal.

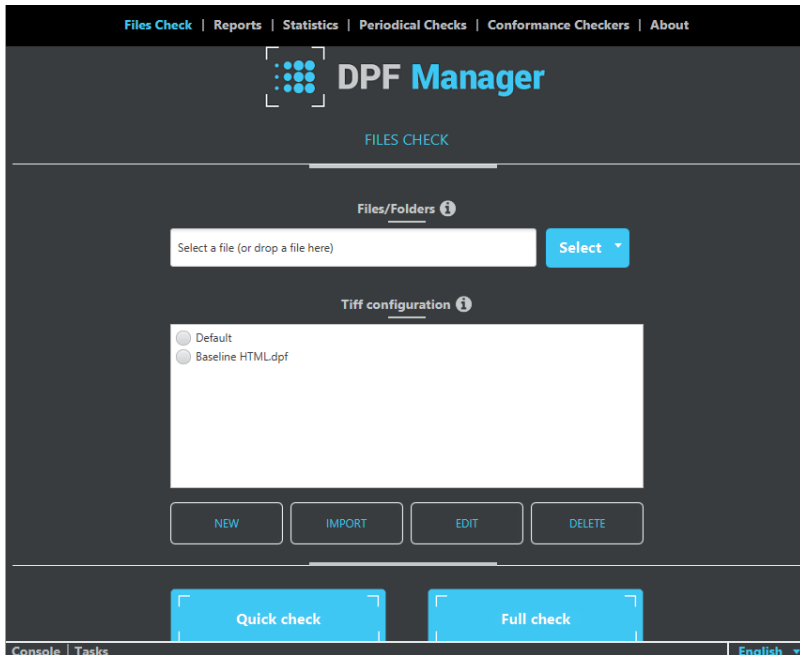
A web user interface is also available at <http://dpfmanager.org/application.html>.

82



VALIDATING TIFF FILES WITH THE GUI

DPF Manager GUI opens with the following screen:



83

Figure 17: DPF Manager GUI, starting screen

In the “**Files/Folders**” section, the input file (or folder, or url, or zip) can be selected by:

- Manually typing its location.
- Dragging the input file/folder into the text box.
- Clicking the “Select” button and uploading the file/folder from the computer.

In order to validate the input files, a configuration needs to be defined. The software comes with a default configuration which verifies that the files conform to the Tiff Baseline 6.0 and generates a report in HTML format. New configurations can be created by clicking the “New” button. All the configurations are displayed in the “**TIFF configuration**” section for easy selection.

Once both the files to be analysed, and the configuration file, have been defined, the **check buttons** starts the process. A **full check** validates the tiff files and shows all the errors found in the report, while a **quick check** only shows if the Tiff files have any error or not.

The language of the GUI can be changed using the bottom-right selector. The creation of a new configuration file consists of five steps. In the first step, the ISO standards to be validated are selected.

84

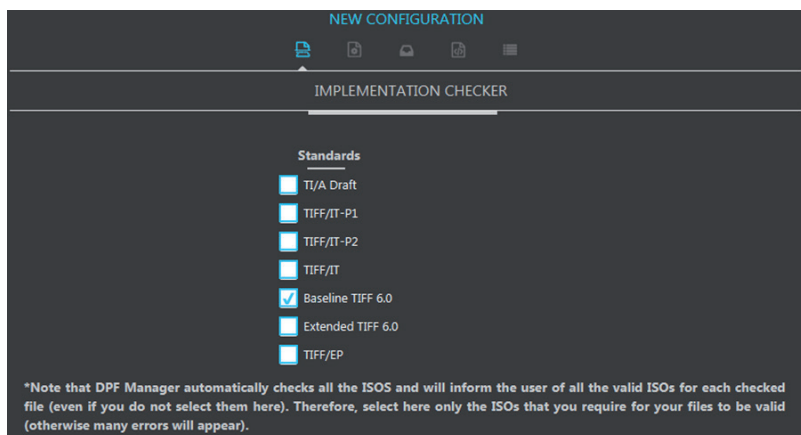


Figure 18: DPF Manager GUI, selection of the standard specification

In the second step, the **policy checker** is configured by adding rules corresponding to specific acceptance criteria of the memory institution. It consists of two parts. In the first part (custom standards), each of the standards selected in the implementation checker step can

be edited. In the second part, the acceptance criteria of the organisation can be defined by adding rules with the desired properties for the TIFF files.

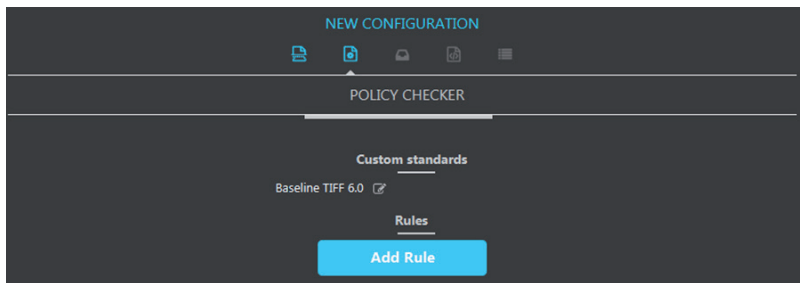


Figure 19: DPF Manager GUI, policy checker interface

When editing an ISO standard, a new panel appears showing all the rules that the standard specifies for a file to be valid. These rules can be disabled in order to create a less restrictive implementation check.

85

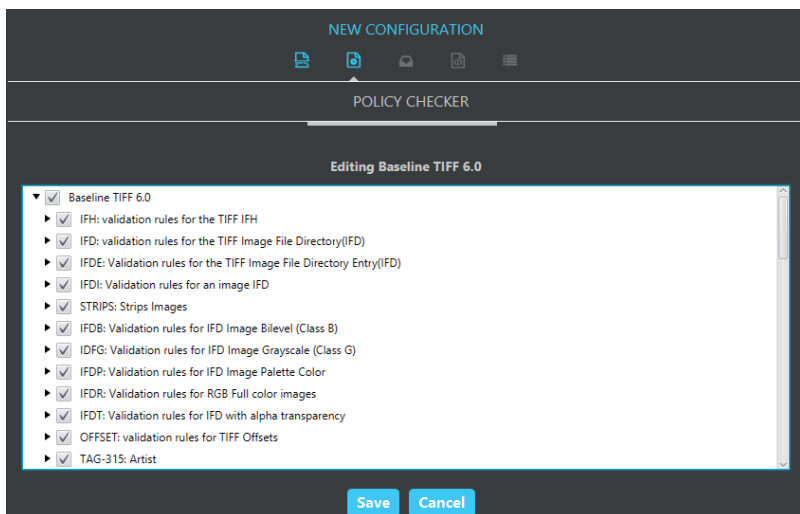


Figure 20: DPF Manager GUI, editing ISO standard rules

In the second part, rules can be added with the desired properties for the TIFF files.

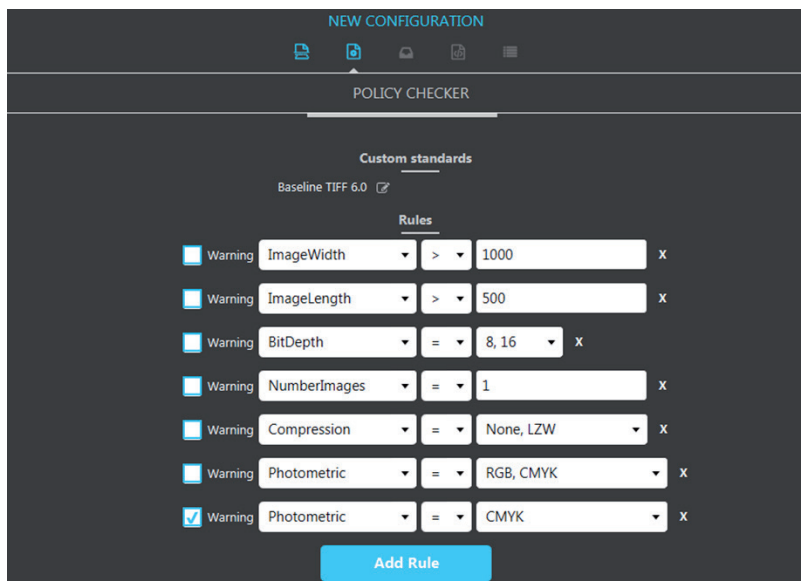


Figure 21: DPF Manager GUI, definition of custom policies

Rules can be defined as “mandatory”, meaning that images not satisfying the rule will produce a policy error, or as “warnings”, meaning that the images not satisfying the rule will produce a warning. The available policies are:

- ImageWidth: Checks the width in pixels of the image.
- ImageLength: Checks the height in pixels of the image.
- LongEdge: Checks the longest value of the above.
- PixelDensity: Checks the resolution, in pixels per centimetre.

- **NumberImages:** Checks the number of images in a single TIFF file.
- **BitDepth:** Number of bits per pixels component (multiples of 2).
- **DPI:** Checks the resolution, in dots per inch.
- **ExtraChannels:** Number of extra pixel components (e.g. transparency).
- **EqualXYResolution:** Checks that the X and Y resolution of the image are the same.
- **Compression:** Compression scheme.
- **Photometric:** Color space of the image data.
- **Planar:** How the pixels components are stored.
- **ByteOrder:** Byte order (little endian, big endian).
- **FileSize:** The size of the file in bytes.
- **ICCProfileClass:** Class of the device ICC Profile.

