DELIVERABLE

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

D3.5 Experience Workshop

Revision: Version 1.0

Authors:

Erwin Verbruggen (Netherlands Institute for Sound and Vision)

Contributors:

Magnus Geber (Riksarkivet)
Claudio Prandoni (Promoter)
Becky McGuinness (OPF)
Bert Lemmens (PACKED)
Eva McEneaney (LGMA)
Marju Niinemaa (KUL)
Sònia Oliveras i Artau (AJ Girona)
Stefan Rohde-Enslin (SMPK)
Anna Kasimati (Greek Film Centre)

Reviewers:

Erik Buelinckx (KIK-IRPA)

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

Revision History

Revision	Date	Author	Organisation	Description
V0.1	10/01/2017	Erwin Verbruggen	Sound and Vision	First draft for review
V1.0	30/01/2017	Erwin Verbruggen	Sound and Vision	Ready for publication

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise.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

TABLE OF CONTENTS

1 EXPERIENCE WORKSHOP REPORT: IMPROVING LONG-TERM PRESERVATION 4

1.1 MORNING SESSION 4
1.2 AFTERNOON SESSION 9
1.3 EXHIBITION AND GOODBYES 13

ANNEX I EXPERIENCE WORKSHOP PARTICIPANTS LIST 14

ANNEX II EXPERIENCE WORKSHOP SLIDES 17

1 EXPERIENCE WORKSHOP REPORT: IMPROVING LONG-TERM PRESERVATION

Le parvis d'entrée de la Gemäldegalerie (Berlin), cropped version of image by Jean-Pierre Dalbéra, CC BY, 2013.

The Experience Workshop[1] took place Berlin on 23 November 2016, in connection with the second Prototype Demonstration. The PREFORMA partners shared their experiences of working with suppliers under R&D service agreements with other memory institutions.

The morning session focused on use cases for conformance checkers from memory institutions, while the afternoon explored the PREFORMA challenge with an overview of the upcoming testing phase. This was followed by presentations from the three suppliers who develop the conformance checkers.

1.1 MORNING SESSION

About PREFORMA: Antonella Fresa (Promoter Srl) & Börje Justrell (Riksarkivet)

Antonella Fresa, technical co-ordinator, welcomed the workshop audience to the PREFORMA Experience workshop by detailing what the Pre-Commercial Procurement (PCP) project, co-funded by the EU, stands for. She explained how the three suppliers to the project had been selected through a competitive tender process and how their software must meet the strict tender as well as memory institution's requirements. She noted that thus far the work carried out by the three suppliers was in line with their expectations and that the PREFORMA team is keen to talk to users for further feedback on the results.

Börje Justrell, project coordinator, introduced the project partners and their approach: all conformance checkers are open source to help establish a community around the software and ensure it is available after the project ends. They are available under two open source licenses: GPLv3 or later and MPLv2 or later. The PREFORMA vision defines three preservation layers: bit

[1] All presentation slides and more photos from the workshop are available at: http://experienceworkshop.PREFORMA-project.eu/programme/.

preservation, logical preservation and semantic preservation. The project is focusing on the logical layer.

Ms. Fresa

Mr. Justrell

Mr. Kulovits

Keynote: Hannes Kulovits (Austrian State Archives)

Hannes Kulovits from the Austrian State Archives continued the day with a keynote talk on *digital preservation and conformance checking at memory institutions*. The organization has been at the forefront of implementing a digital repository for the preservation of digital records. Since the introduction of electronic records management in the federal administration in 2004, all records have been born digital and must be preserved as such. Its mandate to preserve includes everything that is coming from the federal administration. Records are transferred to the Austrian State Archives after a period of 10 years and they are kept restricted for 30-50 years. Objectives: make available records on the website and for long-term preservation. All records that have been created in digital form are legally valid.

One of the main challenges the organisation faces is that the different federal government departments do not all follow the same procedures when supplying their records. There is no consistent policy for records management and there are no file format restrictions – the State Archives has to be able to handle everything. Ministries are free from directives, and can do what they want. There are no file format restrictions. The archive has to be able to handle every single file format, from simple to complex objects. Compared to other European archives, the Austrian State Archives does not have policy powers over records management issues in the federal administration. Consequently, the archive needs to be prepared to accept and preserve a wide variety of file formats of different generations.

One of the most important parts of the preservation process is the ingest part, i.e. to preserve the provenance of a digital object taking into account the various migrations. Conformance checking is a critical step in the ingest and quality assurance processes within its repository. Digital Preservation Managers must have certainty that a digital record or file is what it purports to be to make a well-informed decision on which preservation measures to take. The Austrian State Archives use the OAIS model and the Preservica system for its workflow management, which enables it to use specific tools for migration, configuration and customisation. The two core characteristics that the Austrian State Archives implement in their ecosystem are preservation

planning and preservation operations. It is important to look not only at the file format but also at the technical properties of the digital files to preserve e.g. authenticity.

The preservation planning approach starts from the selection of the content type, passes through the selection of the properties, the detection of high-level issues and the selection of samples to be used for the testing. How difficult is to select and identify the file format? It could be difficult in particular for the old file formats. Are there any license costs? Are they supported by the browsers? Does the file conforms to the standard specifications? It is difficult to extract this information from the file in a way that you can trust it. There are many properties that needs to be checked to do quality assurance, and it also depends on the file format.

The format of the record has been standardised at a federal level with the (XML-based) EDIAKT II standard, which contains a set of metadata describing properties that are used for long-term preservation. On the one hand, it is a very structured format with many metadata, on the other had everything can be described. To solve this problem, the record must conform to a specific structure that they define (e.g. is the number of business cases equal to one? is the record closed? does referenced document exist? etc.). These are simple XPATH expressions which model simple checks. Hannes explained that the significant and technical properties of a file are key to maintaining their authenticity. The properties the archives focus on are:

- **Format and Sub-format:** Assessing the risks for different files formats such as the quality, availability and price of the specification, the number of open source tools available for identification and validation, and the licence cost.
- **Representation Instance Properties:** Whether the object is valid and well-formed, what size is it and is it searchable? Is the metadata valid, and does it conform to the format standard?
- **Information Properties:** How many pages does it have? Does it contain footnotes, or a table of contents? If it's an image, what size it is? How many bits per sample are there?

The archives also look at the record properties and check when the last change was made, what the subject area is and the date it was archived. Hannes gave an example of a use case when they received a digitally signed PDF that rendered different on different computers. It transpired that the font was not embedded. This is a problem for digital preservation as it means the structure of the document can change. The archives mostly use JHOVE and DROID to validate their files, and are planning to evaluate veraPDF and FITS. They are keen to try identification and validation tools as possible on their repository.

Hannes summarised with a few points that are central to the success of digital preservation:

- know WHAT you have,
- be sure that digital files are what they purport to be,
- know WHAT action to take and HOW.

Use cases: Matthias Priem (Vlaams Instituut voor Archivering) & Klas Jadeglans, (Riksarkivet MKC)

Up next we heard use cases from memory institutions, and digitisation and archiving service providers. Conformance checking is considered an important step during ingest workflows and ingest reporting. For digitised material, it is easier to agree on a single, or limited number of formats to ingest, however as born digital content is created by many producers and technology advances quickly, it is difficult to enforce a single or limited number of formats to accept. Conformance checker testing has been carried out by several external organisations. They

presented some of their initial results and gave feedback on how the software could be improved. Also, the PREFORMA memory institutions are testing the tools and looking into adapting their workflows to integrate the project's conformance checkers.

Mr. Priem

Mr. Jadeglans

Matthias Priem explained the current approach for file formats and validation of these formats at VIAA. He illustrated both the achievements to date and challenges they are facing, and how they saw the potential of the work done in PREFORMA. VIAA does digitisation, archiving and dissemination, particularly (but not only) in the broadcasting domain. VIAA provides services but does not manage collections themselves. 500.000 hours of media are currently being digitised and a WWI newspaper project to digitise pages of newspapers is ongoing. Actual ingest in the archives from April 2014 until today has been growing exponentially, in particular when they started ingesting digital born material from public broadcaster VRT. Archiving takes place at two locations and a third location to store material. File management is orchestrated by a MAM system (commercial software). They have licenses to distribute content for education - for schools and scholars to use material stored in the archive. Newspapers of WWI, also available online where it is possible to browse the pages.

File conformance checking is a different process for its digitised and born-digital content. For digitisation, VIAA tries to agree on a single format. Its most common format for the wrapper is MXF, either JPEG2000 or DC-10 wrapped in MXF. VIAA implements a wide testing of the files that arrive in the archive (SIP) with a report assessing the results of the evaluation. This is done either manually with a Google Sheet or through the MAM system itself (which is more limited). Finally, they transcode every file that arrives in the archive. If you are able to transcode a file, you can also decode it. However, they cannot control this validation and transcoding process. For digital-born content people send whatever they want/have and it is difficult to enforce a certain format. Checks are made when the file arrives in the archive (SIP) checking the codec and transcoding the file. Checks are made at different points, mostly through the MAM system, but here they do not control the process.

The organisation considers MediaConch as an interesting candidate to replace this Google Sheet because it does the same things and much more. An in-depth codec check is missing. VIAA is

evaluating FFV1 as a possible candidate to replace JPEG2000, but they want to have hands-on experience on it to evaluate how this would impact on their collections. Experimented with MediaArea with 100 files in JPEG2000 / MXF format (about 3TB): JPEG2000 to MP4, JPEG2000 to FFV1, FFV1 validation through MediaConch, FFV1 to MP4.

Conclusions: decoding to MP4 using ffmpeg seems to be 3 or 4 times faster than starting from JPEG2000. No much difference in space needed. Lossless transcoding seems possible. VIAA is investigating how to embed this in their workflow.

Klas Jadeglans works at the MKC - a digitisation factory that is a department of the National Archives of Sweden. They are producing images, not storing them. Focus on high volume production of digital images (~100.000 images per day): loose sheet, bound books, newspapers, maps etc. They have produced more than 200 million digital images since they switched from microfilm in 1995. The production flow system is a web browser GUI. It is composed both of automatic and human steps/modules. The first part is the production phase, where humans are involved. Here the material is registered, prepared for scanning, and scanned.

The second part is the post-production phase. Here the images are converted in the desired format (JPEG, TIFF, etc.), packaged, and delivered. Here there is no human intervention at all. DPF Manager can be integrated in the post production, before the packaging and delivery to the customers. Question is whether they need to validate something they are creating by themselves (created and added metadata). Answer is yes. DPF Manager will take 16 seconds to check an image, which is quite a long time, but you have batch processes which are very quick (0.45 seconds per image). In conclusion: Klas shared his wishes for the software to better integrate in a real environment:

- make install easier for non-graphical systems;
- totally non-human mode with an OK message if everything was fine;
- option to select output with the possibility to check each single file and to have a summary of the results;
- allow unlimited batch sizes;
- knowledge base: what do I have to do with warnings and errors?
- make it friendly to ‘dummies’;
- very interested in TI/A - who says what is correct and what is not.

Experiences from the PREFORMA memory institutions

Netherlands Institute for Sound and Vision

Erwin Verbruggen from Sound and Vision told about their experiences in the project and how the institution originally hoped to have MXF included as an open format in the tender. Now they are very impressed with the work MediaArea is doing and investigating to what extent the format could be adopted in its repository.

KIK-IRPA

Erik Buelinckx shared the lack of resolve at KIK-IRPA about decision making in the kind of TIFF files they need to handle and preserve. Over the years it has not been clear which specific flavour of TIFF-files have been generated. Only recently a set of guidelines was defined. Photographers currently upload their TIFF files to the server, where there is an automatic procedure to process them. DPF Manager would add a welcome layer to this process to check the quality of the TIFF file.

Digital preservation? Just press 'save'

Benjamin Yousefi, legal and technical adviser at Riksarkivet, did not have a background in digital preservation before he joined the project. When he began to consider the issues, one of the first questions was: how do we determine that a file conforms to the ISO 19005 standard? There are several validators available that try to answer this question, but there are discrepancies in the results so he could not advise which his organisation should use.

ISO standards are interpreted differently so it is difficult to decide which implementation is correct. The PREFORMA Challenge was to establish an objective point of reference. The three suppliers have approached this differently: EasyInnova could not change the specification for TIFF. They have created TI/A as a source for interpreting the specification. MediaInfo have been involved in the creation of the specification for Matroska, acting as a legislator for the format. veraPDF is working with the PDF industry. Through the PDF Association's Technical Working Group, they are resolving ambiguities in the specification. The PREFORMA Challenge is a research project that aims to establish an object point of reference

1.2 AFTERNOON SESSION

The PREFORMA Challenge: Bert Lemmens (PACKED)

Bert Lemmens looked back at the role of memory institutions. They have been preserving paper and ink for decades, and digital preservation is not a new problem. PREFORMA describes digital preservation as **taking precautions enabling long-term access to digital data.** This implies both policy decisions, implementing a sustainability strategy, and practical solutions, deploying tools to preserve and manage of digital data

He then discussed current strategies for digital preservation:

- **Do nothing**: There is often a good reason for this; organisations don't understand the problem or know what action to take. Their strategy is that by doing nothing, they are not doing anything wrong.
- **Conservation**: Apply what you know from preserving analogue material. Put on a shelf, keep at safe, sealed and confidential.
- **Documentation:** Preserve the software manuals and information with machines. In some cases, this is very useful.

All three are very passive strategies. More active strategies include:

- **Migration:** Replace the underlying technology to preserve the content. This is very hard – organisations need to decide which formats they should migrate to and keep up to date. Which formats are newer, better, more sustainable? How do you choose the format?

There is a huge list to choose from, each with their own risks, especially if you don't understand the format.

- **Piggybacking:** Follow what the large organisations are doing
- **Home brewing:** Create your own format or own solutions. This is more common in larger institutions and it creates a whole new range of risks. This approach depends largely on specific people within an organisation. When they leave, the knowledge goes too.
- **Evangelizing:** Look for better alternatives, convince others to do the same.

Memory institutions can lack knowledge about how file formats technically work, and often do not have control over the way they are produced, or the tools to manage the different types. The PREFORMA Challenge Brief was set up to enable memory institutions to gain full control over the technical properties of digital content intended for long term preservation. The main issues lie with the file format specification which describes how the format has been put together. It is a document written in natural language and therefore open to interpretation. There are other issues with the specification. They can be:

- Incomplete
- Inaccessible (closed)
- Planned for obsolescence (client lock in)

Conformance checking is defined as the process of **checking if the technical properties of a digital file are conforming to the specification of the corresponding file format**. Memory institutions need files with consistent properties to ensure authenticity of the content, simplify the management of collections and enable large scale migration and emulation. Three formats were selected for the conformance checker development: one text, one image and one moving image. The specifications were chosen because they are:

- Complete, and you can unambiguously point to one version
- Open – or subject to a nominal charge, but irrevocably royalty free
- Use reference implementations – have test files that show in practice what is and is not valid

Testing: Magnus Geber (Riksarkivet)

Magnus Gerber, Riksarkivet, explained how testing of the conformance checkers is carried out. The suppliers, PREFORMA partners and external members of the digital preservation community have tested the software during development.

The software is hosted on GitHub so it is open: the project and community can track progress and log issues and feature requests. Each supplier has made stable, monthly software releases and has received written feedback from PREFORMA after each formal release. PREFORMA has used a combination of organic and synthetic test files contributed by the partners, suppliers and external sources. The prototyping phase ended on 31 December 2016. A six-month scientific test phase will run from January – June 2017. There is still an open call for external partners to contribute to testing to help improve the software.