87

In the third step, the **format of the report** and the output folder are selected.



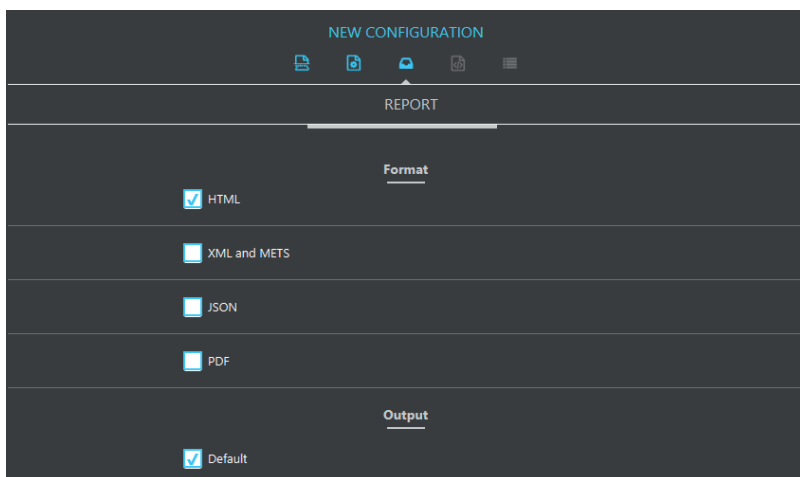


Figure 22: DPF Manager GUI, selection of the output format

88

Step number four is the **metadata fixer**, where the user can specify what changes need to be applied to the image, e.g. adding or removing metadata. A new image will be generated with the modifications specified in this section; the original image will not be altered.

There are four automatic fixes available:

- **Clear Private Data:** This removes all the information related to the GPS coordinates where the photo was taken.
- **Fix non-Ascii tags:** This solves a common error related to text that is not encoded in 7-bit ascii, which is the encoding allowed for TIFF file.
- **Make Baseline Compliant:** Fixes some common errors regarding the Baseline 6 specification.
- **Fix Metadata Inconsistencies:** Resolves incoherencies in the metadata coming from IPTC, XMP and EXIF,

following the guidelines of the Metadata Working Group¹.

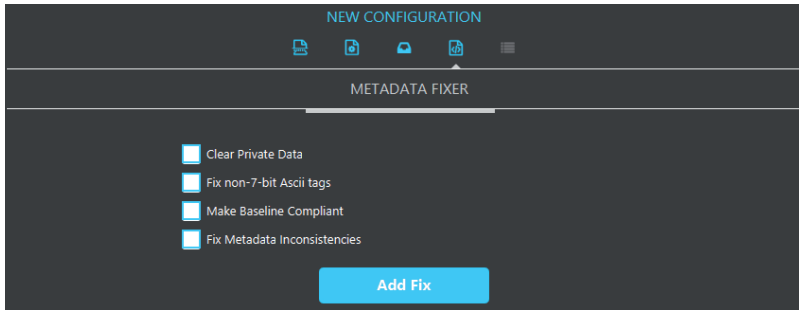


Figure 23: DPF Manager GUI, metadata fixer options

Finally, the last step presents a **summary** of the configuration and allows the user to name it for future reference.

Configurations can be opened and modified by clicking the button **"Edit"** in the main window. The steps are the same as above, and the configuration will be saved either to the same file or to a new one.

External configuration files can also be imported to the configuration files folder through the **"Import"** button in the main window. When a configuration file is imported, the program asks whether it should be saved in the default configuration directory, so that, in future executions, the configuration will appear in the configurations list.

The **"Periodical Checks"** tab allows the user to define checks to be performed periodically.

1. <http://www.metadataworkinggroup.org/specs/>

In order to configure a periodical check, a source (file or folder) must be defined and a configuration file has to be selected. Then, the time period can be set either as daily, weekly or monthly, together with the preferred time.

Periodical checks create tasks in background in the operating system, so it is not necessary for the DPF Manager GUI application to be running for the periodical checks to be executed.

Input path	Configuration	Periodicity
D:\Escritorio\test100\Bilevel.tif	Baseline HTML	Daily, at 12:57

Select a file (or drop a file here) [Select] [Configuration Dropdown]

Time: 13:24
☐ Daily ☐ Weekly ☐ Monthly

[ADD]

Figure 24: DPF Manager GUI, periodical checks options

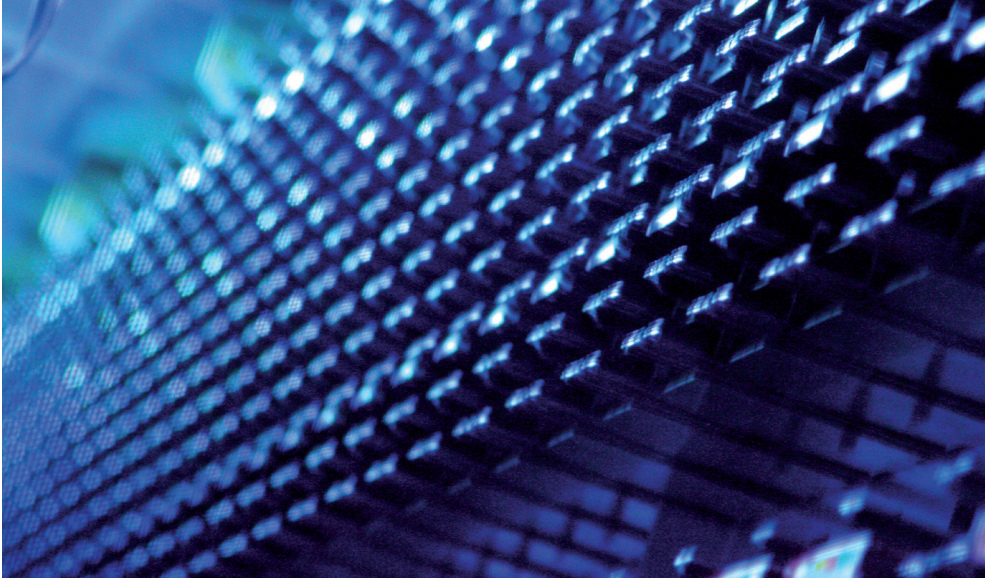
External conformance checkers can be configured in the **"Conformance checkers"** tab.

Enabled	Name	Extensions	Type	Configuration
<input checked="" type="checkbox"/>	DPF Manager	tif, tiff	Internal	Default

Name: [Input] Extensions: [Input]
Location: [Input] [Select] *Enter the extensions separated by comma.
Parameters: [Input] Configuration: [Input] [Select]

*Parameters must contain %input% and %config%

Figure 25: DPF Manager GUI, configuration of external Conformance checkers



By default, only the built-in TIFF conformance checker is set, but it is possible to define additional conformance checkers for other file formats (file extensions) by specifying their location, arguments and configuration file. It is possible also to define more than one conformance checker for the same file extension. In this case, the DPF Manager will automatically leverage the load over all the available conformance checkers of the same format.

91

A **tasks widget** is available at the bottom of the screen to show all the checks that have launched. When a task is running, it can be paused, resumed and cancelled.

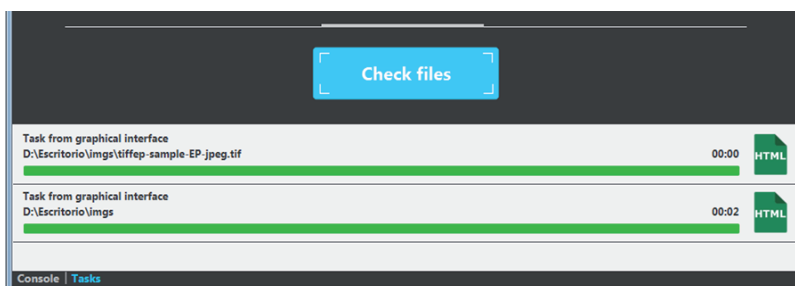


Figure 26: DPF Manager GUI, tasks widget

A **console widget** is also available, which is used as a logging system and displays possible errors that could occur during the execution.

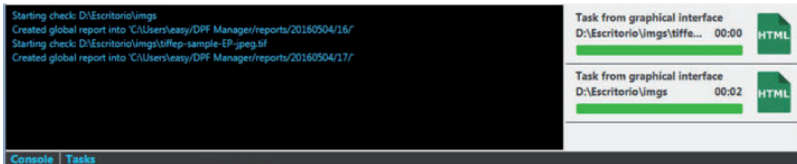


Figure 27: DPF Manager GUI, console widget

When a task is complete, an icon is shown on the left linking to the summary report.

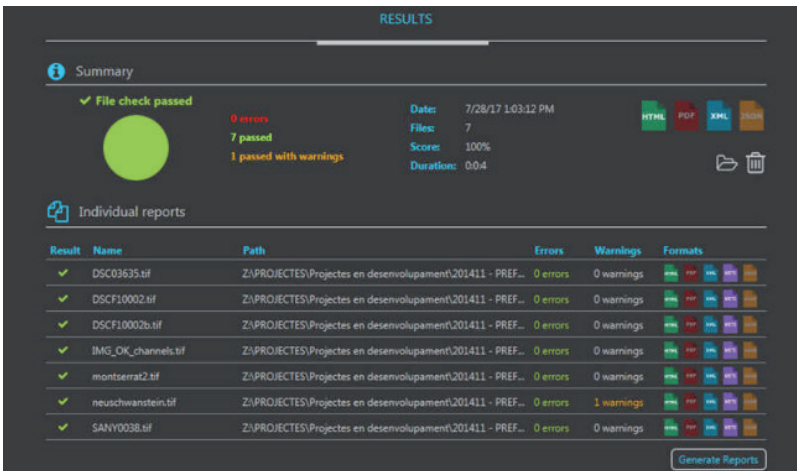
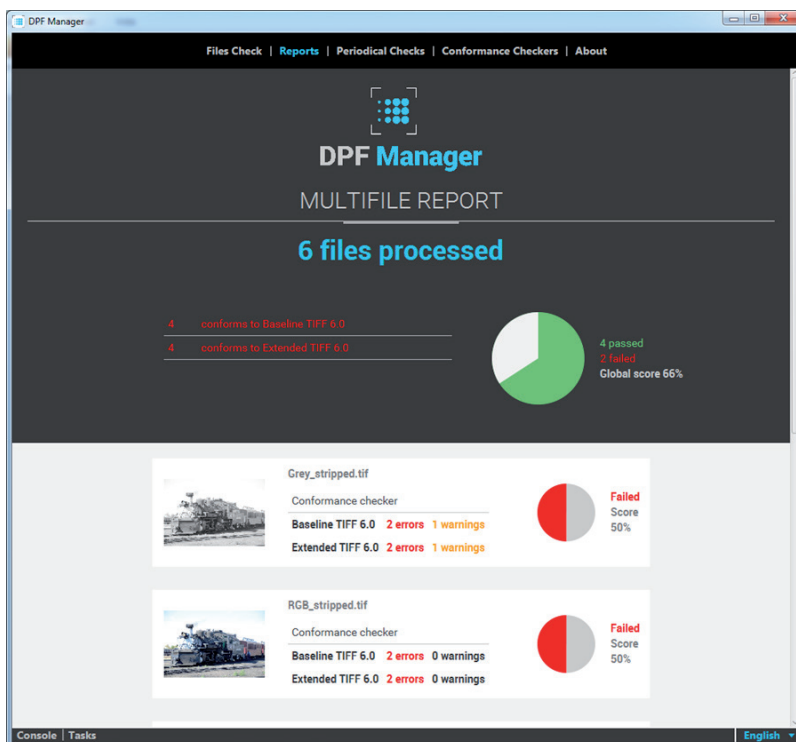


Figure 28: DPF Manager GUI, summary report - overview screen

The summary report presents an overview of the results showing the number of files that conform with each of the selected specifications and policies together with a graphical representation. After that, a table shows all the files analysed, and the errors and warnings found.

In addition, a list of formats is available for the generation of the individual reports.

When quick checks are performed instead of full checks, the errors and warnings columns are not shown, but an option to generate full checks of the individual reports is available, as well as a button to generate full checks of the entire set of files.



93

Figure 29: DPF Manager GUI, summary report – HTML view

The individual report, available by clicking on the respective result, contains all the details of the validation.

The upper part shows the thumbnail, image name, path and number of errors and warnings discovered for each of the selected specifications. Scrolling down, the internal file structure of the TIFF file is presented (right side), together with the list of tags (left side).

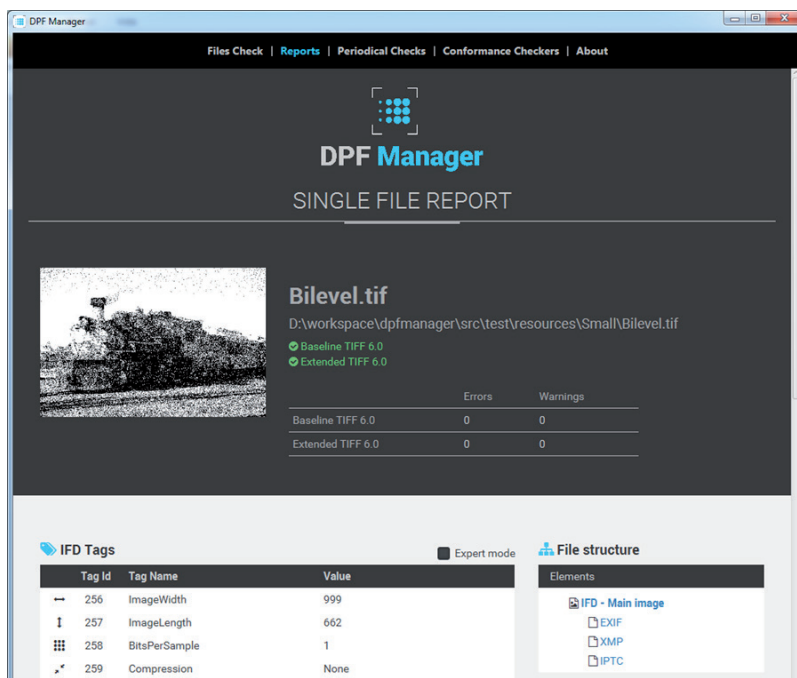
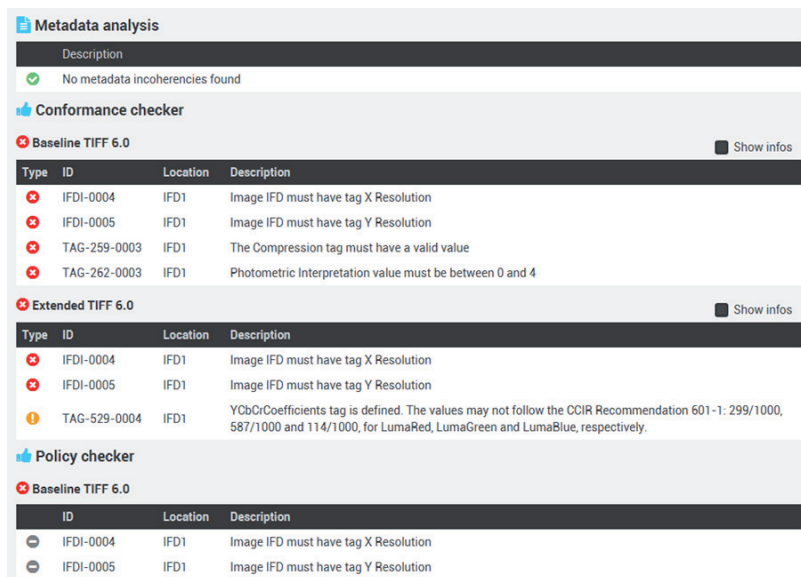


Figure 30: DPF Manager GUI, individual report

Below that, there is a section that analyses the metadata incoherencies, highlighting any inconsistency that has been found in the metadata containers (for example, different authors defined in different containers)².

2. Solving metadata incoherencies is one of the fixes that are available in the metadata fixer options

Finally, the last section lists all the specific errors and warnings that have been discovered for each of the selected specifications and policies. Each error contains information about the rule that has been violated, with the location where the error has been found and a description with suggestions on how to solve it.



Metadata analysis

Description

✓ No metadata incoherencies found

Conformance checker

✗ **Baseline TIFF 6.0** Show infos

Type	ID	Location	Description
✗	IFDI-0004	IFD1	Image IFD must have tag X Resolution
✗	IFDI-0005	IFD1	Image IFD must have tag Y Resolution
✗	TAG-259-0003	IFD1	The Compression tag must have a valid value
✗	TAG-262-0003	IFD1	Photometric Interpretation value must be between 0 and 4

✗ **Extended TIFF 6.0** Show infos

Type	ID	Location	Description
✗	IFDI-0004	IFD1	Image IFD must have tag X Resolution
✗	IFDI-0005	IFD1	Image IFD must have tag Y Resolution
!	TAG-529-0004	IFD1	YCbCrCoefficients tag is defined. The values may not follow the CCIR Recommendation 601-1: 299/1000, 587/1000 and 114/1000, for LumaRed, LumaGreen and LumaBlue, respectively.

Policy checker

✗ **Baseline TIFF 6.0**

ID	Location	Description
IFDI-0004	IFD1	Image IFD must have tag X Resolution
IFDI-0005	IFD1	Image IFD must have tag Y Resolution

Figure 31: DPF Manager GUI, errors and warnings screen

In case a fix has been defined, a secondary report is available showing the differences between the original and the fixed image.

Finally, a **statistics module** is also available to present an overview of all the reports that have been performed with DPF Manager, together with some information about the tags that have been found in the reports. This information is sorted by frequency, the different ISO standards that have been analysed and the policies that have been defined.



Figure 32: DPF Manager GUI, statistics module

96

DPF MANAGER CLI

DPF Manager can be run in command line mode too. The available commands are:

- **check:** Performs a local files check.
- **config:** Manages the configuration files.
- **gui:** Launches graphical user interface.
- **modules:** Manage the conformance checkers.
- **periodic:** Manage periodical checks.
- **remote:** Performs remote file checks (client mode).
- **server:** Launches the server mode.

Each command has its own help function explaining how to invoke it and outlining the associated options and parameters.