Suppliers: veraPDF, DPF Manager, MediaConch

Each of the suppliers gave a presentation about the latest developments in the software and an overview of plans for the future. For more information about each of the conformance checkers visit the PREFORMA Open Source Portal or take a look at the websites.

veraPDF: Carl Wilson (Open Preservation Foundation) & Boris Doubrov (Dual Lab)

Mr. Wilson

Mr. Doubrov

Interpretation of PDF/A standard is not clear and unique and there is a need of a trusted open source project. Industry support guarantees interoperability. Consortium: Open Preservation Foundation (lead), Digital Preservation Coalition (use cases), PDF Association (industry), Dual Lab (lead developer), KEEP Solutions (API). Components: validator, metadata fixer, reporter, policy checker. Specific test files have been created breaking the PDF/A specifications clause by clause. All information and material, including software, documentation, test files, etc. is available on GitHub. The version demonstrated today is 0.26. Version 1.0 is planned for mid-December.

GUI version is very straightforward. You choose a file and a validation profile (or auto-detection). HTML shows a summary of the tests and which ones failed in a condensed way. It is possible to know more on the errors through a dedicated wiki page. You can also select to have a features report, which is available only in XML. You can choose the features you are interested in. There is also a command line interface. -f specifies the validation profile (otherwise automatic detection will be performed). You can run a batch process to check multiple files at the same time. A summary line summarises the result of the validation and it is possible to get a features report too. The GUI will soon support batch processing too, the functionalities will be therefore the same between GUI and command line.

Finally, there is an online service. It also provides the possibility to choose the validation profile and whether you want a feature report. You can choose between HTML, XML and JSON report. There are two releases: PDFBox implementation (based on Apache license) and greenfield implementation, which was introduced in release 0.26 to accomplish with the PREFORMA requirement to license the software under the dual license MPL3+ and GPL2+. A new feature is

a unique Shell which is able to accept as input any kind of file and autodetect which conformance checker to call: veraPDF, DPF Manager and MediaConch. The architecture is plug-in base so to allow extensibility to integrate new features which are not contained in the PDF/A standard specifications, e.g. embedded fonts, ICC profiles, XMP metadata, image compression, etc.

DPF Manager: Miquel Montaner, Victor Muñoz and Xavi Tarrés (Easy Innova); Peter Fornaro (University of Base), and Josep Lluís de la Rosa (University of Girona)

Are you sure that your files are well created? Are you sure that your files are prepared for digital preservation? Consortium: Easy Innova, University of Girona, University of Basel.DPF Manager validate the following specifications: TIFF Baseline 6.0, Extended TIFF, TIFF-EP, TIFF-IT. Format preservation is very important and format migration is very dangerous as it implies transcoding. TI/A is a recommendation how to use the TIFF format at memory institutions. It follows an approach which is similar to PDF/A. A document has been submitted to the ISO TC-171.

DPF Manager GUI allows to select a file or folder to be checked and some configurations. A configuration wizard allows you to create a new configuration profile: you can select which standard specifications you want to validate the files against, you can define your specific policy or acceptance criteria by choosing among a set of pre-defined items (either highlighting them as errors or as warnings in the report), you can select a format for the report (HTML, XML, JSON, PDF), you can define fixers or changers to be applied to the image by choosing among a set of available fixers.

Mr. Montaner

Mr. Fornaro

It is possible to run several tasks in parallel and when a task has finished you can view the report showing the errors and warnings for each file for the implementation and policy checkers. It is possible to view more details about an error by clicking on the specific item. There is a page which shows you all the past reports. It is possible to schedule periodical checks. It is possible to configure external conformance checkers to allow checking PDF and AV files using the other conformance checkers. DPF Manager is available as a GUI and a command line standalone application and also as a web application. Current version is version 3.0. In the next version, several improvements are planned, including integration with Archivematica and a statistical analysis of the TIFF files checked in the memory institution.

MediaConch: Jérôme Martinez (MediaArea.net)

MediaConch can produce two reports, one about the implementation checker and one about the policy checker. It is possible to obtain general information about your file and to inspect the single files. There is a policy editor where you can create your policies. A registry of public policies is also available to share your policies and to browse existing policies that other people and memory institutions have created and shared. There is a fixer which implements simple fixes. MediaConch is not only a standalone product but it is also integrated in third-party systems, e.g. Archivematica. MediaConch provides a GUI, command line, web interface and a REST API.

Output formats are XML (native), Text, HTML and in the future, also PDF. It provides some XSL features that allow you to create a customised report. MediaConch relies on the MediaInfo software and it is licensed under MPL3+ and GPL2+. MediaConch supports MKV, FFV1 and PCM formats but it supports also QuickTime and MOV wrapped in MKV and other formats for the policy checker.

It has a plug-in mechanism that allows integration with other external conformance checkers. Input can be from local hard drive, FTP, HTTP, Amazon S3. Work has been carried out towards the standardisation of MKV and FFV1, e.g. with CELLAR IETF working group. Tests are being carried out with NOA and VIAA. A Matroska research corpus has been created analysing all MKV files from archive.org. Continuous improvements to handle huge collections, in the GUI, to include statistics and for the standardisation of MKV and FFV1 are already planned. MediaArea invites everybody to fork on GitHub and contribute to the improvement of MediaConch, either developing by yourself or asking for additional features to be developed.

Demo: you can upload your files from our file system or by providing an URL or check the files on the server, you can select your policy and the output format of the report. It is possible to change the HTML output by uploading your template. It is possible to download the report. You can view the results of your tests and view more details on each single item. You can create a policy from scratch or starting from the information that you retrieved for a specific item. You can add in your own policy repository every policy that is available in the public registry. It is possible to see the XML or HTML report also of the results of the other conformance checkers (veraPDF and DPF Manager).

1.3 EXHIBITION AND GOODBYES

After the presentations, the three suppliers were available for questions and further demos of their software – a format that had also been used at the Stockholm workshop. It led to interesting one-on-one conversations. For a full list of workshop participants, see Annex I.

ANNEX I EXPERIENCE WORKSHOP PARTICIPANTS LIST

First Name	Last Name	Company	Country
Hannes	Kulovits	Austrian State Archives	Austria
Elke	Meyer	IAEA	Austria
		Austria Count	2
Maksim	Bezrukov	Dual Lab Bel	Belgium
Boris	Doubrov	Dual Lab sprl	Belgium
Bert	Lemmens	Packed	Belgium
Erik	Buelinckx	Royal Institute for Cultural Heritage (KIK-IRPA)	Belgium
Matthias	Priem	VIAA	Belgium
		Belgium Count	5
Alvaro	Malaguti	Rede Nacional de Ensino e Pesquisa (RNP)	Brazil
		Brazil Count	1
René	Mittå	National Archive of Denmark	Denmark
		Denmark Count	1
Marju	Niinemaa	Estonian Ministry of Culture	Estonia
		Estonia Count	1
Jérôme	Martinez	MediaArea	France
Guillaume	Roques	MediaArea	France
		France Count	2
Matthias	Klein	ArchivInForm	Germany
Luisa	Orduno	Atelier Fantarium	Germany
Rainer	Jacobs	Bundesarchiv	Germany
Dietrich	von Seggern	callas software GmbH	Germany
Wolfhard	Hildebrandt	Deutscher Bundestag	Germany
Daniela	Leifert	Deutscher Bundestag	Germany
Thomas	Reidemeister	Deutscher Bundestag	Germany
Philipp	Gerth	Deutsches Archäologisches Institut	Germany
Zoe	Schubert	Deutsches Archäologisches Institut	Germany
Jorge	Urzua	Deutsches Archäologisches Institut	Germany
Monika	Hagedorn-Saupe	Foundation Prussian Heritage	Germany
Stefan	Rohde-Enslin	Foundation Prussian Heritage	Germany
Uwe	Kuehhirt	Fraunhofer IDMT	Germany
DANIEL	EKPAH	Freelance	Germany
Schubert	Schubert	Freelance	Germany
Anna	Schäffler	FU Berlin	Germany
Md Mostafizur	Rahman	German Archaeological Institute	Germany
Maren	Geissler	HELIOS	Germany
Frank	von Hagel	Institut für Museumsforschung	Germany
Marcus	Müller-Oertel	Landesarchiv Berlin	Germany

Martin	Hoppenheit	Landesarchiv Nordrhein-Westfalen	Germany
Ludger	Kemper	LuKe	Germany
Thomas	Zellmann	PDF Association	Germany
Alexander	Hundsdorff	PIQL Germany GmbH	Germany
Bastian	Waag	PIQl Germany GmbH	Germany
Andreas	Weisser	restaumedia.de	Germany
Hans-Joachim	Hübner	SRZ Berlin	Germany
Per	Broman	TU Berlin	Germany
Ulrike	Golas	Universitätsbibliothek der Technischen Universität Berlin	Germany
Yvonne	Tunnat	ZBW Kiel	Germany
		Germany Count	30
Marios	Phinikettos	National Technical University of Athens	Greece
		Greece Count	1
Maurizio	Fedele	Onlus ICT Ad Duas Lauros	Italy
Stefano	Fedele	Onlus ICT Ad Duas Lauros	Italy
Antonio	Pallotti	Onlus ICT Ad Duas Lauros	Italy
Antonella	Fresa	Promoter Srl	Italy
Claudio	Prandoni	Promoter Srl	Italy
		Italy Count	5
Erwin	Verbruggen	Netherlands Institute for Sound and Vision	Netherlands
		Netherlands Count	1
Kamil	Rutkowski	DI Factory	Poland
		Poland Count	1
Stanislav	Rogozhin	ABBYY	Russian Federation
		Russian Federation Count	1
Lenka Bazalova	Bazalova	The University library in Bratislava	Slowakia
Lucia	Kelemenová	The University Library in Bratislava	Slowakia
Milan	Rakús	The University Library in Bratislava	Slowakia
Juraj	Strnisko	The University Library in Bratislava	Slowakia
		Slowakia Count	4
Sònia	Oliveras i Artau	Ajuntament de Girona	Spain
Miquel	Montaner	Easy Innova	Spain
Víctor	Muñoz	Easy Innova	Spain
Xavi	Tarrés	Easy Innova	Spain
Peplluís	de la Rosa	University of Girona	Spain
		Spain Count	5
Magnus	Geber	Riksarkivet	Sweden
Klas	Jadeglans	Riksarkivet	Sweden
Borje	Justrell	Riksarkivet	Sweden

Benjamin	Yousefi	Riksarkivet	Sweden
Jan	Karlsson	SMHI	Sweden
Lisa	Hammar	Swedish Meteorological and Hydrological Institute (SMHI)	Sweden
		Sweden Count	6
Peter	Fornaro	University of Basel	Switzerland
		Switzerland Count	1
Olena	Chaikovska	Kyiv National University of Culture and Arts	Ukraine
		Ukraine Count	1
Joachim	Jung	Open Preservation Foundation	United Kingdom
Becky	McGuinness	Open Preservation Foundation	United Kingdom
Carl	Wilson	Open Preservation Foundation	United Kingdom
		United Kingdom Count	3
		Grand Count	71

ANNEX II EXPERIENCE WORKSHOP SLIDES

PREservation FORMATs for culture information/e-archives

Introduction to PREFORMA

Borje Justrell,
Swedish National Archives
Coordinator

Project Identity Card

- PREFORMA is a **Pre-Commercial Procurement** project co-funded by the European Commission under FP7-ICT Programme.
- **Start date**: 1 January 2014
- **Duration**: 48 month (end date: 31 December 2017)
- **Total budget for the procurement**: 2.805.000 EUR
- **Website**: www.preforma-project.eu
- **Contacts:**
  - Project Coordinator: Borje Justrell, Riksarkivet, borje.justrell@riksarkivet.se
  - Technical Coordinator: Antonella Fresa, Promoter Srl, fresa@promoter.it
  - Communication Coordinator: Claudio Prandoni, Promoter Srl, prandoni@promoter.it

Project Partners

- RIKSARKIVET, Sweden **Project Coordinator and memory institution**
- PROMOTER SRL, Italy **Technical and Communication Coordinator**

- **Technical partners**
  - PACKED EXPERTISECENTRUM DIGITAAL ERFGOED VZW, Belgium
  - FRAUNHOFER-GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V, Germany
  - HOGSKOLAN I SKOVDE (University of Skovde), Sweden
  - UNIVERSITA DEGLI STUDI DI PADOVA, Italy

- **Memory institutions**
  - STICHTING NEDERLANDS INSTITUUT VOOR BEELD EN GELUID, Netherlands
  - Koninklijk Instituut voor het Kunstpatrimonium, Belgium
  - GREEK FILM CENTRE AE, Greece
  - LOCAL GOVERNMENT MANAGEMENT AGENCY-AN GHNIOMHAIREACHT BAINISTIOCHTA RIALTAIS AITIUIL, Ireland
  - STIFTUNG PREUSSISCHER KULTURBESITZ, Germany
  - AYUNTAMIENTO DE GIRONA, Spain
  - Eesti Vabariigi Kultuuriministeerium, Estonia
  - KUNGLIGA BIBLIOTEKET, Sweden

Project Concept

- Memory institutions are facing **increasing transfers** of electronic documents and other digital media content for long-term preservation.
- Data content are normally stored in specific **file formats** for documents, images, sound, video etc., and these files are usually produced by software from different vendors.
- Even if the transferred files are in standard formats, **the correct implementation of standards** cannot be guaranteed:
  - The software used for the production of the electronic files is not in control neither by the institutions that produces them nor by the memory institutions.
  - Conformance tests of transfers are done by memory institutions, but are not totally reliable; different software for testing could end up in different results.
- This poses problems in **long-term preservation**. Data objects meant for preservation, passing through an uncontrolled generative process, can jeopardise the whole preservation exercise.

Pre-Commercial Procurement (PCP)

- Pre-Commercial Procurement (PCP) is a competition-like procurement method. It enables public sector bodies to engage with innovative businesses and other interested parties in development projects heading at innovative solutions that address specific public sector challenges and needs.
- These innovative solutions are created through a phased procurement of development contracts to reduce risk.
- Pre-Commercial Procurement (PCP) is becoming more and more common within the public sectors of the European Union.

Project Aim and Objectives

- **The aim:** to implement good quality files in various standard formats for preserving content in a long term.
- **The main objective:** to give memory institutions full control of the process of conformity tests of files to be ingested into archives.
- **The main objective of the PCP launched by PREFORMA:** to develop an *open source software* for the management of the whole conformance test process, supporting a range of standards, addressing the needs of any memory institution or other organisation with a preservation task.

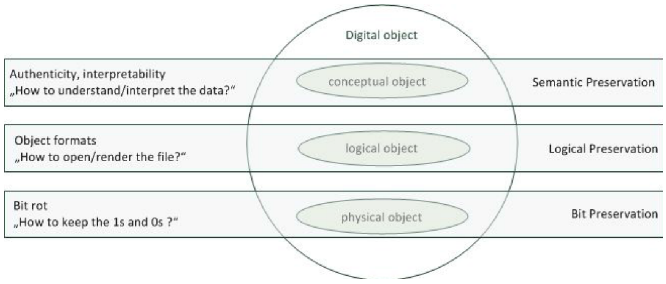
Open Source Approach

- PREFORMA is following an **open source approach**, with the aim of establishing a sustainable research and development community comprising a wide range of contributors and users from different stakeholder groups.
- The open source nature ensures long-term availability of the software, beyond the memory institutions and suppliers involved in PREFORMA.
- Licenses
  - All **software** developed during the PREFORMA project will be provided under two specific open source licenses: "GPLv3 or later" and "MPLv2 or later".
  - All **digital assets** developed during the PREFORMA project will be provided under Creative Commons CC-BY v4.0, and in open file formats.