6.4. MEDIACONCH: CONFORMANCE CHECKING FOR AUDIO-VISUAL FILES

INTRODUCTION AND TARGET FORMATS

Video files can be very complicated, because audio-visual files are often made up of at least three formats: a container format, a video encoding format, and an audio encoding format. As a result, checking for errors can be three times as difficult. By using MediaConch to check files against their specifications, cultural heritage institutions can be confident that their video files are technically correct and will be able to successfully play back and transcode to different formats far into the future.

MediaConch works with virtually all audio-visual files, but has been developed with three file formats in mind:

- Matroska: A multimedia container, or wrapper, format that holds encoded video and audio streams as well as supplemental metadata. Matroska files are based upon EBML, a binary file format similar to XML.
- FFV1: A lossless, open source video encoding format developed by FFmpeg.
- LPCM: A method for encoding audio. It is an uncompressed audio format.

Why these formats? Matroska and FFV1 are open file formats. Their specifications are freely available and openly licensed. Continued development is open and available to the public, historical context and conversations surrounding the specification are publicly accessible, and use of the formats and their specifications is without

charge and can be used by any person, institution or company. They can also be improved upon by any person, contingent on the standards body with the responsibility to collectively approve changes.

Matroska as an audio-visual file format has been in use since 2002, with widespread internet usage. Matroska has been adopted as the foundation of Google's Webm format, which a file format optimized specifically for web-streaming. Some of Matroska's features - such as subtitle management, chaptering, extensible structured metadata, file attachments, and broad support of audio-visual encodings - have facilitated its adoption in a number of media communities. Matroska has also been implemented into many home media environments such as Xbox and Playstation and works "out of the box" in the Windows 10 operating system.

98 The Matroska wrapper is organised into sectional elements, and each element may have a dedicated checksum associated with it. This is one of the primary reasons why it is deemed such a suitable format for digital preservation. Discrete sections of a file can be checked for errors, which means error detection can be specific to a region (as opposed to having to identify errors within the entire file). For example, a checksum mismatch specific to the descriptive metadata section of the file can be assessed and corrected without having to do quality control and analysis on the file's content streams. Considering the potentially vast file sizes and complexity of audio-visual files, this can greatly reduce not only the time required to analyse and repair files, but also the amount of data throughput over a network, as well as requiring less computing power. The Matroska format features embeddable technical and descriptive metadata so that contextual information about the file can be embedded within the file itself, not just provided alongside in a different type of document.

FFV1 is an efficient, lossless video encoding format that is designed in a manner responsive to the requirements of digital preservation. FFV1 has rapid traction in both the development and digital preservation communities and

is widely and freely distributed with the ubiquitous FFmpeg and libav libraries for video processing. FFV1's lossless compression algorithm allows for a reduction in file size without loss of quality. Additionally, FFV1 version 3 is a very flexible encoding format, allowing adjustments to the encoding process based on different priorities such as size efficiency, data resilience, or encoding speed. FFV1 is a strong candidate for video files undergoing file format normalization prior to the OAIS-compliant repository ingestion phase. For example, Artefactual's Archivematica (a free and open source digital preservation system) recommends pre- and post-normalization FFV1+MKV validation methods.

Linear PCM is a simple, uncompressed representation of an audio wave. LPCM's strength is in its simplicity and lack of compression; it is a pure data format with widespread usage and is the de facto standard for digital audio.

WHAT IS MEDIACONCH ABOUT?

99

MediaConch is an extensible, open source software project that analyses preservation-level, audio-visual files for use in memory institutions, providing a detailed report of a file's technical metadata and other related information. MediaConch validates files down to the bit-level, ensuring every part of a video file is exactly what it claims to be. MediaConch can be used during file creation or ingestion, after a file migration, during a quality analysis or quality control phase, or as part of routine file check-ups.

Video files are based in time, so comprehensive quality control historically would require at least as much time as the length of the video file, a practice unsustainable when dealing with thousands of files. Automated testing with MediaConch allows to quickly detect common errors, without the need for individual inspection. Therefore, preservationists can spend their time focusing only on files that fail validation and need more granular care while feeling confident that the other files are healthy and comply with their standardisations. MediaConch also



produces warnings about potential risks related to compatibility issues, in case files may have difficulty playing back in certain contexts but not in others. It also has the capability to correct minor bit-level file errors that result from digital storage entropy.

MediaConch provides 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. Additionally, MediaConch's policy feature can create in-house policies specific to the needs of the cultural heritage institution, for all the audio-visual formats used by these institutions. MediaConch policies are built within the MediaConch GUI and can be exported for use by other users or implemented into a command-line batch processing script. MediaConchOnline provides an online public directory of shared policies from other MediaConch users.

MediaConch is currently being developed by the MediaArea team, notable for the creation of open source audio-visual file metadata reporting software: MediaInfo.

INSTALLING MEDIACONCH

MediaConch is available for use as a command line interface (CLI), a graphical user interface (GUI), a web user interface (WebUI), and in fully automated server mode.

The command line tool can be downloaded at <http://www.preforma-project.eu/mediaconch-download.html> or built from the MediaConch source code on most operating systems. For Mac and Linux Homebrew (Linuxbrew) users, Mediaconch CLI can also be installed using the ``brew`` command: ``brew install mediaconch`` will install the command line tool. For Ubuntu users, MediaConch CLI is available directly in the distribution and can be installed using the Ubuntu command ``apt install mediaconch``. To begin using the command line tool, type ``mediaconch -h`` into the terminal to get helpful commands.

The graphical user interface can be downloaded at <http://www.preforma-project.eu/mediaconch-download.html> or built from the MediaConch source code on most operating systems. For Ubuntu users, MediaConch GUI is available directly in the distribution and can be installed using any package manager. For example, in the Ubuntu software center (or Synaptic, etc), one must simply enter ``MediaConch`` in the search box. For Mac users, MediaConch GUI is available directly in the Mac App Store by entering ``MediaConch`` in the search box.

The web user interface can be previewed at <https://mediaarea.net/MediaConchOnline/> or installed from source at <https://github.com/MediaArea/MediaConchOnline>.

For Ubuntu users, MediaConch Server is available directly in the distribution package and can be installed using the Ubuntu command ``apt install mediaconch-server``. The server features the ability to add a watch folder, including fully automated processes; it is planned to add the ability to email results to users.

VALIDATING VIDEO FILES WITH MEDIACONCH

MediaConch works with virtually any file format, but is created to work specifically with Matroska-wrapped FFV1 and LPCM encoded video files. For these files, validation is performed according to the specifications of each of these formats. If, for example, an AVI-wrapped FFV1 and LPCM encoded video file is used, thorough file validation

will be performed on the FFV1 and LPCM portions and the AVI file will be minimally checked for validation and approved if these validations pass. MediaConch works with sister projects veraPDF and DPF Manager to perform thorough validation on PDF and TIFF file formats, too.

MediaConch consists of three main sections: “Checker,” “Policies,” and “Display.”

By default, MediaConch opens in in the “**Checker**” section. Files from various sources can be checked with MediaConch, regardless of the chosen interface. “Check local file” allows to select a file or files from a local computer. “Check online file” allows to select a file using a URL path. “Check local folder” allows to select an entire folder from a local computer or volume.

For any of these selections, it is possible to select a custom policy by choosing it from either an existing policy in MediaConch, or from an imported XSLT or Schematron policy file. MediaConch comes preloaded with several policies that can either be used as they are, or as a starting point for creating custom policies for similar use-cases.

102

MediaConch

Checker Policies Public Policies Display Settings Help

Check files

Check local file Check online file Check local folder

Policy: No policy Display: MediaConchHtml Verbosity: 5

☐ Enable fixer

Select files: Choose Files No file selected

Check files

Results

Show 10 entries Search:

Files	Implem	Policy	MediaInfo	MediaTrace	Status
No data available in table					

Showing 0 to 0 of 0 entries

Previous Next

Figure 33: MediaConch GUI, Checker screen

After selecting file, policy, display, and verbosity and clicking “Check files”, the following reports are generated below in the results section: Implementation, Policy, MediaInfo, and MediaTrace.

The implementation report declares whether a particular file is either VALID or NOT VALID according to the specification. If a file corresponds to one of the standards and is not valid, the reason for the failure is provided. The policy report compares the file against the assigned policy. Mediainfo (high-level overview of the content of the file) and MediaTrace (binary analysis of the file) reports are produced too.

Implementation and policy reports can be displayed or downloaded in the following formats:

- HTML
- Text
- Text with Unicode support
- XML

103

In the command line interface, reports can also be viewed using the “Simple” or “CSV” flags. The simple flag will create a display with “pass” or “fail” for each report segment. The CSV flag, intended specifically for use with policies, will generate a table of results in CSV format.

In the GUI and WebUI, reports can be downloaded by either clicking on the down arrow found directly to the right of each report, or by clicking on the “Download” button located at the bottom right of each report in View Mode. All formats can be exported and stored as sidecar metadata, PREMIS object, or a preservationist may choose to store their preservation materials.

Implementation and policy reports give a high-level pass/fail view. Hovering over these results, “see more” and “download” buttons will appear, just like the MediaInfo/MediaTrace reports.