Overall R&D Objective (The PREFORMA Challenge)

- Develop an **open source conformance checker** that:
  - checks if a file complies with standard specifications
  - checks if a file complies with the acceptance criteria of the memory institution
  - reports back to human and software agents
  - perform simple fixes
- Establish an ecosystem around an **open source reference implementation** that:
  - generates useful feedback for those who control software
  - advances improvement of the standard specification
  - advances development of new business cases for managing preservation files

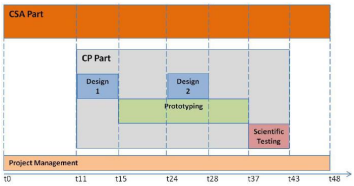
Vision



Target Users and Stakeholders

- **Memory institutions** and cultural heritage organisations, involved in (or planning) digital culture initiatives.
- **Developers** contributing code for the PREFORMA open source tools.
- **Research organisations** providing technical advice to cultural stakeholders.
- **Standardisation bodies** maintaining the technical specifications of the preservation formats covered in PREFORMA.
- **Funding agencies**, such as Ministries of Culture and national/regional administrations, that own and manage digitisation programmes and may endorse the use of the PREFORMA tools in the digitisation process.
- Other **projects** in the digital cultural heritage domain.
- **Any other organisation** planning for long-term preservation of digital content.

Project Implementation Schedule



- **Design phase** (4 months): November 2014 – February 2015
- **Prototyping phase** (22 months): March 2015 – December 2016
  - First prototypes: March 2015 – October 2015
  - Re-design: November 2015 – February 2016
  - Second prototype: March 2016 – December 2016
- **Testing phase** (6 months): January 2017 – June 2017

PREFORMA Suppliers in the Prototyping Phase

1. **veraPDF Consortium** (led by Open Preservation Foundation and PDF Association) – The PDF/A conformance checker accepted industry-wide (PDF/A)
2. **EasyInnova** – Digital Preservation Formats Manager (TIFF)
3. **MediaArea** – PREFORMA MediaConch - CONformance CHecking for audiovisual files (MKV|FFV1|LPCM)

Need for Open Source Communities

Memory institutions have requirements for very long life-cycles of digital assets.

Therefore, it is necessary to establish sustainable communities related to each Open Source software (OSS) solution in which open file formats are implemented.

Events

Today: Experience Workshop in Berlin in connection with demonstrations of the outcomes of the Protyping phase

Forthcoming:

- A **Workshop** that will take place in Padua in March 2017 in connection with prototype demonstrations
- **Training seminars**, planned for Spring 2017
- **Final Conference** that will take place in Stockholm in Autumn 2017 to present the final results of the project.

PREFORMA General Presentation

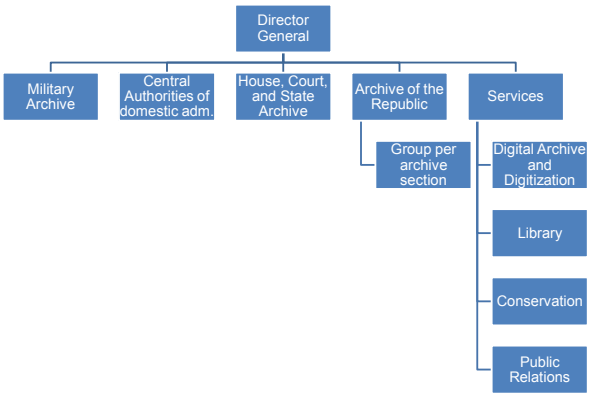
Digital preservation and conformance checking at memory institutions

Hannes Kulovits
Austrian State Archives

Austrian State Archives

- Federal Act on the Safekeeping, Storage and Use of Archival Holdings of the Federal Government (1999)
- Archival holdings of the federal government
- Key role as 'living archive': Federal Government Departments have to offer current records to the archive (13 ministries plus subordinate agencies)
- Transfer to archive after a period of 10 years
- Access dates (30, 50 years)

Austrian State Archives



Digital archival holdings

- „Made-digital“
  - Photographs
  - Paper records
  - Maps
  - Charters
  - Parchments
  - Seals
- „Born-digital“
  - Electronic records (>= 2003)

Electronic Records Management

- Electronic records management system on federal level (all agencies) in 2004
- Legislation adapted accordingly
- Since 2004 all records are born-digital
- Only the electronic record is legally valid
- Central Electronic Records Management System (Fabasoft eGovSuite)
- Operated by Federal Data Center
- 2013: 11.000 users, 18 TB (net), > 1m files/yr

Challenges

- Ministries free from directives
- No uniform records management policy across ministries
- Responsibilities for records management within agencies differ
- No file format restrictions
- Software malfunction
- Certain gap between number of offered records and appraised records

Challenges

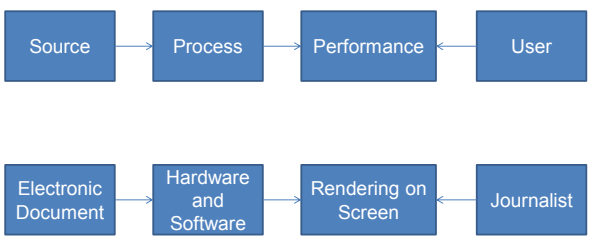
- *"We do not currently have specifications for these older file formats."*
- *"It is likely that those employees who had significant knowledge of these formats are no longer with Microsoft."*

(Tony Hey, Corporate Vice President
Microsoft Research)

Fundamental Requirements

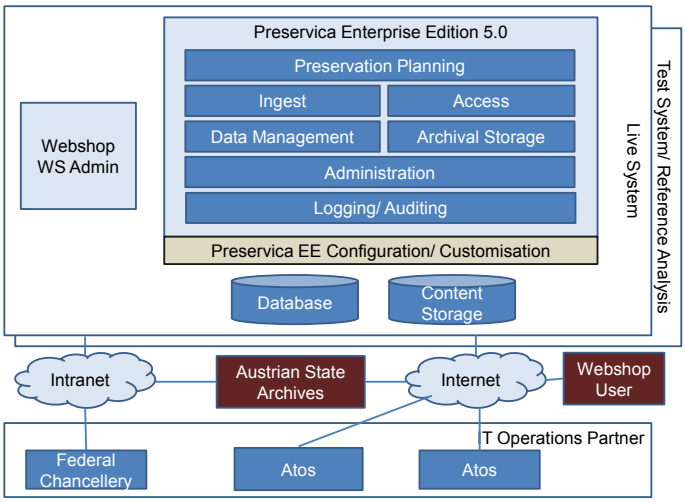
- Archival holdings must be
  - Authentic
  - Reliable
  - Have integrity
  - Useable
  - Interpretable
- Information/Digital Preservation

Performance of a Digital Document

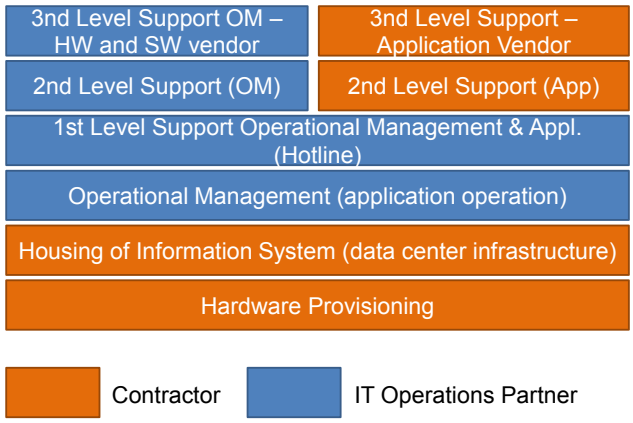


Digital Preservation Goals

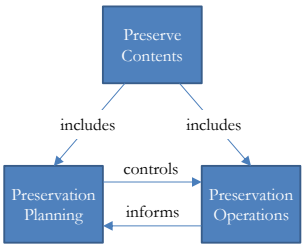
1. **Acquire content** from producers in accordance to the mandate
2. **Deliver** authentic, complete, usable and understandable **objects** to designated user community
3. Faithfully **preserve provenance** of all objects
4. Authentically **preserve objects** for the specified time horizon, securing their integrity and protecting them from threats
5. **React to changes** in the environment timely in order to keep objects accessible and understandable
6. Ensure repository **sustainability**: mandate, technical, financial, operational, communities
7. Build **trust** in the depositors, the designated community and other stakeholders
8. Maximize **efficiency** in all operations



Roles and responsibilities



Preservation operations



- Analysis ("characterisation")
  - Identification
  - Feature extraction
  - Validation
- Actions
  - Migration
  - Emulation
  - Others
- Quality Assurance
  - Authenticity
  - Significant properties
- Metadata
- Reporting

Which two files are similar?

- Consider three files A, B, C

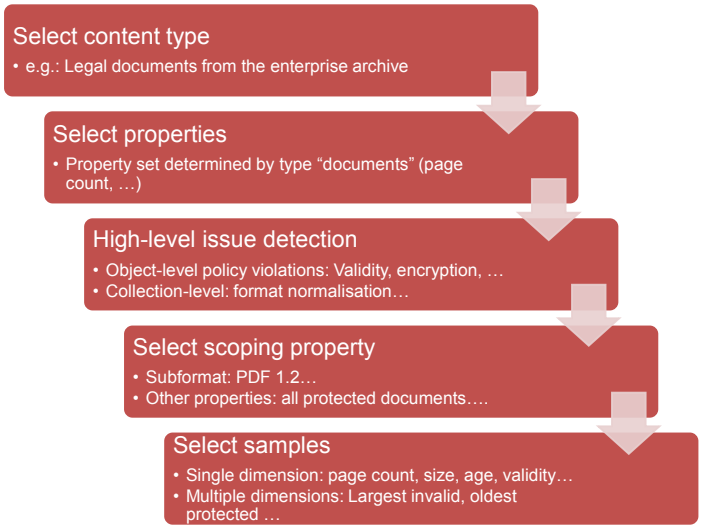
	A	B	C
Format	PDF 1.2	PDF 1.2	PDF 1.4

Which two files are similar?

- Consider three files A, B, C

	A	B	C
Format	PDF 1.2	PDF 1.2	PDF 1.4
Page count	20	1.700	40
Encryption	Yes	No	Yes
File size	1MB	65 MB	2 MB
Valid	no	yes	No
Well-formed	Yes	yes	Yes
Digital signature	no	yes	no

… file format is just another property.



Properties of interest

- Format and Sub-format
- Representation Instance Properties
- Information Properties

Format Risk Factors

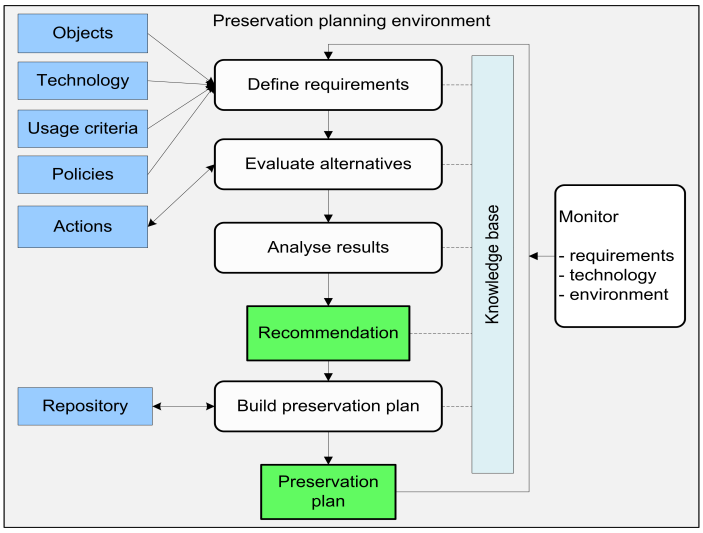
- Number of free and open source tools available
- Availability of documentation
- Quality of documentation
- Standardised
- Identification possibilities
- Validation possibilities
- Intellectual Property Rights
- License costs
- Native browser support

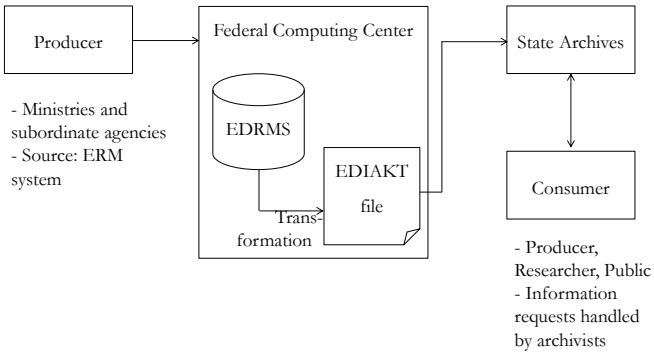
Representation Instance Properties

- Object well-formed
- Object valid
- File size
- Compression
- Document searchable
- Document machine-readable
- Embedded metadata valid
  - E.g. EXIF
- Format conforms

Information Properties

- Document
  - Number of pages
  - Number of characters
  - Footnotes
  - Table of contents
  - Header, Footer
- Image
  - Width, Height
  - Bits per sample
  - Colour space

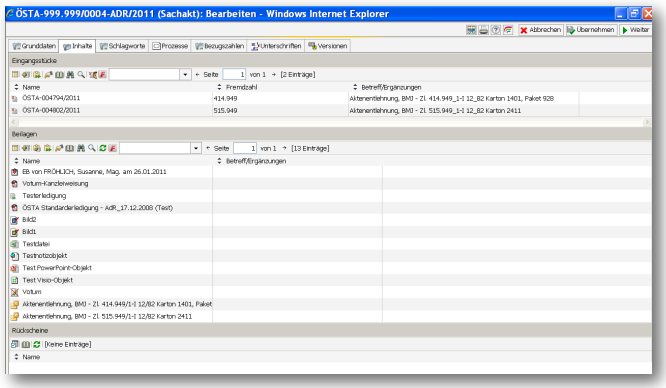




Records have properties

- Unique Identifier
- Sender
- Receiver
- Organisational Unit
- Status (Approved/Closed/Cancelled/…)
- Keywords
- Last Change
- Subject Area
- Cassation Period
- Archive Date
- Archive Annotation
- … and hundreds more

From ERM System to the Archive



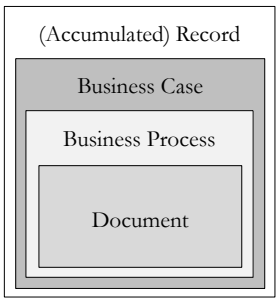
From ERM System to the Archive

Transformation (EDIAKT II)

```
<?xml version="1.0" encoding="UTF-8"?>
<Ediakt
    xmlns="http://reference.e-government.gv.at/namespace/ediakt/20
    xmlns:p="http://reference.e-government.gv.at/namespace/personc
                              xmlns:xsi="http://www.w3.org/2001/XMLS
  <Header>
    <Receiver>
      <Authority>
        <p:CorporateBody>
          <p:FullName>Österreichisches Staatsarchiv</p:FullName>
        </p:CorporateBody>
      </Authority>
    </Receiver>
    <Sender>
      <p:CorporateBody>
        <p:FullName>ELAK im Bund</p:FullName>
      </p:CorporateBody>
    </Sender>
    <Purpose>
      <Type>for archive</Type>
    </Purpose>
  </Header>
  <MetaData>
    <Identifier>
      <Identification>BMUKK-X1.000/0070-IT/1c/2009</Identification
      <ManagementIndikator>BMUKK-BMWF</ManagementIndikator>
      <OrganisationalUnit>BMUKK-BMWF - IT/1c (BMUKK - IT/1c)</Orga
      <SubjectArea>X1.000/Test</SubjectArea>
      <OrdinalNumber>70</OrdinalNumber>
      <Year>2009</Year>
      <TechnicalID>COO.3000.110.7.1918325</TechnicalID>
    </Identifier>
```

[http://reference.e-government.gv.at/Q-EDIAKT_XML-Schema_zu_Ediakt.739.0.html]

EDIAKT II



XSD Validation not enough

- Number of Business Cases equals one?
- Is record closed?
- Does referenced document exist?
- Second manifestation for defined files exists?
- Does workflow exist?
- Does workflow conform to RM standard?
- ...

Use Case

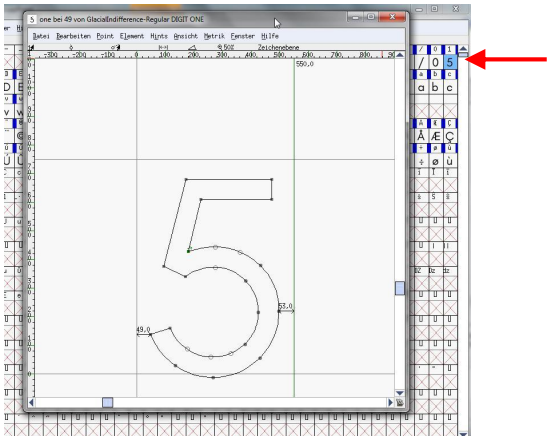
- One PDF that renders differently on different computers
- PDF is digitally signed

- The rendered PDFs are absolutely identical
- Verify using checksums

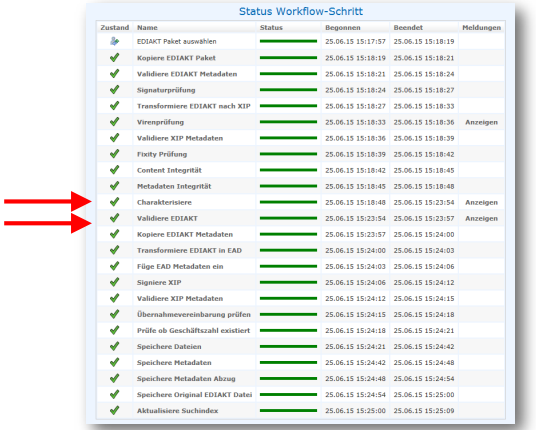
Problem

- Font is not embedded
- During rendering, application loads fonts from operating system
- If the font doesn't exist it is replaced by a similar one (defined in font)
- Structure of the document might change
- Prone to attacks

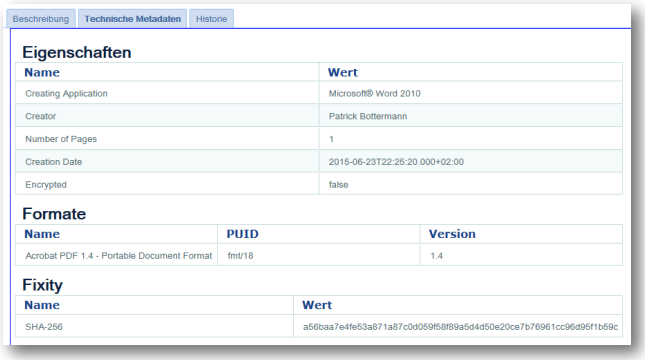
Scenario



Ingest



Technical Metadata



Tools for Characterization and Validation

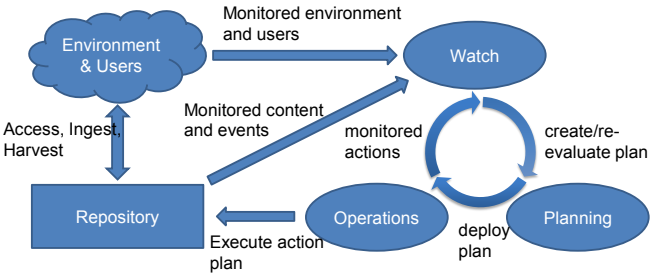
```
JhoveView (Rel. 1.6, 2011-01-04)
 Date: 2015-06-25 15:38:37 MESZ
 RepresentationInformation: Contract-123-pdfa.pdf
  ReportingModule: PDF-hul, Rel. 1.8 (2009-05-22)
  LastModified: 2015-06-23 22:24:56 MESZ
  Size: 120945
  Format: PDF
  Version: 1.4
  Status: Well-Formed and valid
...
 TrueType:
     Font:
      Name: F1
      BaseFont: ABCDEE+Glacial Indifference
...
 FontDescriptor:
       FontName: ABCDEE+GlacialÃIndifference
       Flags: Nonsymbolic
       FontBBox: -32, -250, 937, 750
       FontFile2: true
...
```

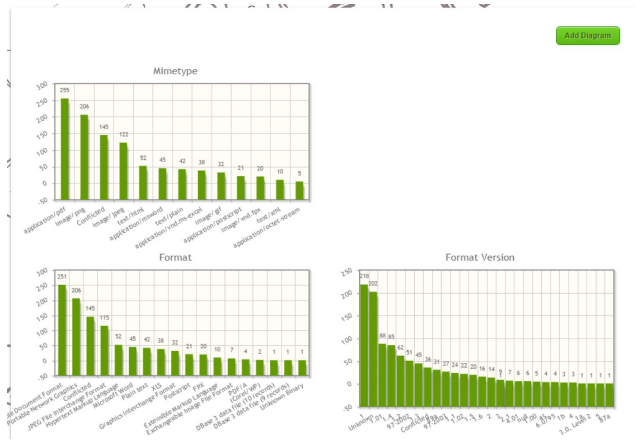
- JHove
- FITS
- (DROID)
- veraPDF
- DPF M.
- ...

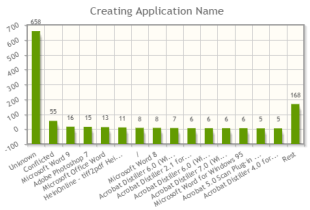
Conclusions

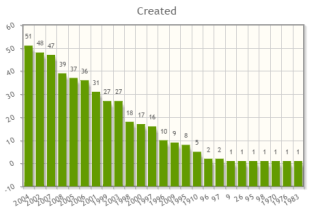
- Central to the success of digital preservation efforts is to
  - know WHAT you have
  - be sure that digital files are what they purport to be
  - know WHAT action to take and HOW
- Well-established capabilities
  - Preservation Planning
  - Preservation Operations

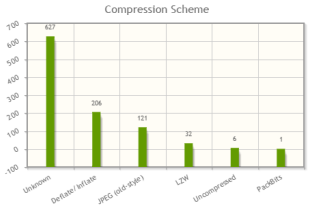
Digital preservation lifecycle

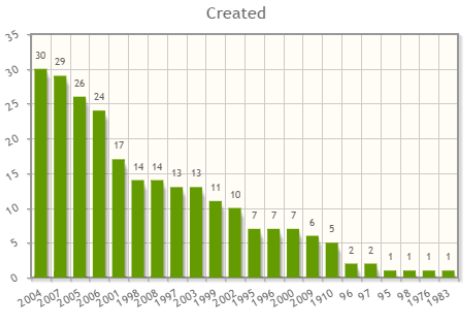


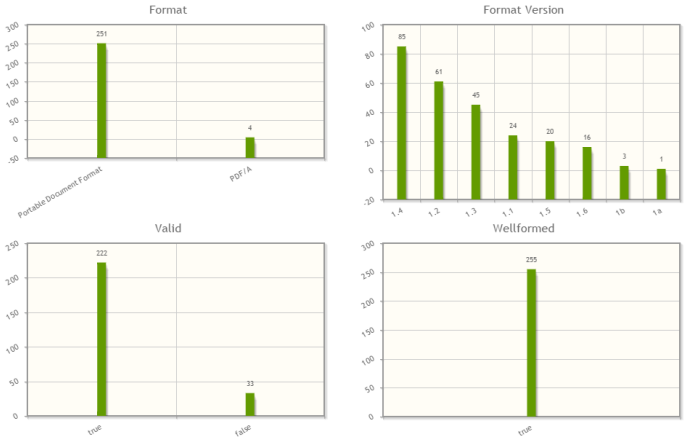












Thank you for your attention!

hannes.kulovits@oesta.gv.at

Preforma Workshop Berlin

About me

- Matthias Priem
- matthias.priem@viaa.be
- t: @matthiaspriem

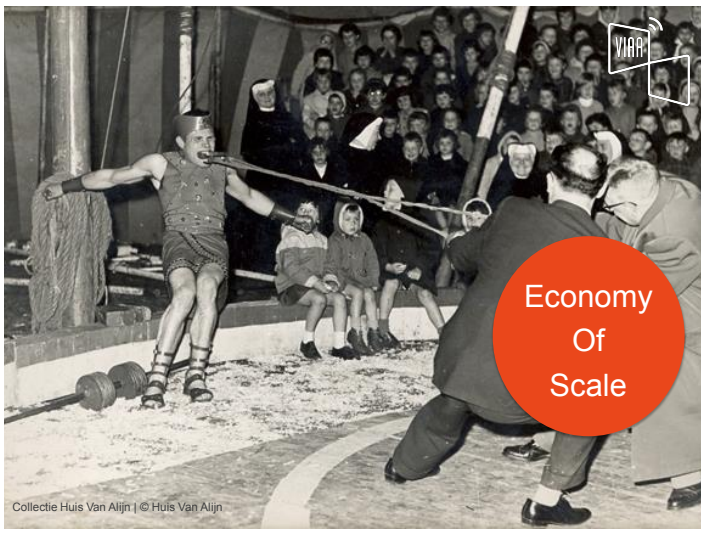
- Was IT Manager at iMinds, Ugent
- Work at VIAA since 3 years
- Project manager archive


21 Dec. '12
FOUNDED


DIGITISATION
ARCHIVING
DISSEMINATION


No collection management
We work for 100+ org's
SERVICE PROVIDER
CITY ARCHIVES
BROADCAST
CULTURAL HERITAGE
vrt
ADVN
ARGOS
KU LEUVEN
S.M.A.K.
TV OOST
Musea Brugge
IN FLANDERS FIELDS MUSEUM
TVL
ROB
MSK GENT
M HKA


Not (only) a Broadcast Archive



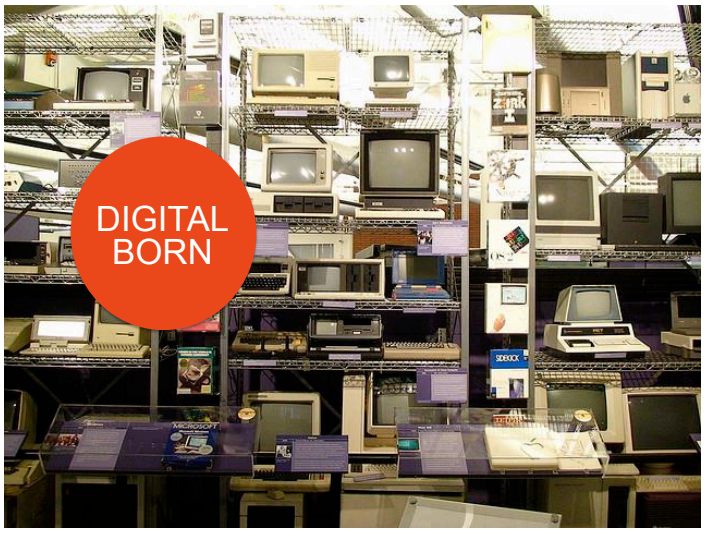
DIGITISATION ARCHIVING DISSEMINATION

DIGITISATION

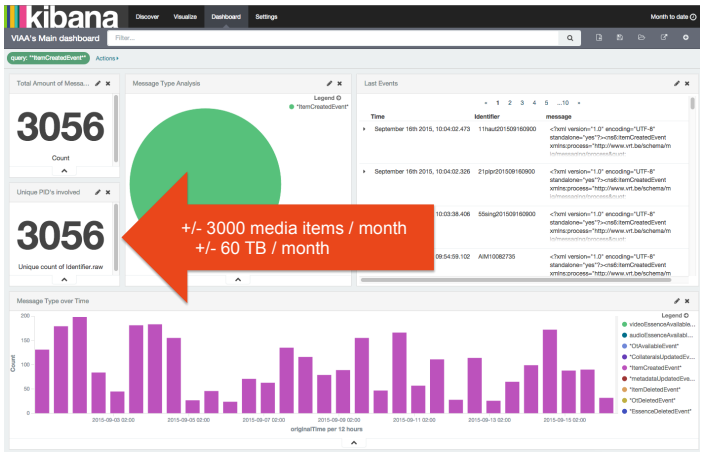
INVENTORY OF ANALOGUE CARRIERS (2014)

	HERITAGE + MEDIA		HERITAGE	MEDIA
Type materiaal/formaat	# DRAGERS TOTAAL	UREN TOTAAL	# dragers aandeel %	# dragers aandeel %
Film	74.173	26.709	11%	89%
Video analog	150.389	228.773	26%	74%
Video digital	225.578	156.189	5%	95%
Audio analogue	165.883	78.370	36%	64%
Audio digital	34.870	18.112	54%	46%
TOTAL	**650.893**	**508.153**	**21%**	**79%**

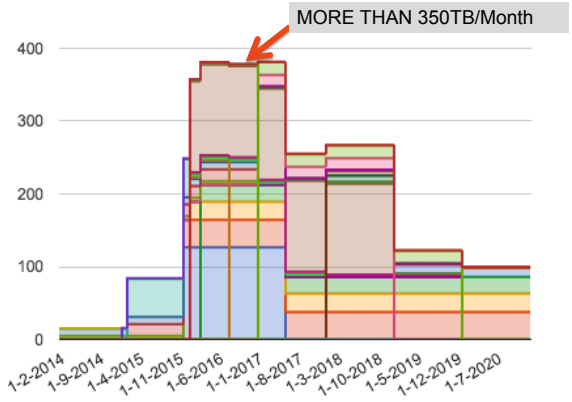
Evergem's Yzerblad: komt uit de loopgrachten als 't past **(S.I. 1917 – 1918)** Collectie Erfgoedbibliotheek Hendrik Conscience | © Vlaamse Erfgoedbibliotheek – Foto: Stefan Tavernier