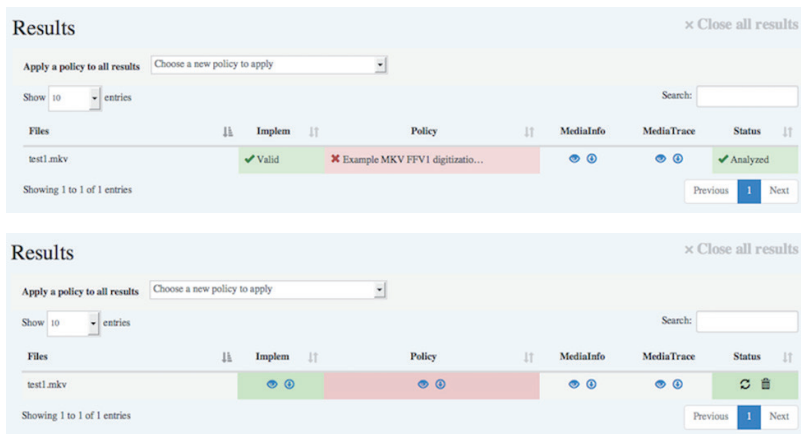


Figure 34: MediaConch GUI, implementation and policy report

104

For MediaInfo reports, there is an additional “Create policy from MediaInfo report” button. This can be used to create a new policy based on an existing file.

In MediaConch, users can develop their own policies and share them with other memory institutions, as they can be easily exported and imported between instances or frameworks.

In the “Policies” section of the GUI or WebUI software, a preservationist can create customised policy tests to check for conformance to a specific set of standards that a collection must adhere to. It is also possible to import previously generated policy sets in either XSL or Schematron format.



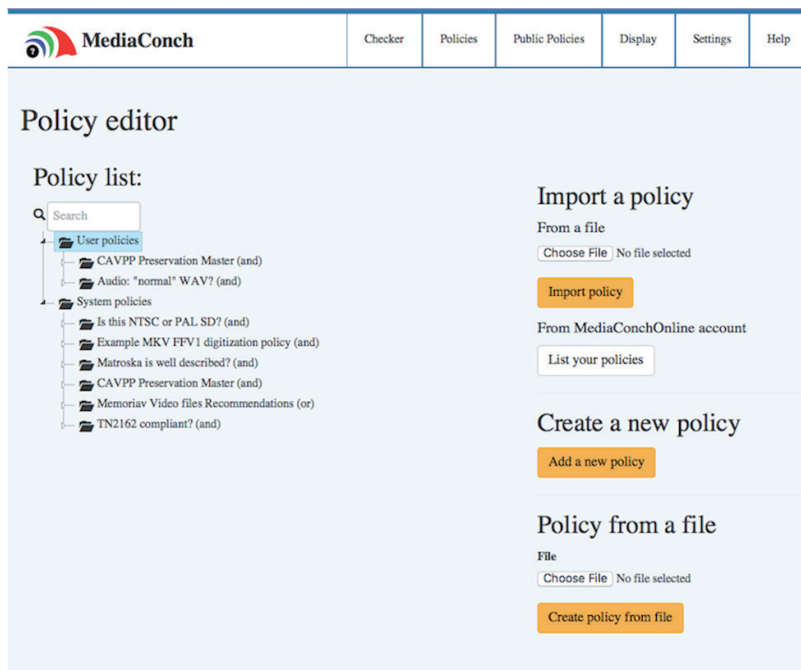


Figure 35: MediaConch GUI, Policy editor

Video files are broken down into different “streams”, which a preservationist can select depending on the section of the video they are looking to create a policy for. These streams include General, Video, Audio, Image, Text, Menu, or Other.

Policy sets consist of individual rules and assertions. A policy may contain one or more rules, and rules may consist of one or more asserts. Rules and assertions typically contain a metadata field (e.g., “Format”), that field’s associated metadata stream type (e.g., “General”), a validator (e.g., “is_equal”), and a desired value (e.g., “Matroska”). In brief, a preservationist would select a type, a field, a validator, and a value. Rules and assertions are automatically saved during creation,

but it is possible to duplicate or delete them using the associated buttons on each rule/assertion window.

The “**Display**” section allows to apply various display XSLs for use with policy and implementation check reports in the Checker section.

Once a display XSL is imported, it can be instantly used by selecting the display in the “Choose a Display” dropdown menu when checking files in the Checker section.

MediaConch includes features for checking and implementing fixity within video files, including some methods to correct small errors, with a focus on Matroska files. For example, some files may have an incorrect Matroska Segment element size (e.g. set to 0 instead of the actual size). As a result, a file with this problem is playable on some video player software that support corrections of this bug, but not on some other players (e.g. VLC can play the file, but Windows 10 Media Player cannot). This is an example of an interoperability issue resulting in inconsistent playback of a file. The fixer feature in MediaConch is able to resolve this issue.

MediaConch can also check for “bit flipping” and correct it if found, saving the file from damage caused by an unintentional state switch between 0 and 1, a problem which can occur to bits stored for a long period of time. This can be especially helpful and efficient when video streams are encoded with FFV1 and each FFV1 frame/slice is “protected” with a CRC checksum.

MEDIACONCH CLI

Running `mediaconch -h` in the terminal window will display the primary options available for the command line tool. `mediaconch -ah` will display only advanced options.

Simply, `mediaconch` can be run by following this pattern:
`mediaconch [-Options...] FilePath1 [FilePath2...]`

Here are a few handy commands and an explanation of what they do:

- `mediaconch [FilePath]` - This is the simplest command, and it will test and return the validation results on screen.
- `mediaconch [FilePath] > output.txt` - The above output can be very verbose, possibly too verbose for a terminal window, so this command will send the output to a text file called `output.txt`.
- `mediaconch -fs [FilePath]` - This will export a simple pass/fail output for each test performed on the file.
- `mediaconch -p my_policy.xml [FilePath]` - This will test the file against a local policy. Policies must be created via MediaConchOnline or the MediaConch GUI and can be exported to use anywhere.

With the command line tool, MediaConch can be used on thousands of files at once, alleviating the need for the preservationist to waste time inspecting healthy files.

To check multiple files, a user can simply run ``mediaconch *.mkv`` to run implementation testing on all Matroska files in a folder.

7. TAKING CONTROL OF CONFORMITY TESTS PROCESS OF DIGITAL FILES: AN ACTION PLAN

The general steps to consider when preserving digital objects long-term are described more in detail in Chapter 4. This chapter discusses the actions to be taken by memory institutions and other organisations when preserving digital objects long-term in order to establish a process of controlled conformity tests.

For organisations to gain control over the technical properties of their digital objects, the basic instrument, presented in this Handbook, is the Conformance checkers developed in the PREFORMA project. Their function is to guarantee that data objects are produced according to standards, tested for conformity, and (if needed) re-processed for corrections.

But digital objects need to be preserved in a context to be understandable and usable to future users. Therefore, a set of non-technical issues must be involved and decided upon when managing digital holdings and collections. In summary, these issues are the governance principles of a successful preservation process and should address three key phases:

- To establish a sustainable strategy for preserving digital objects
- To take policy decisions
- To deploy tools that allow for the implementation of these decisions

In the following sub-sections, a number of actions are proposed that connect to these key phases.

7.1. BASIC GOVERNANCE DECISIONS

The average time span of digital preservation solutions is usually:

- Short-term preservation – solutions that are used for a fleeting time (normally a maximum five years)
- Medium-term preservation – solutions that are used during a system's lifetime (normally around ten years)
- Long-term preservation – solutions that are used after the originating system's lifetime (number of years unspecified).

In order words, digital preservation is about taking measures in advance, regardless of whether the aim is to preserve the digital files and their data short-term, or beyond the lifetime of current technology. To cope with that, digital objects need to be selected for active preservation treatment at an early stage, otherwise they run the risk of being lost or unusable. Digital preservation, therefore, requires decisions about:

- **A vision**, i.e. Why we do it? What are the objectives? The vision is strictly connected with the specific nature of the institution, its vocation, its purposes and its audiences, and should be a central consideration at the highest level of decision making.
- **A policy**, i.e. How do we want to achieve our goals? Policies are also specific to the institution, and they derive from its vision that of the digital archive's life and role. Then the PREFORMA tools can support the characterisation of the policies, in terms of parameters that can be processed by the software.
- **A governance model**, i.e. What do we do to obtain successful results? Governance models include taking decisions about the preservation strategy to be implemented by the institution, and adopting a set of practical solutions and tools to preserve and manage digital data. To be successful, they often need to be implemented step-by-step.
- **The scope of the preservation**, which includes questions like:

- What needs to be preserved (images, text, videos, datasets) and what can be omitted?
- What are the targets of the preserved data and for what purposes should preservation be guaranteed (e.g. heritage amateurs, researchers, educational scope, maintenance of public records)?
- If other memory institutions are involved, should joint efforts be a desirable solution and if so, how should be coordinated?
- Is the private sector interested in the data to be preserved and how will that affect the situation?

Then, a set of technical decisions should be taken at different operational levels:

- How will the preservation workflow be implemented in the institution?
- Which file formats should be considered?
- What are the properties of the digital files that are relevant for the scope of the digital archive?
- How will files be stored and by whom (e.g. in-house at an internal data centre or by out-sourcing to external e-infrastructural services)?
- How will the preserved files be made discoverable and retrievable by the targeted audiences?
- For how long should the files be preserved (e.g. for 1 year, 5 years, 10 years, 100 years, forever)?

However, decisions on governance issues are not enough. A set of practical solutions must also be adopted, which include tools and procedures that keep the underlying technology alive or to update it.