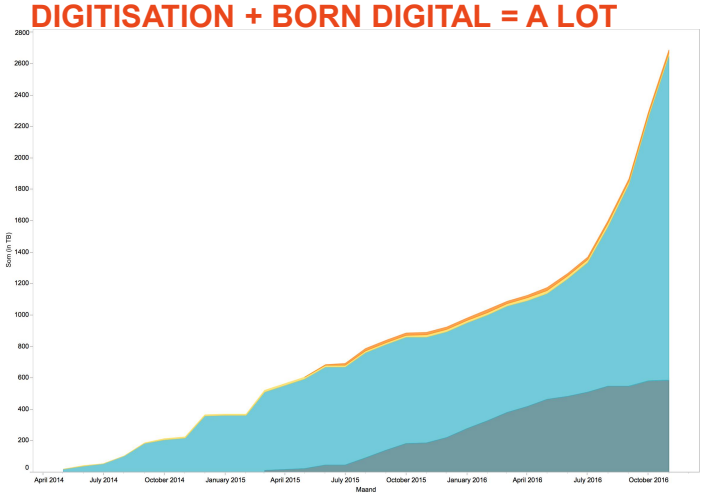
DIGITAL BORN @ VRT

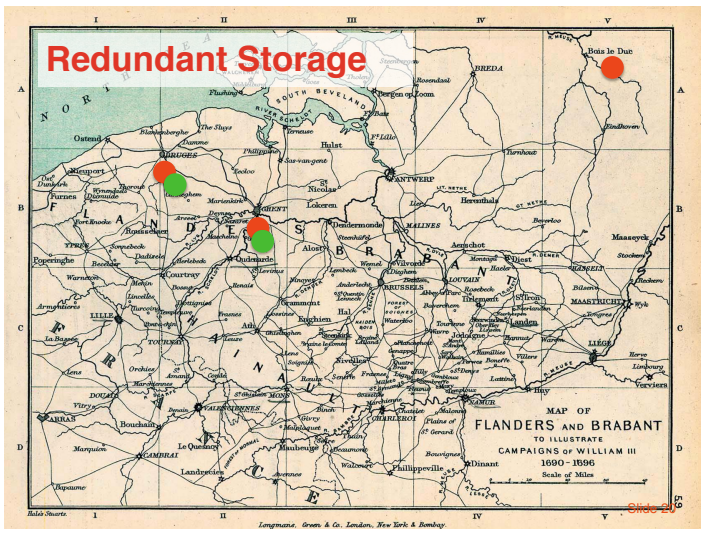


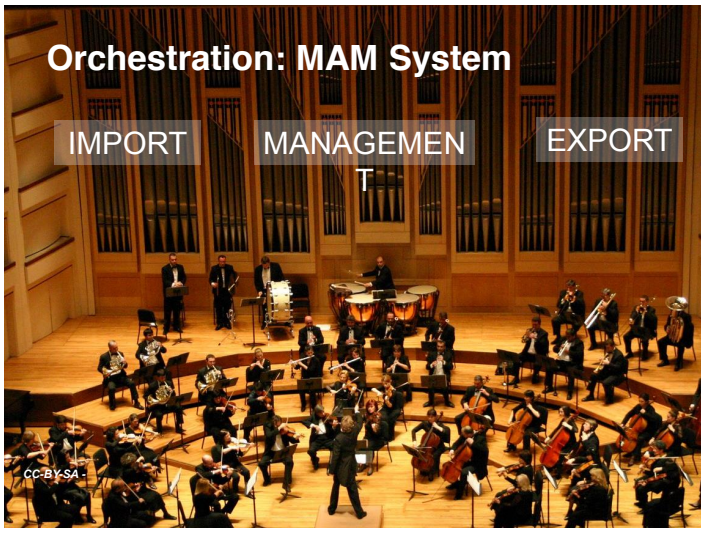
DIGITISATION + BORN DIGITAL = A LOT



DIGITISATION + BORN DIGITAL = A LOT







Orchestration: MAM System

1. IMPORT WORKFLOWS & INGEST REPORTING

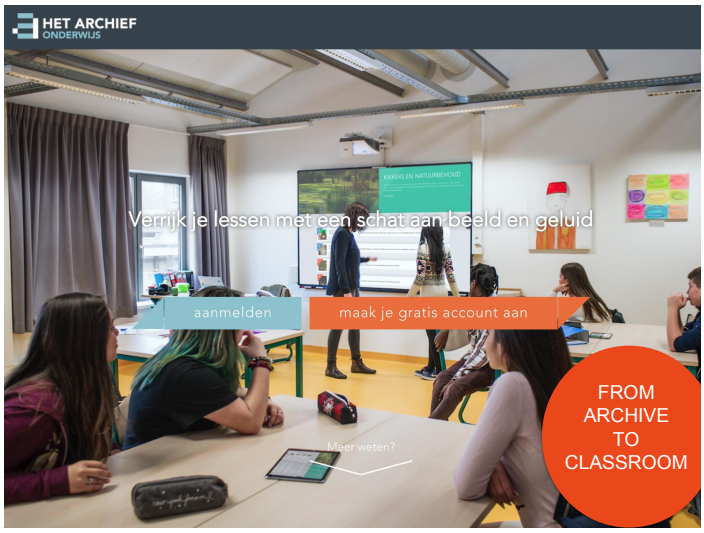
CONFORMANCE CHECKING!

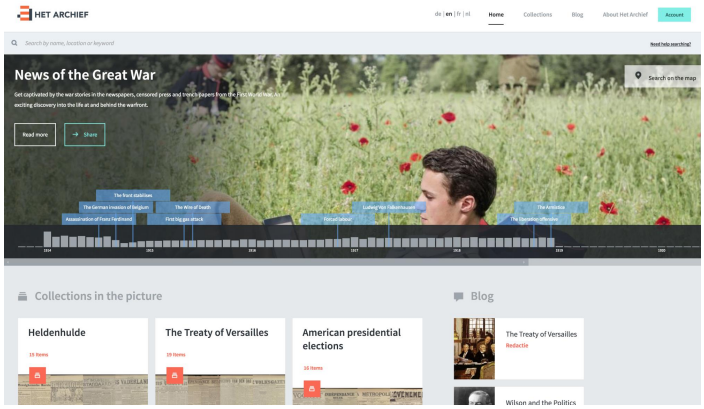
multi-tenant MAM SYSTEM

2. MANAGEMENT
- Transcoding (ffmpeg)
- Partial retieve
- Metadata
- Time-based Annotation
- Search (~SOLR)
- Manual QC
- Storage Management

3. EXPORT WORKFLOW (OAI-PMH) , **EXIT PATH**

DISSEMINATION





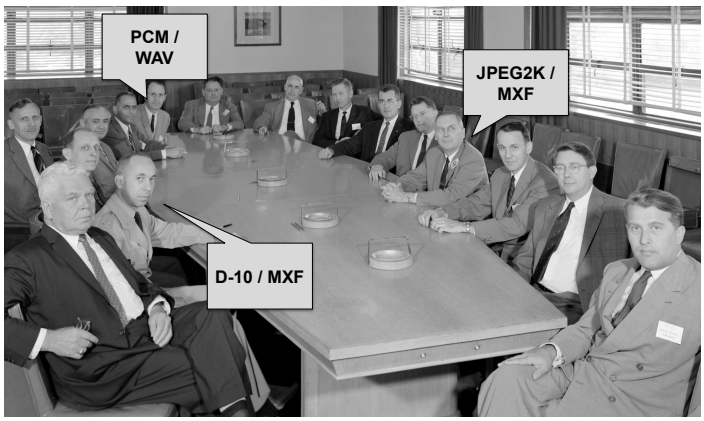
VIAA

File conformance checking @ VIAA

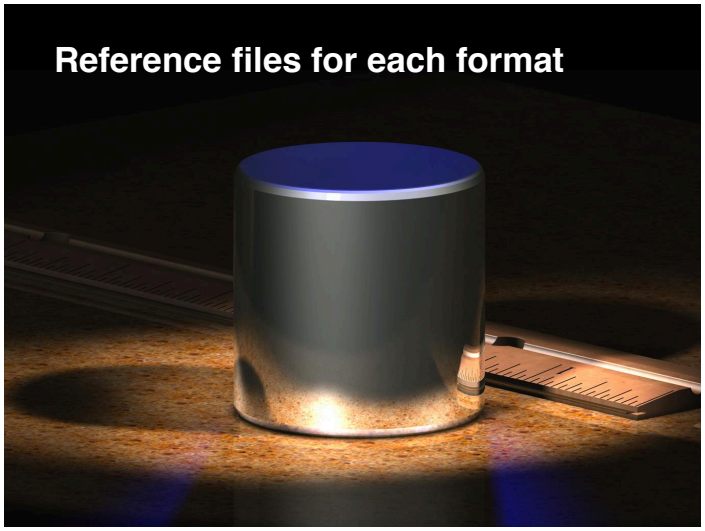
File conformance checking : overview

- Digitization
  - Try to agree on a single(*) digitization format for everything
  - Conformance checking before the project starts
  - When files arrive at the archive (SIP)
- Born digital
  - When files arrive at the archive (SIP)

Agreeing on digitization formats



Reference files for each format

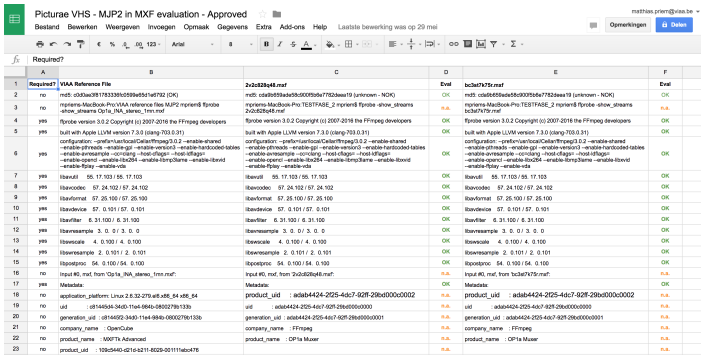


Before we start a project

- Tender demanding the file format
- Test and pilot phase
- Service provider provides test files
- VIAA does the check

Lingotto Test track - CC-BY-2.0 - Jean-Pierre Dalbéra

Before we start a project



Picturae VHS - MJP2 in MXF evaluation - Approved

	A	B	C	D	E	F
1	Required?	VIAA Reference File	2v2c828q48.mxf	Eval	bc3a97b75c.mxf	Eval
2	no	md5: c0d0ae3f81783336fc0599e65d1e6792 (OK)	md5: cda9b659ade58c900f5b6e7782deea19 (unknown - NOK)	OK	md5: cda9b659ade58c900f5b6e7782deea19 (unknown - NOK)	OK
3	no	mpriems-MacBook-Pro:VIAA reference files MJP2 mprieм$ ffprobe -show_streams Op1a_INA_stereo_1mn.mxf	mpriems-MacBook-Pro:TESTFASE_2 mprieм$ ffprobe -show_streams 2v2c828q48.mxf	n.a.	mpriems-MacBook-Pro:TESTFASE_2 mprieм$ ffprobe -show_streams bc3a97b75c.mxf	n.a.
4	yes	ffprobe version 3.0.2 Copyright (c) 2007-2016 the FFmpeg developers	ffprobe version 3.0.2 Copyright (c) 2007-2016 the FFmpeg developers	OK	ffprobe version 3.0.2 Copyright (c) 2007-2016 the FFmpeg developers	OK
5	yes	built with Apple LLVM version 7.3.0 (clang-703.0.31)	built with Apple LLVM version 7.3.0 (clang-703.0.31)	OK	built with Apple LLVM version 7.3.0 (clang-703.0.31)	OK
6	yes	configuration: --prefix=/usr/local/Cellar/ffmpeg/3.0.2 --enable-shared --enable-pthreads --enable-gpl --enable-version3 --enable-hardcoded-tables --enable-avresample --cc=clang --host-cflags= --host-ldflags= --enable-opencl --enable-libx264 --enable-libmp3lame --enable-libxvid --enable-ffplay --enable-vda	configuration: --prefix=/usr/local/Cellar/ffmpeg/3.0.2 --enable-shared --enable-pthreads --enable-gpl --enable-version3 --enable-hardcoded-tables --enable-avresample --cc=clang --host-cflags= --host-ldflags= --enable-opencl --enable-libx264 --enable-libmp3lame --enable-libxvid --enable-ffplay --enable-vda	OK	configuration: --prefix=/usr/local/Cellar/ffmpeg/3.0.2 --enable-shared --enable-pthreads --enable-gpl --enable-version3 --enable-hardcoded-tables --enable-avresample --cc=clang --host-cflags= --host-ldflags= --enable-opencl --enable-libx264 --enable-libmp3lame --enable-libxvid --enable-ffplay --enable-vda	OK
7	yes	libavutil 55. 17.103 / 55. 17.103	libavutil 55. 17.103 / 55. 17.103	OK	libavutil 55. 17.103 / 55. 17.103	OK
8	yes	libavcodec 57. 24.102 / 57. 24.102	libavcodec 57. 24.102 / 57. 24.102	OK	libavcodec 57. 24.102 / 57. 24.102	OK
9	yes	libavformat 57. 25.100 / 57. 25.100	libavformat 57. 25.100 / 57. 25.100	OK	libavformat 57. 25.100 / 57. 25.100	OK
10	yes	libavdevice 57. 0.101 / 57. 0.101	libavdevice 57. 0.101 / 57. 0.101	OK	libavdevice 57. 0.101 / 57. 0.101	OK
11	yes	libavfilter 6. 31.100 / 6. 31.100	libavfilter 6. 31.100 / 6. 31.100	OK	libavfilter 6. 31.100 / 6. 31.100	OK
12	yes	libavresample 3. 0. 0 / 3. 0. 0	libavresample 3. 0. 0 / 3. 0. 0	OK	libavresample 3. 0. 0 / 3. 0. 0	OK
13	yes	libswscale 4. 0.100 / 4. 0.100	libswscale 4. 0.100 / 4. 0.100	OK	libswscale 4. 0.100 / 4. 0.100	OK
14	yes	libswresample 2. 0.101 / 2. 0.101	libswresample 2. 0.101 / 2. 0.101	OK	libswresample 2. 0.101 / 2. 0.101	OK
15	yes	libpostproc 54. 0.100 / 54. 0.100	libpostproc 54. 0.100 / 54. 0.100	OK	libpostproc 54. 0.100 / 54. 0.100	OK
16	no	Input #0, mxf, from 'Op1a_INA_stereo_1mn.mxf':	Input #0, mxf, from '2v2c828q48.mxf':	n.a.	Input #0, mxf, from 'bc3a97b75c.mxf':	n.a.
17	yes	Metadata:	Metadata:	OK	Metadata:	OK
18	no	application_platform: Linux 2.6.32-279.el6.x86_64 x86_64	product_uid : adab4424-2f25-4dc7-92ff-29bd000c0002	n.a.	product_uid : adab4424-2f25-4dc7-92ff-29bd000c0002	n.a.
19	no	uid : c8144fd4-34d0-11e4-984b-080027f9133b	uid : adab4424-2f25-4dc7-92ff-29bd000c0000	n.a.	uid : adab4424-2f25-4dc7-92ff-29bd000c0000	n.a.
20	no	generation_uid : c8146ff2-34d0-11e4-984b-080027f9133b	generation_uid : adab4424-2f25-4dc7-92ff-29bd000c0001	n.a.	generation_uid : adab4424-2f25-4dc7-92ff-29bd000c0001	n.a.
21	no	company_name : OpenCube	company_name : FFmpeg	n.a.	company_name : FFmpeg	n.a.
22	no	product_name : MXF7k Advanced	product_name : OP1a Muxer	n.a.	product_name : OP1a Muxer	n.a.
23	no	product_uid : 109c5440-d21d-b211-8029-001111abc476		n.a.		n.a.

Before we start a project

Required?	VIAA Reference File	2v2c828q48.mxf	Eval
no	md5: c0d0ae3f81783336fc0599e65d1e6792 (OK)	md5: cda9b659ade58c900f5b6e7782deea19 (unknown - NOK)	OK
no	mpriems-MacBook-Pro:VIAA reference files MJP2 mprieм$ ffprobe -show_streams Op1a_INA_stereo_1mn.mxf	mpriems-MacBook-Pro:TESTFASE_2 mprieм$ ffprobe -show_streams 2v2c828q48.mxf	n.a.
yes	ffprobe version 3.0.2 Copyright (c) 2007-2016 the FFmpeg developers	ffprobe version 3.0.2 Copyright (c) 2007-2016 the FFmpeg developers	OK
yes	built with Apple LLVM version 7.3.0 (clang-703.0.31)	built with Apple LLVM version 7.3.0 (clang-703.0.31)	OK
yes	configuration: --prefix=/usr/local/Cellar/ffmpeg/3.0.2 --enable-shared --enable-pthreads --enable-gpl --enable-version3 --enable-hardcoded-tables --enable-avresample --cc=clang --host-cflags= --host-ldflags= --enable-opencl --enable-libx264 --enable-libmp3lame --enable-libxvid --enable-ffplay --enable-vda	configuration: --prefix=/usr/local/Cellar/ffmpeg/3.0.2 --enable-shared --enable-pthreads --enable-gpl --enable-version3 --enable-hardcoded-tables --enable-avresample --cc=clang --host-cflags= --host-ldflags= --enable-opencl --enable-libx264 --enable-libmp3lame --enable-libxvid --enable-ffplay --enable-vda	OK
yes	libavutil 55. 17.103 / 55. 17.103	libavutil 55. 17.103 / 55. 17.103	OK
yes	libavcodec 57. 24.102 / 57. 24.102	libavcodec 57. 24.102 / 57. 24.102	OK
yes	libavformat 57. 25.100 / 57. 25.100	libavformat 57. 25.100 / 57. 25.100	OK
yes	libavdevice 57. 0.101 / 57. 0.101	libavdevice 57. 0.101 / 57. 0.101	OK
yes	libavfilter 6. 31.100 / 6. 31.100	libavfilter 6. 31.100 / 6. 31.100	OK
yes	libavr	libavresample 3. 0. 0 / 3. 0. 0	OK
yes	li	libswscale 4. 0.100 / 4. 0.100	OK
yes		libswresample 2. 0.101 / 2. 0.101	OK
yes		libpostproc 54. 0.100 / 54. 0.100	OK
no	nxf':	Input #0, mxf, from '2v2c828q48.mxf':	n.a.

(kind of) MANUAL MEDIACONCH ...

WHEN SIPS ARRIVE

VIAA

Event ID	PID	Event	Status	Date	Comment	Content Provider
	3t9			from / until		
2546111	3t9d50gs73	PUBLISHED	OK	2015-10-20 17:22:08		Ring TV
2546110	3t9d50gs73	QC_MANUAL	OK	2015-10-20 17:22:07	outcome:OK, , , , user:lissa.janssens@ringtv.be(592af7c6-1375-4ace-8e5e-9cdded3f787e),	Ring TV
575174	3t9d50gs73	TAPE_VAULTED	OK	2015-04-01 18:43:09	V00151L6	Ring TV
505112	3t9d50gs73	ARCHIVED_ON_BACKUP	OK	2015-03-07 08:57:05	tape=G00151L6	Ring TV
505088	3t9d50gs73	ARCHIVED_ON_TAPE	OK	2015-03-07 08:36:33	tape=O00151L6	Ring TV
505058	3t9d50gs73	ARCHIVED_ON_VAULT	OK	2015-03-07 08:15:16	tape=V00151L6	Ring TV
504850	3t9d50gs73	TRANSCODING	OK	2015-03-07 05:16:56	pathToVideo:http://archief-media.viaa.be/viaa/RINGTV/392fa4abb3404ec0a1ff814eb947b345359e013b082a476e97d2f7af24ada176/browse.mp4	Ring TV
500769	3t9d50gs73	MD5_CHECK	OK	2015-03-03 20:25:07		Ring TV
500768	3t9d50gs73	CODEC_CHECK	OK	2015-03-03 20:25:07		Ring TV
500767	3t9d50gs73	SIDECAR_CHECK	OK	2015-03-03 20:25:07		
483063	3t9d50gs73	SIP_VALIDATION	OK	2015-03-02 17:53:56	QC on tape	
483062	3t9d50gs73	SIP_VALIDATION	OK	2015-03-02 17:53:56	Sidecar on tape	
483061	3t9d50gs73	SIP_DETECTED	OK	2015-03-02 17:53:56		

INGEST PROCES

WHEN SIPS ARRIVE

VIAA

Event ID	PID	Event	Status	Date	Comment	Content Provider
	3t9			from / until		
2546111	3t9d50gs73	PUBLISHED	OK	2015-10-20 17:22:08		Ring TV
2546110	3t9d50gs73	QC_MANUAL	OK	2015-10-20 17:22:07	outcome:OK, , , , user:lissa.janssens@ringtv.be(592af7c6-1375-4ace-8e5e-9cdded3f787e),	Ring TV
575174	3t9d50gs73	TAPE_VAULTED	OK	2015-04-01 18:43:09	V00151L6	Ring TV
505112	3t9d50gs73	ARCHIVED_ON_BACKUP	OK	2015-03-07 08:57:05	tape=G00151L6	Ring TV
505088	3t9d50gs73	ARCHIVED_ON_TAPE	OK	2015-03-07 08:36:33	tape=O00151L6	Ring TV
505058	3t9d50gs73	ARCHIVED_ON_VAULT	OK	2015-03-07 08:15:16	tape=V00151L6	Ring TV
504850	3t9d50gs73	TRANSCODING	OK	2015-03-07 05:16:56	pathToVideo:http://archief-media.viaa.be/viaa/RINGTV/392fa4abb3404ec0a1ff814eb947b345359e013b082a476e97d2f7af24ada176/browse.mp4	Ring TV
500769	3t9d50gs73	MD5_CHECK	OK			Ring TV
500768	3t9d50gs73	CODEC_CHECK	OK			Ring TV
500767	3t9d50gs73	SIDECAR_CHECK	OK			
483063	3t9d50gs73	SIP_VALIDATION	OK	2015-03-02 17:53:56	QC on tape	
483062	3t9d50gs73	SIP_VALIDATION	OK	2015-03-02 17:53:56	Sidecar on tape	
483061	3t9d50gs73	SIP_DETECTED	OK	2015-03-02 17:53:56		

1. Basic check against the reference file

INGEST PROCES

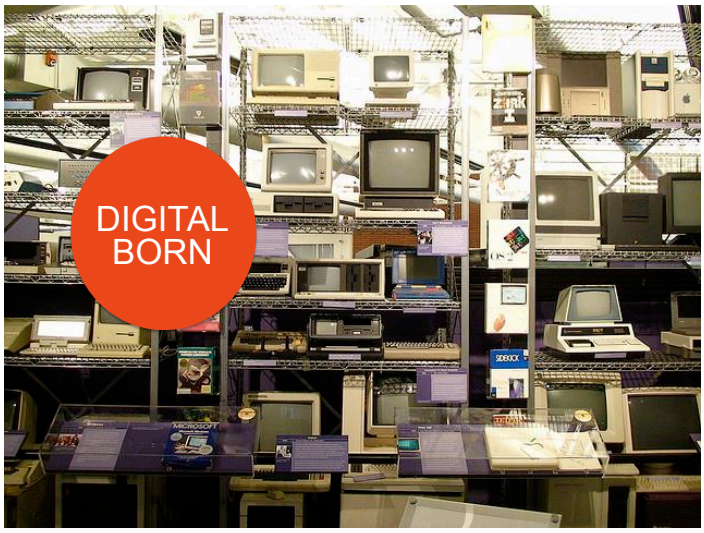
WHEN SIPS ARRIVE

VIAA

Event ID	PID	Event	Status	Date	Comment	Content Provider
	3t9			from / until		
2546111	3t9d50gs73	PUBLISHED	OK	2015-10-20 17:22:08		Ring TV
2546110	3t9d50gs73	QC_MANUAL	OK	2015-10-20 17:22:07	outcome:OK, , , , user:lissa.janssens@ringtv.be(592af7c6-1375-4ace-8e5e-9cdded3f787e),	Ring TV
575174	3t9d50gs73	TAPE_VAULTED	OK	2015-04-01 18:43:09	V00151L6	Ring TV
505112	3t9d50gs73	ARCHIVED_ON_BACKUP	OK	2015-03-07 08:57:05	tape=G00151L6	Ring TV
505088	3t9d50gs73	ARCHIVED_ON_TAPE	OK	2015-03-07 08:36:33	tape=O00151L6	Ring TV
505058	3t9d50gs73	ARCHIVED_ON_VAULT	OK			
504850	3t9d50gs73	TRANSCODING	OK			
500769	3t9d50gs73	MD5_CHECK	OK	2015-03-03 20:25:07		Ring TV
500768	3t9d50gs73	CODEC_CHECK	OK	2015-03-03 20:25:07		
500767	3t9d50gs73	SIDECAR_CHECK	OK	2015-03-03 20:25:07		
483063	3t9d50gs73	SIP_VALIDATION	OK	2015-03-02 17:53:56	QC on tape	
483062	3t9d50gs73	SIP_VALIDATION	OK	2015-03-02 17:53:56	Sidecar on tape	
483061	3t9d50gs73	SIP_DETECTED	OK	2015-03-02 17:53:56		

2. This step also includes an internal MXF check

INGEST PROCES



DIGITAL BORN CONTENT

- MUCH more diverse
- Difficult to enforce a certain format
  - No other file available
  - Many producers (e.g. broadcast)
  - Fast technological advances
- Only checks are done during ingest
  - Upon SIP arrival
- Record technical information and store for future use

Current status

- We do check the conformance at several instances.
- But:
  - Missing in-depth analysis (we check metadata)
  - Some tools are closed inside the MAM
    no real control over what happens / the output
- New formats and digital born content are a challenge

Our interest in PREFORMA

- Conformance checking / file validation
  - Mediaconch is potentially interesting for the manual step when we start
  - In depth codec check is missing
  - May use this inside the ingest workflow
- Work around standardisation of FFv1
- Results in open source = ☺

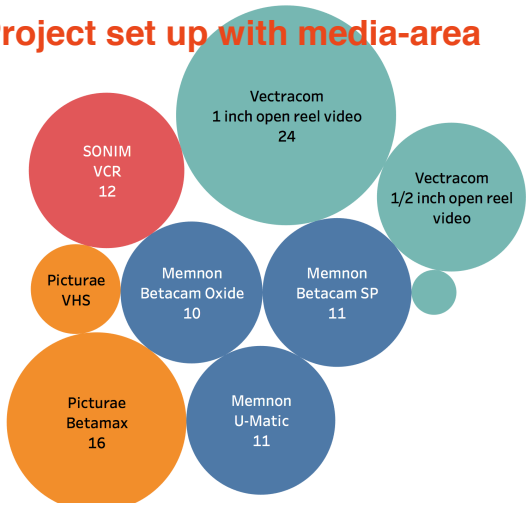
Project set up with media-area

- Hands on experience with FFv1
  - How easily can we create an FFv1 file?
  - Can we convert jp2k to FFv1 without data-loss?
  - What about
    - Storage needs of the hi-res
    - Processing speed (when creating an mp4 for instance)
  - Based on VIAA material
- Hands-on experience with mediaconch
  - How can we get this into our flows?

Project set up with media-area

- Testing material:
  - 100 files in jpeg2000 / mxf format (about 3TB)
  - Cross-section of everything we have
    - Different digitization firms
    - Different sources (analogue sources)
    - Different owners (VIAA content partners)
- 1 machine with 18 cores to process the data.
- FFMPEG

Project set up with media-area



Project set up with media-area

- Tests (~benchmarking):
  - 1. jpeg2000 to mp4
  - 2. jpeg2000 to ffv1 (with framemd5 check)
  - 3. Ffv1 validation in mediaconch
  - 4. FFv1 to mp4
- Results
  - Insight in resulting files
  - Insight in processing capacity
  - Feasiblity of further implementation in VIAA archival flows

Some preliminary results

- Decoding to mp4 *using ffmpeg* : seems like 3 or 4 times faster when starting from FFv1.
- Jpeg2000 vs FFv1 : 10% less space needed
- Lossless transcoding appears possible
- todo :investigate if and how to embed this in our workflows

=> results will be put out in the open in the next couple of weeks.

PREFORMA

Conformance checking in a high volume production line

Klas Jadeglans
Riksarkivet, MKC
IT Architect

PREFORMA

Table of content

- About MKC
- Production flow system
- Implementation of DPF Manager
- Experiences
- Wishes

Riksarkivet
Experience Workshop
Berlin, 2016-11-23

About MKC

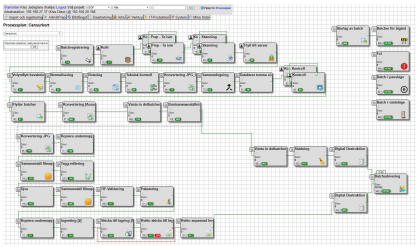
A Digitization factory for the cultural heritage

About MKC

- A part of the Swedish National Archives
- A national government resource for digitizing
- About 65 employees
- Located in the center of Sweden
- Focus on high volume production of digital images
- ~100.000 images per day
- ...loose sheet, bound books, newspapers, maps etc
- Have produced more than 200 Million digital images since we switched from microfilm in 1995

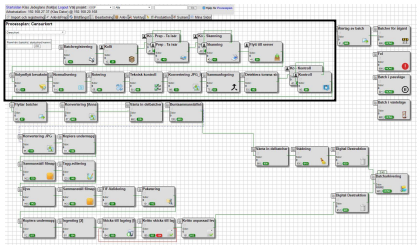
- We do not archive any images ourselves...
- ...we deliver all images to others that archive

Production Flow System



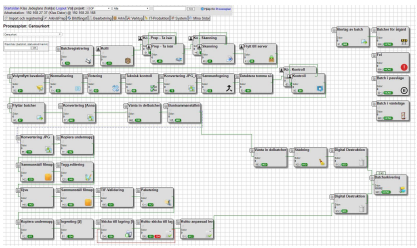
- Web browser GUI for human operators
- Autonomous "workers" for automatic operations
- 17 virtual linux servers, 1 Windows app
- Designed by me, built by me and my collegues
- Integrates to some 3rd-party components

Production Flow System

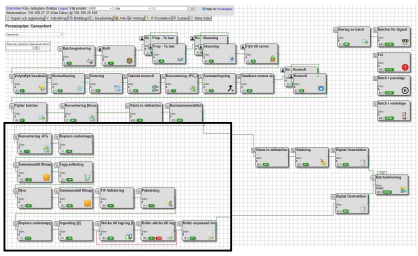


- Production:
  - Registration, preparation, scanning, quality control etc

Production Flow System

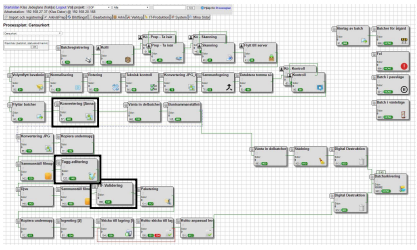


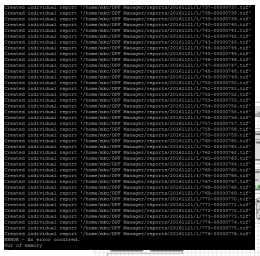
Production Flow System



- **Post production:**
  - **Image conversion, OCR, packaging, delivery etc**

Implementation of DPF Manager





Performance:

- Only greyscale images in test:
  - one tif: 16 seconds
  - batch(776 tifs): 355 seconds
  - 0.45 seconds/image
- Really good
- Unfortunally ”out of memory”
- …summary was 321.000 rows
- …since ALL 776 images had errors

PREFORMA

Wishes

- **Make install easier for Non-graphical systems**
- **"Totally non-human" mode**
  - **Ultra silent**
  - **command line return code**
- **Option to select output:**
  - **summary**
  - **individual files**
  - **only final conclusion (number of pass/fail)**
- **Unlimited batch sizes**
- **Knowledge base**
- **Make it "for dummies"**

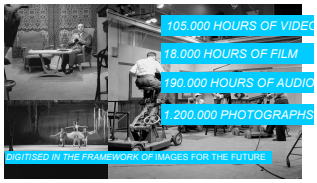
- **Finally:**
  - **TIA**

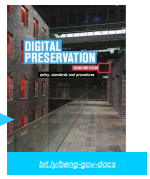
PREFORMA

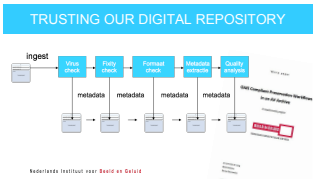
Thank you!