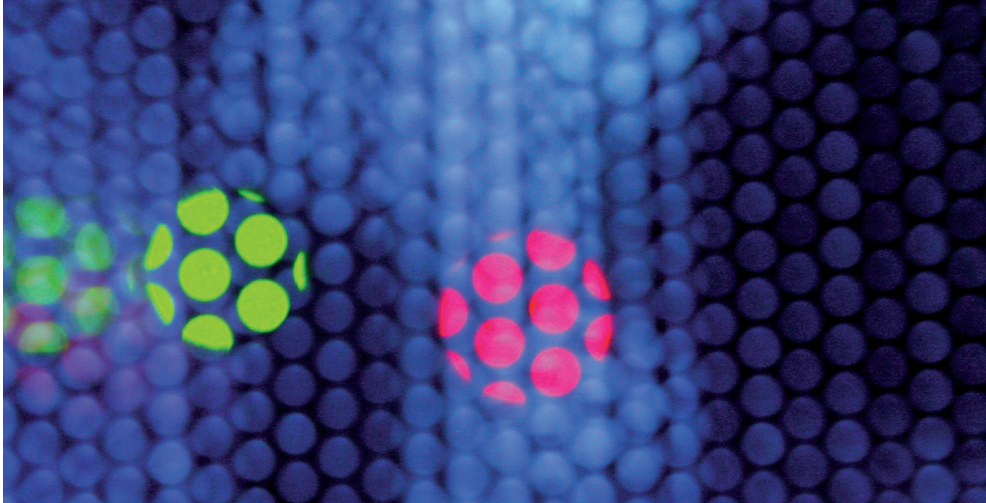
Today, a broad range of preservation software and hardware tools are available, and institutions can combine and tailor the digital preservation components according to their needs and context. To define an initial set of critical system requirements is a good start in establishing necessary conditions for a suitable and practical preservation environment. Examples of such requirements are:

- Assorted issues, like reliability and robustness, ease of use, scalability and flexibility
- Content and metadata issues, like mechanisms for integration and automaton of appraisal and ingestion of digital objects
- Performance issues, like scalability
- Trust issues, like authenticity and integrity of data, and identification of digital objects in danger of becoming inaccessible due to technical obsolescence
- Infrastructure-related issues, like dependencies of external systems and sources
- Hardware-related issues, like support of multiple storage media and devices, back-up and restore functions.

7.2. DEFINING POLICY RULES

This section will focus on the policy decisions to be addressed by the memory institution and how the PREFORMA tools can be of help in this process.

The definition of policy rules will allow the memory institutions to carry out a policy-based assessment,



which is a fundamental pre-requisite for an effective implementation of any preservation strategy.

Policy decisions may answer the following questions, that are at the basis of the preservation strategy:

- In what legal framework do you operate?
- How do you put in place the necessary trust-building processes?
- What are the standards to be supported in your digital archive?
- What channels do you use to give access to digital objects?
- What information should be documented (e.g. frequency of usage, typologies of users, accesses)

113

The policy requirements are specific to each institution and normally, are only partly covered by file format standard specifications. Therefore, the PREFORMA tools allow memory institutions to define their own specific policy requirements.

Through very easy-to-use user interfaces, the digital archive manager can define the policy rules of his or her specific institutions, or if needed, a selected part of its holdings. The characterisation of the individual

policy requirements will be associated with the specific content type of the digital archive or part of it, and with the workflow of the institution. Then, each conformance checker will verify how the format of the files corresponds to the policy rules established by the institution and will produce a report about any experienced deviation.

7.3. IMPLEMENTING THE PREFORMA CONFORMANCE CHECKERS

114

Detailed descriptions on how to install the three Conformance checkers developed in the PREFORMA project are provided in Chapter 6. This sub-section will concentrate on the process of implementing them.

After 20 years of digitisation, many memory institutions in Europe are still lacking strategies and operational solutions for digital curation. For those institutions that already have digital objects to preserve, but no process or systems in place for dealing with them, the situation can easily get out-of-control and become unmanageable.

A **first step**, before starting the process of implementing the Conformance checkers, would be to make sure that at least the basic digital preservation issues have been considered, including strategy and policy, formats to use, and storage technology.

A **second step** would be to identify and address internal constraints that in one way or another could set the implementation process at risk, for example:

- Dominance of an analogue mind-set among the curators

- Lack of business awareness to achieve economies of scale; doing nothing or just enough to solve today's problems, will in the longer-term be a waste of resources
- Lack of internal support in setting priorities for digital preservation and for developing strategies and routines (especially long-term); Also, clarification of internal roles is always needed (e.g. how should responsibilities be divided between administrators and information managers and their IT counterparts?)
- Lack of training for staff
- Limited IT resources (staff, hardware and software)

A **third step** would be, given the outcome of earlier steps, to choose (firstly) which Conformance checker or checkers to implement, and (secondly) an alternative for its implementation. The alternatives to choose between are:

- Stand alone, which allow for packaging the conformance checkers in an executable to be run on a PC; this is preferable for smaller organisations without a centralised IT infrastructure (in-house or out-sourced)
- Networked, which allow for deployment in network-based solutions for digital repositories (dedicated server, cloud solutions)
- Integration in legacy systems, which allow for plugging them into proprietary legacy systems via API.

To get the chosen Conformance checker not only up and running, but also to fully explore its usefulness, requires more than just technical skills. Besides knowing what to check, why and how, conformance checking also requires knowledge about the chosen format or formats,

and this is something that many curators of digital objects do not possess.

A recommended **fourth step** would, for that reason, be to ensure that the implementation goes hand-in-hand with an internal build-up of robust know-how and skills in preservation formats. Tutorials and practical handbooks are, to some extent, already in place, but will increase in number as the use of the PREFORMA Conformance checkers expands.

National memory institutions and policymakers on various levels will have an important impact on the implementation process by taking on the responsibility to organise training courses and seminars.

8. CONCLUSIONS

This chapter gives a cursory summing up of the discussions in the previous chapters, by briefly describing the current situation in digital preservation and suggesting some recommendations about issues connected with long-term preservation of digital objects.

8.1. DIGITAL PRESERVATION TODAY AT GLANCE

The **current situation** in digital preservation can briefly be described as follows:

117

The results of 20 years of digitisation represent a considerable challenge for memory institutions, although the significance of preserving digital information is understood by most users of IT-technology. Many memory institutions in Europe have neither implemented internal strategies and operational solutions for long-term digital preservation, nor safeguarded their digital holdings and collections according to international standards for digital archiving. This could be devastating for the possibilities for future access and use of the digital objects they have in their custody.

It is an established fact that the amount of digital information worldwide is rapidly growing. Thus, the number of transfers of digital objects to memory institutions is expected to continuously increase over time, which will cause higher costs for preserving digital objects long-term. The economic challenges of carrying out digital preservation are already substantial and can

easily become unmanageable, especially for institutions who are already receiving digital objects but lack specific programmes, processes and systems for treating them long-term.

Many smaller memory organisations and institutions with digital preservation programmes, but little or no expertise in the preservation field, are in a situation where they must rely on market-conform technology. Thus, they run the risk of having limited control over the digital objects they should preserve and may not know the exact needs or requirements when setting requirements for long-term sustainability of file formats.

On the positive side, there are several strategies available today that allow digital objects be accessible and useable over time and exist beyond the lifetime of the systems that created them. Most notable, is the migration strategy, often used in combination with the OAIS reference model.

118

Standardised file formats are also normally used so that the digital objects to be preserved avoid technical obsolescence and remain accessible and usable into the future. However, a disadvantage is that today's solutions for validating file formats are not totally reliable, which could be dangerous for the entire preservation process.

To solve this situation the PREFORMA project provides a full set of tools to support memory institutions in implementing high-quality standardised file formats. These tools will give memory institutions full control of the process of the conformity tests of files to be ingested into their archives, by giving a guarantee that the content is produced according to standards.

The PREFORMA solution is also in line with all preservation strategies that exist today.

8.2. RECOMMENDATIONS

Firstly, the process of preserving digital objects long-term includes several fundamental steps to consider. These are:

- Selecting the data to preserve
- Keeping the data alive, but also the meaning of it, its context and its dependencies to other systems and data
- Maintaining trust in the data
- Establishing and maintaining good governance of the data

Secondly, preserving digital objects long-term also requires that institutions:

- **develop a preservation strategy:** Such a strategy will be a combination of approaches that consider the specificity of the institution, e.g. the type of digital objects, their life duration, ownerships and rights, and ways of access.
- **take decisions** in accordance with the strategy and deploy tools to implement these decisions.

Thirdly, when memory institutions **integrate the PREFORMA tools in their preservation programme**, they can take full control of the process of validating file formats. For those institutions who have to migrate digital objects from old repository systems to new ones, this validation will also ensure quality of content and will document the risks associated with file formats.

BACK TO THE FUTURE? DIGITAL PRESERVATION NEEDS OF FUTURE USERS ANTICIPATED TODAY

By Milena Dobrevá (UCL Qatar) and Raivo Ruusalepp (National Library of Estonia)



The rapid increase in the volume and complexity of digital objects and data requires solutions that tackle not only the current needs in digital asset management, but also provide sustainability and interoperability into the future.