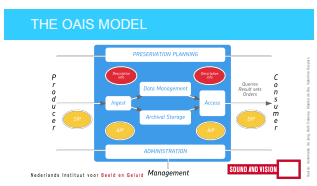
Klas Jadeglans
Riksarkivet, MKC
klas.jadeglans@riksarkivet.se

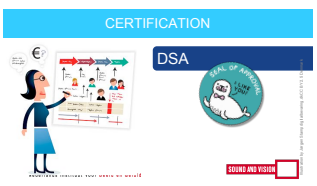
Riksarkivet Experience Workshop Berlin, 2016-11-23











KIK-IRPA experiences
(Berlin workshop 23/11/2016)

- Situation
  - BALaT = Belgian Art Links and Tools
    - http://balat.kikirpa.be
  - Archive: 1.000.000 photo (negatives)
  - Printed photographs digitised in the 90s
    - Low to mid-resolution
  - Since end 90s: digital born
  - Since 2003: high resolution photography of old negatives

KIK-IRPA experiences
(Berlin workshop 23/11/2016)

- Problem
  - Wide range of "different" TIFF's
  - Lack of internal procedure
    - Photographers decided themselves for many years
  - Evolution of software (version, programs, …)
  - Where, when and how to tackle?

PREFORMA General Presentation

KIK-IRPA experiences
(Berlin workshop 23/11/2016)

- Preforma
  - Adapt workflow to integrate checker
  - Situation now:
    - Photographer/operator follows basic procedure
      - TIFF v6 baseline / not compressed / Adobe RGB / 16 bits / 350 ppi
    - Final file dropped in inbox on server
      - Derivatives are automatically created (thumbnails, downloadable mid-resolution JPEGs, etc…)

PREFORMA General Presentation

KIK-IRPA experiences
(Berlin workshop 23/11/2016)

- Preforma
  - Situation future:
    - Idem
      - but conformance checker at "inbox on server"-level => IT intervention/installation
    - Photographers/operators can also use local checker
    - Bulk control existing TIFFs
      - IT intervention
    - …

PREFORMA General Presentation

benjamin.yousefi@riksarkivet.se

Legal and technical adviser

(the Swedish National Archives)

Riksarkivet

Evaluation of suppliers and their software, **open source**, development tools, platform, forums, community, documentation, tutorials, deployment, standards and standardization, patent and license, industry support, marketing, *governance*, sustainability, commercial feasibility, quality of software, test files, interoperability, modularity, interface, portability, scalability...

Riksarkivet

Riksarkivet

How do we create a PDF/A in accordance with the ISO 19005?

How do we confirm that a PDF/A is in accordance with the ISO 19005?

Riksarkivet

Problem	Consequence	Example
no authoritative source	Användningen av uttryck och kon cept skiljer sig källa från källa, vilket föranleder osäkerhet kring gällande uttryck och koncept.	Några exempel från textkodning: • *character set* kan åsyfta *repertoire*, *code* eller *encoding*; • *code position*, kan ha följande synonymer: *code number, code value, code element, code point, code set value* eller bara *code*.
there can be **several terms or sources for one and the same standatd, specifikation or recommendation**	En osäkerhet kring informa tionsinnehållet samt relationen mellan källorna och deras dignitet.	• JPEG standarden återfinns som ISO/IEC och ITU-T Recommendation, och JFIF som Ecma International TR. • Windows- eller CP- 1252 associeras felaktigt till ISO 8859-1 (Latin-1) vilket motsvaras egentligen av Windows- eller CP- 28591, men vars felaktiga association kan bli standardiserad genom HTML 5.
there could be many **different versions of a standard, specification or recommendation**	Risk att att felaktiga källor, föråldrade eller icke-normerande källor an vänds.	PKCS #7 Cryptographic Message Syntax Standard (CMS) återfinns som IETF informativ RFC 2315 (PKCS #7), standardiserad i RFC 2630 (CMS), er satt av RFC 3369 (CMS) och 3370 (CMS algoritmer). RFC 3369, ersatt av 3852 (CMS), och uppdaterad genom RFC 4853 (multipla signatä rer) och 5083 ("Authenticated-Enveloped-Data" innehållstyp), men ersatt av RFC 5652 (CMS). RFC 3370 (CMS algoritmer) uppdaterad genom 5754 (använda SHA2 i CMS).
access to sources could be limited by fees, secrecy (trade secrets), or other obstacles.	Ett förlitande på sekundära och tertiära källor, och egna semantiska konstruk tioner utifrån analys av sekundära och tertiära källor, eller observationer eller experiment.	• Flertalet ISO -standarder är avgiftsbelagda. • Dokumentation av proprietära format, protokoll, och API, kan vara helt eller delvis opublicera de, eller om publicerade, bristfälliga.
different interpretation of sources	Samma koncept får olika uttryck, vilket påverkar implementeringen av källan till programkod.	RFC använder "bör inte" vilket tolkas av vissa som "ska inte" och av andra som "kan".
different implementation of sources	Kod och program kan ge uttryck för vari erande tillvägagångssätt, funktioner och beteende, ibland med samma benämning, vilket kan avvika från den ursprungliga definitionen.	• Renderingen av HTML-dokument varierar beroende på vilken webbläsaren, och vilken version av webbläsaren, som används. • PDF/A-1a framställt i LibreOffice 4.0 skiljer sig från PDF/A-1a framställt i LibreOffice 5.0.
meaning changes with time	Tidigare uppfattningar får en annan me ning med tiden, såsom tolkningar, definitioner, principer.	• Språk som kompileras och tolkas; AOT, och JIT; byte code, machine code (bättre exempel?).

variation of code

causes

variation in production, creation, reproducing and recreating

causes

variation of data and information

Riksarkivet

The PREFORMA Challenge to establish an object point of reference

Riksarkivet

Analog preservation?

PREFORMA

SEVENTH FRAMEWORK PROGRAMME

Digital preservation?

Digital preservation definition

Taking **precautions** enabling **long-term access** to digital data.

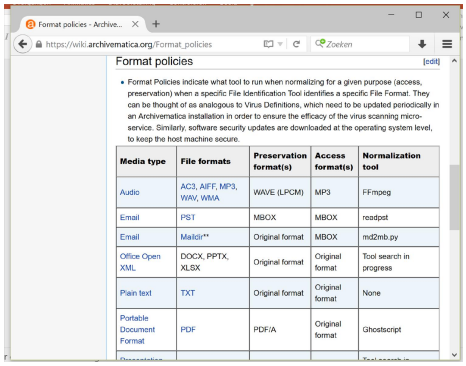
This implies both **policy decisions**, implementing a sustainability strategy, and **practical solutions**, deploying tools to preserve and manage of digital data.

#1 Do Nothing

#2 Conservation

#3 Documentation

#4 Migration?



Format policies

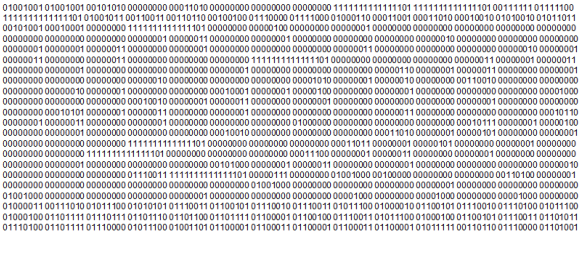
- Format Policies indicate what tool to run when normalizing for a given purpose (access, preservation) when a specific File Identification Tool identifies a specific File Format. They can be thought of as analogous to Virus Definitions, which need to be updated periodically in an Archivematica installation in order to ensure the efficacy of the virus scanning micro-service. Similarly, software security updates are downloaded at the operating system level, to keep the host machine secure.

Media type	File formats	Preservation format(s)	Access format(s)	Normalization tool
Audio	AC3, AIFF, MP3, WAV, WMA	WAVE (LPCM)	MP3	FFmpeg
Email	PST	MBOX	MBOX	readpst
Email	Maildir**	Original format	MBOX	md2mb.py
Office Open XML	DOCX, PPTX, XLSX	Original format	Original format	Tool search in progress
Plain text	TXT	Original format	Original format	None
Portable Document Format	PDF	PDF/A	Original format	Ghostscript

Migrate Where To?

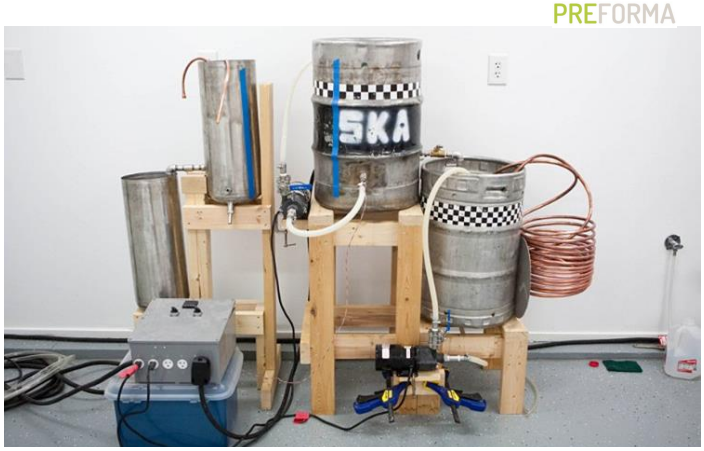
AS-07 | AVC | AVI | DCP | Dirac | DPX | | FLAC | FFV1 | IMX | JPEG | JPEG2000 | LPCM | MKV | MOV | MPEG2 | MPEG4 | MPEG-AF | MXF | OGG | PDF | PDF 1.4 | PDF/A1 | PDF/A2 | PDF/A3 | PNG | RAW | Theora | TIFF 6.0 | VP8 | VP9 | WebM | XDCAM HD422

How? Risks?



#1 Piggybacking

#2 Homebrewing



#3 Evangelising

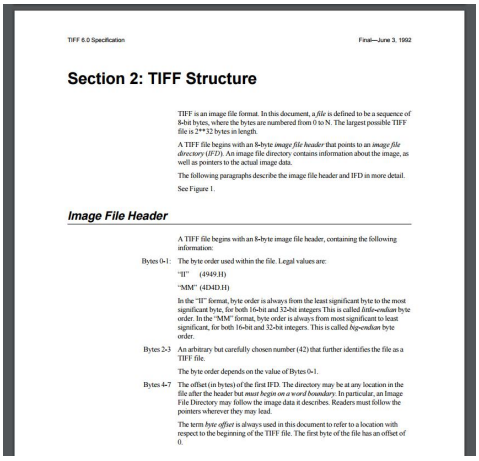
Memory institutions lack:

- knowledge how files technically work
- control over the way files are produced
- tools to manage the abundance and wild fauna and flora of files

Challenge Brief

*Empower memory institutions to **gain** full control over the **technical properties** of **digital content** intended for long-term preservation.*

Specifications



TIFF 6.0 Specification — Final—June 3, 1992

Section 2: TIFF Structure

TIFF is an image file format. In this document, a *file* is defined to be a sequence of 8-bit bytes, where the bytes are numbered from 0 to N. The largest possible TIFF file is 2**32 bytes in length.

A TIFF file begins with an 8-byte *image file header* that points to an *image file directory* (*IFD*). An image file directory contains information about the image, as well as pointers to the actual image data.

The following paragraphs describe the image file header and IFD in more detail.

See Figure 1.

Image File Header

A TIFF file begins with an 8-byte image file header, containing the following information:

Bytes 0-1: The byte order used within the file. Legal values are:

"II" (4949.H)

"MM" (4D4D.H)

In the "II" format, byte order is always from the least significant byte to the most significant byte, for both 16-bit and 32-bit integers This is called *little-endian* byte order. In the "MM" format, byte order is always from most significant to least significant, for both 16-bit and 32-bit integers. This is called *big-endian* byte order.

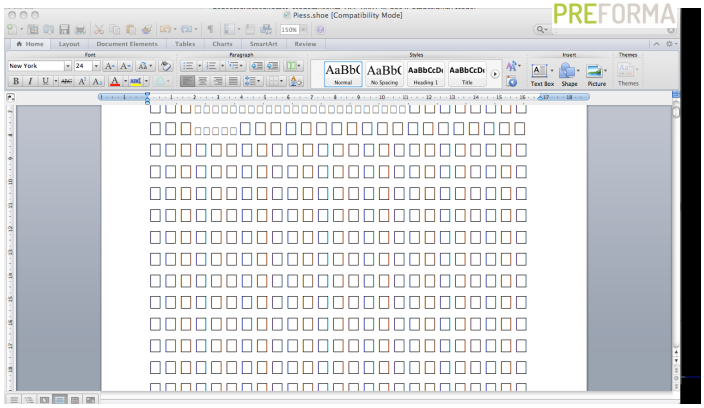
Bytes 2-3 An arbitrary but carefully chosen number (42) that further identifies the file as a TIFF file.

The byte order depends on the value of Bytes 0-1.

Bytes 4-7 The offset (in bytes) of the first IFD. The directory may be at any location in the file after the header but *must begin on a word boundary*. In particular, an Image File Directory may follow the image data it describes. Readers must follow the pointers wherever they may lead.

The term *byte offset* is always used in this document to refer to a location with respect to the beginning of the TIFF file. The first byte of the file has an offset of 0.

Ambiguities > errors



Why don't developers read the specification properly?

- Lazy programming > 'when-it-opens-it's-valid' fallacy.
- Specification are incomplete
- Developers don't have access to the specification (closed)
- Planned obsolescence > make clients dependent (software lock-in)

Why do you need files with consistent properties?

- ensure the **authenticity** of the content
- simplify **management** of digital collections
- enable large scale **migration** and **emulation**

Conformance Checking

The process of checking if the **technical properties** of a digital file are **conform** with the **specification** of the corresponding file format

#1 Develop a conformance checker

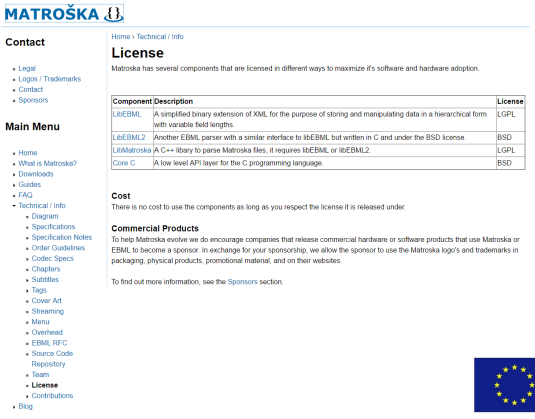
- **Implementation** checker > compliant with specification?
- **Policy** checker > compliant with acceptance criteria
- **Reporter** > readable for human and software agents
- **Fixer** > solves simple errors/ambiguities

Which Formats? Specifications?