Although there is a general awareness that our collective digital holdings need to be preserved for posterity, this is a complex challenge, especially in the memory institutions sector, where many institutions are still facing the difficulty of formulating their strategy. One example of this is the fact that written strategies on long term preservation are still not developed in the majority of the memory institutions in Europe. An ENUMERATE survey found in 2015 that 26% of the respondents among 724 responding European institutions had a written digital preservation strategy in place, see [Enumerate 26:2015]; the responding institutions are likely to be among those

which are more active in the 'digital' domain. This survey also found that 47% of the institutions are not following internationally recognised standards in this area, with only 21% having their own archive. In the group of respondents, the national libraries were the most advanced in strategic thinking, 90% had strategies and infrastructure in place to handle digital preservation.

These results indicate that digital preservation continues to be an area where many institutions struggle. While the evidence from such surveys provides a glimpse into the reality, the key to advancement in this area is in understanding the reasons for this slow adoption of standards and solutions, and for expanding institutional-level policy work in the domain of digital preservation. Understanding the reasons would require a different type of study, but the anecdotal evidence puts forward one recurring theme: the current services and functions of memory institutions are constantly under stress due to the current rate of technological advancement, and compared to other technologically-driven areas, digital preservation is notoriously complex to understand, strategize, and implement. The abundance of frameworks and approaches is testament to experimental field, and as such it requires strong technological support team within the institutions – and not all institutions have such teams at their disposal.

121

Thus, our professional community, which needs to implement preservation solutions, has arrived at a time when we need a shift in the approach to digital preservation. Once people are aware of the problem, after the stage of developing multiple frameworks, approaches, and technological breakthroughs, the next step is to provide the cultural sector with solutions which are easy to implement, monitor, and integrate in the existing technical infrastructures.

Within this context, PREFORMA made a major step towards providing much needed and easy-to-use tools for addressing the specific issues: the conformity of formats of digital objects to be preserved, and how to take a holistic approach understanding the points of view

of different stakeholders. Let us imagine that some of these stakeholders are meeting at a digital preservation event; those who have spent some years working on digital preservation can easily imagine the following dialogue:

Memory institution professional: We are well prepared to preserve our analogue collections. Though this expansion of the digitisation programmes is putting us under huge pressure to become digital asset managers on top of everything else we must do! What is the best solution which would not require too much staff effort and resources on our side?

Developer (external): Haven't you tried this new preservation system yet?

Developer (in house [if there is one]): But that would require us to make substantial changes to our existing digital asset management, and the back-end interfaces.

Standardisation specialist: What are you all talking about? This new preservation system is not even following OAIS in detail. How are you going to make sure your digital archive will be interoperable with the national framework which is still being developed?

Policy maker: It is clear you are working in a very complex environment.

(Meaning: "Oh dear! Do these people know what they are doing? I am sure they do, but how are they going to achieve it? Should we wait until they have a robust solution before approving the funding for their next large-scale digitisation project? What do we need to do to ensure that these collections do not end up hidden and potentially lost?") With this tough question the policy maker decides to top ups on coffee and leaves the group.

Memory institution professional: Yes, we are aware there are many tools and different standards, but we cannot invest in expanding our team. Are there any free solutions that do not require us to change our own environment, but will help in specific tasks? For example, when we ingested the collection of our regional branch, we discovered that about 2% of the objects would not render. It looks like a minor issue, but this means that 1 in 50 objects will be unusable and we started getting complaints about objects which cannot render from genealogists who were browsing our digital collection on a systematic basis. How can we make sure that we ingest digital objects that do not have any issues?

The different points of view and vested interests are there, but how easy it is to resolve them? If we speak of end users (which is different from stakeholders), the traditional approach of technological solutions designers of is to capture the needs and expectations in a formal way, which is translated into functional requirements. These determine the specific solution will be doing, and how exactly it will be interacting with the users, as well as the data it uses.

123

The needs of the stakeholders do not necessarily mean they will be using a specific solution, but they still translate into requirements. For example, if standardisation bodies expect that digital preservation systems follow internationally or nationally agreed standards, this would influence the functional requirements. In this sense, PREFORMA is pushing the boundaries of digital preservation thinking and understanding. Multiple digital preservation projects and initiatives were addressing the complexity of this domain answering to the concerns and needs of one or two of those stakeholders. A substantial advantage in PREFORMA is that it not only develops tools, but also looks at the complex needs of various stakeholders, while also considering the end users.

Why is understanding the roles and needs of stakeholders so important and why it is so important at this point in time? To answer this question, we should have a clear

idea about the development of digital preservation so far, and also predict what the main drivers that will further push developments in this area will be. After all, what we preserve today would need to not only be stored and available in the future, but also, to be usable. This means that future memory institution professionals, standardisation specialists, and developers will continue the work that is being done today.

If we were to know – or at least to guess – what the requirements of the future would be, we would arrive at an additional layer of functional requirements.

124



Figure 36: Layers of requirements and expectations

Although we assume that every layer of requirements and/or expectations is expanding on the previous one, the actual picture would be more complex with partial overlaps – and even potentially contradictory requirements in the future, compared to what we can capture today. For example, standardisation is one of the key guiding principles today, but the push for interoperability can distort its boundaries, implementing solutions that aim to integrate any objects independent of how ‘standardised’ their formats are.

This leads us to a perceived ‘slushiness’ of what digital preservation really means. On one hand, we can say that the ‘acid test’ in digital preservation is the ability to use a digital object in the future. But ‘use’ itself is a complex notion – do we mean using it exactly in the same way as it was used originally? Or in a different way? And if we mean a different way, what do we mean exactly? (Not knowing exactly what this means, we cannot translate our thinking into functional requirements). How can we do better, when our present understanding of future technologies and modes of use are limited? Unfortunately, as a professional community, we are not equipped with crystal balls or other magical tools that can help us see what future end users will want, or even more, what the stakeholders will want. What will the memory institutions of the future look like? Will there be developers? Or will there be gigantic aggregators and information brokers taking over functions currently implemented by memory institutions?

Looking into the past might seem easy, as it is well-documented in the academic literature [for a survey of the EC-funded projects see Strodl et al. (2011) and Harvey (2015)]. Presently, we can say that preservation has advanced and addressed different types of objects, research data, software, and processes. It also tried different strategies for preservation (see Fig. 37), as already discussed in Section 3 of this report [Dobrev, Ruusalepp (2012)].

125



Figure 37: Main preservation strategies of today

While we cannot predict all the solutions required in the future, we can definitely summarise what factors will influence the needs of the future. We can approach this in two ways. The first one is to look at the constituents of the digital content area – and here we can use the Digital Library Reference Model [DELOS DLRM 2007], which summarised the main domains and their structure a decade ago.

DELOS DRM domain

What will change in it?

Content

Content will continue to diversify and become more complex in terms of objects which combine different formats and datasets. The care for datasets will be expanding with the growing interest to open science and research data management. Content also will be enriched with data coming in growing quantities from the users. The handling of this type of content is still experimental and debatable but the expansion of participatory approaches will push towards collecting and storing user generated and contributed content

Users

The end users will diversify in terms of geographic origin and skills. If we look at the current typical uses of digital cultural content for research, teaching, and leisure, probably these broad categories will stay but the way people communicate with digital content will be changing according to the specific context of use and tasks. The ease of use and re-use of digital materials will be the guiding drive for development

Architecture

The architectures will definitely change and move to larger scale solutions, to mention at least one obvious development

DELOS DRM domain	What will change in it?
Functionality	The functionality of today is mostly to retrieve objects from a digital archive. We can expect further blending of functionality with the actual context of use where the user does not necessarily have to realise they are retrieving an archival item which takes several steps before they can use the object
Quality	The ideas of quality are expanding and changing with the progress in various domains. Currently the preservation community is concentrating on quality of individual objects – in the future we could expect that the quality of the overall user experience will be more in the focus
Policy	Here one can expect growth; if about ¾ of the European institutions who responded to the Enumerate survey had policies in the preservation area, there is a scope to expand the policy design and inception. An interesting question is what will be changing in the nature of the policies. This will be linked to the changes which are happening in the overall memory sector where the mission and remits of memory institutions are currently undergoing major shifts

127

Table 6: DLRM domains and anticipated changes which will need to be accommodated in the future

The second way we could explore the trends for the future is to look at areas outside of the digital library domain. Here we can list several factors that will be changing in the near future:

- **Changes in the stakeholders involved and their roles** as already mentioned, will influence the future expectations on preservation solutions.
- **Scale (big data).** While memory institutions are still defining their strategies, the ongoing large-scale digitisation pushes the boundaries to big data, which will have an added volatility with the advancement of participatory approaches in cultural heritage. Another big influencer will be the aggregators. Considering that some 10% of European cultural heritage is estimated to be available in digital form, there is a substantial space for growth.
- **Open science.** The call for sharing research data, methodologies, and publications and the growing involvement of citizen scientists, citizen historians, citizen archivists and curators will push the boundaries of preservation as well. For example, in the area of digital humanities and arts, the consultation with digitised primary sources is a robust part of the research lifecycle. This again puts the focus on the contexts of use.
- **Costs.** This one is hardly a surprise, but with all the large-scale work and expensive technological environments comes the issue of how much it actually costs to preserve. The EC already funded projects like 4C (n.d.) that looked into cost models and we can expect more work in this direction.
- **New skills.** The EC is researching the training needs of data curators, data librarians and data scientists. The requirements for new skills are often discussed in relevant conferences and is addressed in the academic literature – see e.g. Harvie, Mahard (2013). Preparing new higher education programmes, as well as professional training, will require continuous upskilling to keep up-to-date with new developments. The educational component is key for robust professional service in the future.

Finally, digital preservation discourse has, for a long time, been focussed on longevity and sustainability issues of file formats, workflows, tools, processes, etc. This has resulted in a plethora of tools and services that are tailored for a specific object type, support a particular task, or require a specialised skill-set to implement. Not only has it left preservationists, especially those new to the discipline, perplexed about what works best for what preservation situation, it has also diverted the attention away from the systems level. The core condition of digital preservation is that information needs to live longer than the system(s) that created it. Increasingly, software is required along with the information it contains for evidence and preservation purposes. Academic journals require researchers to submit not just the underlying data, but also automated workflows used to process the data and arrive at the reported results, and software for enacting the workflows. Smartphone apps that are used to publish newspapers and magazines fall under the collection remit of libraries. Publishers increasingly demand automated interfaces for bulk deposit of digital legal deposit copies of their publications.

129

When defined as a systems-level issue, digital preservation becomes largely an interoperability challenge. For example, content objects and services based on them in one repository system can be migrated to a new repository system by defining interoperability requirements and standards that support them. An example of this approach is offered by the recent eARK project [eARK (n.d.)] that took interoperability as the binding concept for the preservation toolset it developed and has established an Archival Standards Board (<http://www.dasboard.eu/>) to ensure longevity of the standards that support the interoperability. The UNESCO initiative PERSIST is defining solutions for software preservation and interoperability of different platforms (<https://unescopersist.org/tag/digital-preservation/>).

Standards-based interoperability as a means of overcoming obsolescence of systems brings resilience into preservation spotlight. Rather than longevity or sustainability of digital information, the next stage

of maturity of digital preservation domain should focus on **resilience** of systems and information created by them. Resilience in this context can be defined as the capacity to prepare for disruptions, recover from shocks and stresses, and adapt and grow from a disruptive experience. The first generation of digital repositories and digital preservation systems are reaching an age where they can be categorised as legacy software. The process of replacing them will be undertaken by many memory and academic institutions in the coming years. Defining resilience conditions for preservation systems as interoperability requirements when migrating between systems, would help to conceptualise digital preservation in new ways and ensure that this domain is future-proof. Building PREFORMA tools into the migration process between repositories would be an excellent way of ensuring quality of content and documenting risks with file formats.

This brief discussion on factors that will shape digital preservation in the future shows an exciting field of opportunities. Within this area, PREFORMA is an interesting development since it presents a solution that is consistent with all existing preservation strategies (see table 7).

Approach	Relevance of PREFORMA tools
Incremental	PREFORMA tools are useful for processing of objects in batch mode. This is relevant to all tasks of mass ingestion of digital objects into a digital archive. The tools can easily fit into the discussion of pre-ingest requirements stipulated in OAIS and PAIS.
Techno-centric	PREFORMA tools would be part of media renewal while checking the state of objects which are being copied

Approach	Relevance of PREFORMA tools
Analytical	PREFORMA tools can be used as part of the digital forensic tools which are used to establish the nature and state of a specific digital object
Durable digital objects	The concept of PREFORMA tools is applicable to durable digital objects as well – in this case PREFORMA tools can be integrated into the workflow which enhances a digital object to become durable, allowing to do checks on format compliance

Table7: PREFORMA tools in the context of various preservation strategies

Finally, where does this all lead when we consider the current situation and needs of memory institutions? The need for efficient and easy to implement solutions is one of the key present needs. PREFORMA demonstrated how the expectations of different stakeholders can be managed with tools that address the specific issue of format conformance . Its approach, as an innovative procurement framework, can be used as a model for further digital preservation work. In its approach, PREFORMA addresses diverse needs and also demonstrates that is solutions blend well with all preservation strategies that had been tried and tested so far.

131

REFERENCES

- 4C project (n.d.) Collaboration to Calculate the Costs of Curation. Project website <http://www.4cproject.eu/>
- DELOS DLRM (2007) The DELOS Digital Library Reference Model. Version 0.96, November 2007, http://delosw.isti.cnr.it/files/pdf/ReferenceModel/DELOS_DLRReferenceModel_096.pdf

Dobreva M, Ruusalepp R (2012) Digital preservation: interoperability ad modum. In: Chowdhury GG, Foo S (eds) Digital libraries and information access: research perspectives. Facet, London, pp 193-215

eARK project (n.d). eARK project website. <http://www.eark-project.com/>

Harvey, R. (2015). 'The Last Decade of Digital Preservation: A Personal View from Australia', Preservation, Digital Technology and Culture v.44 no.1: 22-30.

Harvey, R., Mahard, M. (2013). Mapping the preservation landscape for the twenty-first century In: Preservation, Digital Technology and Culture, 42, 5 - 16

Nauta, G.J, vn den Heuvel, W. (2015) Survey Report on Digitisation in European Cultural Heritage Institutions 2015. DEN Foundation (NL) on behalf of Europeana/ENUMERATE, Public version, June 2015.

Strodl S., Petrov, P., Rauber, A. (2011). Research on digital preservation within projects co-funded by the European Union in the ICT programme. Technical report, Vienna University of Technology, May 2011.

ABBREVIATIONS

AIP - Archival Information Package

API - Application Programme Interface

AVI - Audio Video Interleaved

CCITT - Consultative Committee for International Telephony and Telegraphy

CLI - Command Line Interface

CMYK - Cyan, Magenta, Yellow, Key black

CSV - Comma-Separated Values

DCH-RP - Digital Cultural Heritage - Roadmap for Preservation

DEI - The Department of Information Engineering of the University of Padua

DIP - Dissemination Information Package

DManager - The PREFORMA tool to check the conformity of TIFF files

DURAARK - Durable Architectural Knowledge

EBML - Extensible Binary Meta Language

EAD - Encoded Archival Description

EC - European Commission

EU - European Union

EXIF - Exchangeable Image File Format

FFmpeg - The free software project that produces libraries and programs for handling multimedia data

FFV1 - FF video codec 1, a lossless intra-frame video codec

FP7 - Seventh Framework Programme of the European Commission for the research and technological development.

GPL - General Public
Licence

GUI - Graphic User
Interface

HTML - Hyper Text Markup
Language

KIK-IRPA - The Royal
Institute for Cultural
Heritage in Belgium

ICC - International Color
Consortium

IDC - International Data
Corporation

IEC - International
Electrotechnical
Commission

IIM - Information
Interchange Model

IPTC - International Press
Telecommunication Council

ISO 14721 - Space data
and information transfer
systems --Open archival
information system (OAIS)
- Reference model

IT - Information
Technology

Libav - Cross-platform
tools and libraries to
convert, manipulate and
stream a wide range of

multimedia formats and
protocols

LPCM - Linear pulse code
modulation, a method
for digitally encoding
uncompressed audio
information

LZW - Lempel-Ziv-Welch, a
universal lossless data
compression algorithm

Matroska MKV - Open
standard video container
file format

MediaConch - The PREFORMA
tool to check the
conformity of audiovisual
files

METS - Metadata Ecoding
and Transmission Standard

MIX - NISO Metadata for
images in XML

MPL - Mozilla Public
Licence

NISO - National
Information Standards
Organisation

OAIS - Open Archival
Information System

OSS - Open Source Software

PCP - Pre-Commercial
Procurement

PDF/A – ISO-standardized version of the Portable Document Format (PDF)

PPI – Public Procurement of Innovation

PREMIS – Preservation Metadata Maintenance Activity

R&D – Research and Development

SGDAP – The Records Management, Archives and Publications Service of the Girona City Council

SIP – Submission Information Package

SME – Small and Medium Enterprise

SPK – The Prussian Cultural Heritage Foundation

TIFF – Tagged Image File Format

TWG – Technical Working Group

veraPDF – The PREFORMA tool to check the conformity of PDF files

VLC – Mediaplayer, free and open-source, portable and cross-platform written by the VideoLAN project

VR – Virtual Reality

WebM – Open, royalty-free, media file format designed for the web

WebUI – Web User Interface

XML – eXtensible Markup Language

XMP – Extensible Metadata Platform

XSLT – Extensible Stylesheet Language Transformations

AUTHORS

Ashley Blewer, MediaArea.net

Peter Bubestinger-Steindl, AudioVisual Research & Development (AV-RD)

Milena Dobрева-McPherson, University College London (UCL) Qatar

Boris Doubrov, Dual Lab

Antonella Fresa, Promoter SRL

Lars Ilshammar, National Library of Sweden

Duff Johnson, PDF Association

Börje Justrell, National Archives of Sweden

Rolf Källman, National Archives of Sweden

Jérôme Martinez, MediaArea.net

Becky McGuinness, Open Preservation Foundation

Miquel Montaner, University of Girona

Víctor Muñoz, University of Girona

Bengt Neiss, National Library of Sweden

Bert Lemmens, PACKED VZW

Josep Lluís De la Rosa, University of Girona

Claudio Prandoni, Aedeka SRL

Da Guillaume Roques, MediaArea.net

David Rice, MediaArea.net

Raivo Ruuselapp, National Library of Estonia

Xavier Tarrés Bonet, University of Girona

Erwin Verbruggen, Netherlands Institute for Sound and Vision

Carl Wilson, Open Preservation Foundation

Benjamin Yousefi, National Archives of Sweden