MXF | MPEG | IMX | XDCAM HD422 | DPX | DCP | JPEG2000 | MOV | MPEG2 | AVI | MPEG4 | AVC | PDF 1.4 | PDF/A1 | TIFF 6.0 | JPEG | RAW | AS-07 | MPEG|AF | PDF | MKV | FFV1 | OGG | Dirac | PNG | WebM | VP8 | OGG | Theora | PDF/A2 | PDF/A3 | LPCM

#1. Use complete specifications

#2. Use open specifications

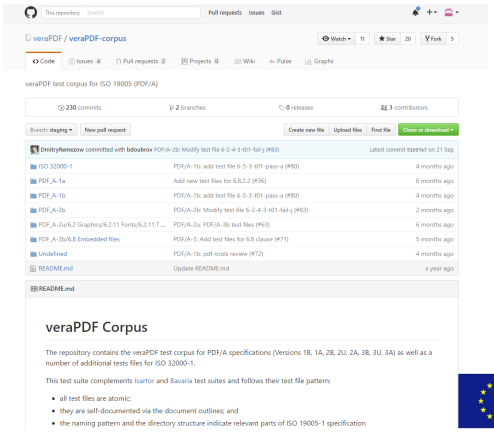


#2. Free/Libre?

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

European Interoperability Framework for Pan-European eGovernment Service (version 1.0 2004)

#3. Use reference implementations



What we have chosen (eventually…)

- TEXT (strengthen the consensus)
  - ISO 32000-1:2008 (PDF 1.7)
  - ISO 19005-1:2005 (PDF/A-1)
  - ISO 19005-2:2011 (PDF/A-2)
  - ISO 19005-3:2012 (PDF/A-3)
- IMAGE (improve the consensus)
  - ISO 12234-2:2001 (TIFF/EP)
  - ISO 12369:2004 (TIFF/IT)
- MOVING IMAGE (virgin path…)
  - OGG / MKV
  - FFV1 / Dirac / ISO 15444-1 (JPEG2000 core coding system)
  - LPCM

#2. Establish an ecosystem around an reference implementation

- Improve the **specification** > standardization initiatives
- Address those who **control** software > community building
- Advance **open source business models** > new services to support producers and archivists

Open Source Approach

<opensource.preforma-project.eu>

Aim:

- Establish a **sustainable research and development community**
- Ensure **long-term availability of the software** beyond memory institutions and suppliers involved.

Licenses:

- All **software** developed during the PREFORMA project will be provided under two specific open source licenses: "**GPLv3 or later**" and "**MPLv2 or later**".
- All **digital assets** developed during the PREFORMA project will be provided under **Creative Commons CC-BY v4.0**, and in **open file formats**.

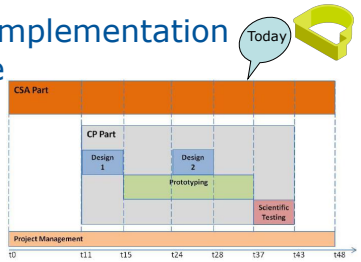
3 Projects

Testing the PREFORMA prototypes

Magnus Geber
Riksarkivet
(National Archives of Sweden)
magnus.geber@riksarkivet.se

Riksarkivet

Project Implementation Schedule



- **Design phase** (4 months): November 2014 – February 2015
- **Prototyping phase** (22 months): March 2015 – December 2016
  - First prototypes: March 2015 – October 2015
  - Re-design: November 2015 – February 2016
  - Second prototype: March 2016 – December 2016
- **Evaluation phase** (6 months): January 2017 – June 2017

Test initiatied by

- Suppliers
  - During development in prototype phase
- Project partners
  - During development i prototype phase
  - Formal evaluation in WP7 after prototype phase

Test preformed by

- Suppliers
- Project partners
- External contributors
  - Both concerning test initiated by Suppliers and Project

Media Partner
DIGITALICULTURE
www.digitalmeetsculture.net
PREFORMA workshop
Berlin, 23 November 2016

Continuous releases

- Frequent releases
- Stable monthly releases
- Formal releases, intermediate and final x 2
  - Formal report documents

Test feedback from project to suppliers

- Ongoing via Github
- Upcoming during prototype phase meetings
- After each formal release
  - Evaluation committee, groups for each format
  - Structured analyze
  - Written feedback

Media Partner
DIGITALICULTURE
www.digitalmeetsculture.net
PREFORMA workshop
Berlin, 23 November 2016

Test beds

- Test integration with legacy and business systems
- Test in production environment
- Example of system types
  - Electronic records management
  - Scanning production
  - E-archive systems
- Cases
  - MKC
  - E-SPACE
  - PACKED

Media Partner
DIGITALICULTURE
www.digitalmeetsculture.net
PREFORMA workshop
Berlin, 23 November 2016

Test files

- Types
  Synthetic files - Organic files
  - Training/Development files
  - Examining files
  - Evaluation, for WP7 Evaluation
  - Demonstration files, open for public
- Source of file
  - Created
  - Gathered by Suppliers
  - Gathered by Project
  - Provide by project partners
  - Provide by external sources
- Organization
  - Metadata form
  - Vault, cloud storage
  - Managed by internal dispatchers
  - Files saved for WP7

Media Partner
DIGITALICULTURE
www.digitalmeetsculture.net
PREFORMA workshop
Berlin, 23 November 2016

Scientific evaluation WP7

- Jan-June 2017
- Specific method, University of Padua
- Establishment of classes
- Establishment of ground truth
  http://www.dpfmanager.org/#download
- Three taskforces
  - Leaders from University of Padua
  - Domain experts
  - External experts
- Preparation started during autumn 2016

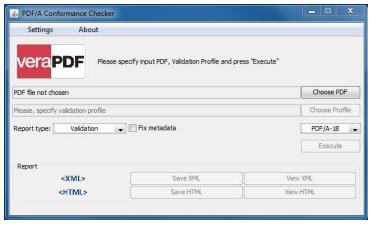
Still possible with testing of external contributors

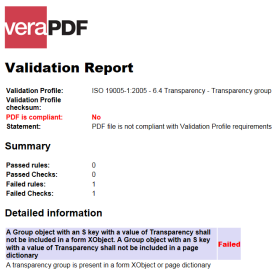
Preforma Open Source Portal

Media Partner
DIGITALICULTURE
www.digitalmeetsculture.net
PREFORMA workshop
Berlin, 23 November 2016

VeraPDF PDF/A

Software: http://verapdf.org/software/

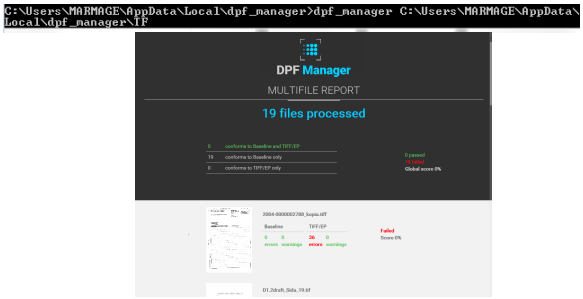




Media Partner
DIGITALICULTURE
www.digitalmeetsculture.net
PREFORMA workshop
Berlin, 23 November 2016

DBF Manager (EasyInnova) Tiff

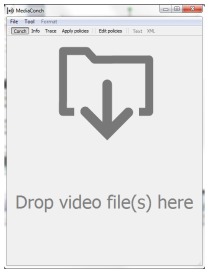
Software: http://www.dpfmanager.org/#download

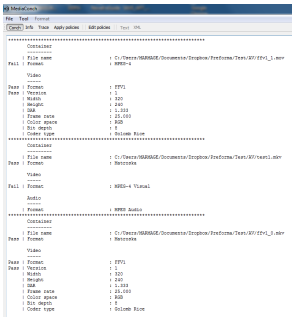


Media Partner
DIGITALICULTURE
www.digitalmeetsculture.net
PREFORMA workshop
Berlin, 23 November 2016

MediaConch (MediaArea) AV

Mjukvara: https://mediaarea.net/MediaConch/download.html





Test feedback from extermal contributors

- Github
- Contact relevant Supplier
- Contact Project, info@preforma-project.eu
- Contact Project partner of choice

Media Partner
DIGITALICULTURE
www.digitalmeetsculture.net

PREFORMA workshop
Berlin, 23 November 2016

veraPDF: industry supported, open source PDF/A validation for digital preservationists

PREFORMA Experience Workshop, Berlin
23 November

Why veraPDF?

- PDF/A, and the standards on which PDF/A is based, are complex; agreement on meaning isn't always clear
- Need for a **trusted** open source tool
- A single **commercial** entity cannot define conformance with PDF/A
- Industry assistance guarantees the means of **interoperability** (it's PDF's holy grail, after all!)

veraPDF consortium

- **Digital preservationists**
  - **Lead**: Open Preservation Foundation
  - Digital Preservation Coalition
  - KEEP Solutions
- **PDF technology industry**
  - **Lead**: PDF Association
  - Dual Lab (veraPDF lead developer)

Functionality & Quality

- **Functional requirements**
  - PDF/A conformance checker
  - PDF features extraction (characterization)
  - Reliable batch processing
  - Clear reporting
  - Ensure accurate use of PDF/A metadata
- **Quality**
  - Test corpora
  - Open source community
  - Proven quality control practices

Other components

- **Policy Checker**
  - Extract PDF Features with various data from the Document (font names, page count and boundaries, security info, images, etc)
  - Verify the Report against Policy requirements in Schematron syntax
- **Metadata Fixer**
  - Fix any incorrect claims of specific PDF/A validity
  - Add PDF/A identification to an otherwise conforming PDF/A document

veraPDF test corpus stats

- Part 1 Level B: 179 test files complementing Isartor
- Part 2 Level B: 223 test files complementing BFO
- Part 3 Level B: 12 extra tests on embedded files
- Level U: 3 test files on Unicode character map
- Level A: 7 test files on tagging, predefined roles
- XMP 2004: 367 tests on predefined schemas
- XMP 2005: 549 tests on predefined schemas

Industry support & transparency

- **Industry support**
  - PDF Association's Validation Technical Working Group
  - Review test documents
  - Discuss and resolve ambiguities in the specifications
- **Transparency of development practices**
  - Open grammar and validation profiles
  - Dual license scheme: MPLv2+ and GPLv3+
  - Availability of functional and technical specifications
  - All materials available at GitHub open repositoty: github.org/verapdf

Today: beta version 0.26

- Validation of all PDF/A parts and conformance levels
- Java API, REST API, CLI, GUI integration interfaces
- Cross-platform GUI installer
- Validation rules wiki, demo web site
- Test corpus with over 1000+ newly generated files
- Available at: http://downloads.verapdf.org
- All sources at: https://github.com/veraPDF
- **The 1.0 release is planned for mid-December**

Helping to understand PDF/A

- Open format of validation profiles:
  - All 8 validation profiles for 1b,1a,2b,2u,2a,3b,3u,3a
  - Each consists of ~100 atomic rules
- The source code for the PDF model as well as all validation profiles is openly available at github:
  - https://github.com/veraPDF/veraPDF-model
  - https://github.com/veraPDF/veraPDF-validation-profiles
- Documentation: Wiki
  - https://github.com/veraPDF/veraPDF-validation-profiles/wiki

Resolution of ambiguities

- **Identification**:
  - 27 cases formally reported to the mailing list
  - discussed at regular TWG calls
  - formally resolved at joint TWG / ISO committee meetings
  - When applicable to future parts of PDF/A, recommendations are forwarded to the PDF/A Project Leader
- **Distribution**:
  - Resolutions posted on the veraPDF mailing list
- **Documentation**:
  - Included into validation Wiki at https://github.com/veraPDF/veraPDF-validation-profiles/wiki

PDF parser implementation

- veraPDF's proof of concept was implemented using PDFBox, an open-source Java library under Apache
- PREFORMA's licensing requirements mandate dual licensing (GPLv3 and MPLv2); no suitable codebase with such licensing was available
- Accordingly, the veraPDF consortium was obliged to develop a greenfield implementation of a PDF parser
- **The first beta of the new greenfield PDF parser is included into the latest 0.26 release**

Interoperability: PREFORMA suppliers

- Two other suppliers:
  - Easy Innova - TIFF validation
  - MediaArea - video streams validation
- Uniform shell:
  - detects file type and forwards it to the corresponding validation tool.
- Embedded video files:
  - veraPDF provides a sample plug-in for validating embedded video files (AVI, MKV) via MediaArea validator

Extensibility via plug-ins

- PDF/A specifications refer to relevant PDF specifications, which rely on a number of external standards. Validating embedded fonts, ICC profiles, XMP metadata, image compression and more is crucial to establishing archival quality for the whole document
- veraPDF provides a plug-in mechanism to access the embedded ICC profiles, fonts, images, attachments
- Collaboration with experts in relevant technologies is necessary for complete coverage

Today: validation extents update

- **JPEG2000**: plug-in based on Jpylyzer python script
- **Other images**: none, looking for contributors
- **Fonts**: partnering with Compart AG (Germany)
- **XMP**: validated internally by veraPDF
- **ICC profiles**: plug-in based on the official ICC validator
- **Digital signatures**: none, looking for contributors
- **Video attachments:** plug-in based on MediaArea

Development processes

- **Github** is the repository for both code and test corpora
  https://github.com/verapdf
- **Travis** manages the continuous builds
  https://travis-ci.org/veraPDF
- **Jenkins** manages automated tests and deployment
  http://jenkins.opf-labs.org/job/veraPDF-library/
- **Sonar** monitors code quality
  http://sonar.opf-labs.org/dashboard/index/8021

Get involved!

- Join the Open Preservation Foundation http://openpreservation.org/join/
- Download the candidate test suite files from GitHub
- Share your own test files
- Test the code posted on GitHub, and report issues
- Develop plug-ins for validating embedded data or PDF features not related to PDF/A
- Think about how you could use industry supported open source PDF/A validation

Stay in touch

- http://verapdf.org/
- http://verapdf.org/subscribe/ (news)
- users@lists.verapdf.org (Q&A, discussions)
- https://twitter.com/_verapdf
- https://github.com/veraPDF
- info@verapdf.org

Thank you

Questions?

DPF Manager

The open source COMMUNITY

EXPERIENCE WORKSHOP | PREFORMA
Berlin, November 23th, 2016
#PreformaBerlin2016

Dr. Miquel Montaner
CTO at Easy Innova

Dr. Peter Fornaro
Managing Director at University of Basel

Dr. Víctor Muñoz
R&D Manager at Easy Innova

Xavi Tarrés
Project Manager at Easy Innova

Prof. Dr. Josep Lluís de la Rosa
Full Professor at University of Girona

www.easyinnova.com

PREFORMA

Are you sure…?

Initial Presentation

INDEX

1. Consortium & People Involved
2. Scope of the Project
3. What is DPF Manager?
4. TIFF Specificacions

TI/A initiative
DPF Manager
Community

Dr. MIQUEL MONTANER
CTO at Easy Innova
miquel@easyinnova.com

EASY INNOVA
www.easyinnova.com

PREFORMA

Consortium & People Involved

Easy Innova, S.L. (Spain)
Spin-off of the University of Girona

Certified by: TECNIO

Member of: AENOR

Know-how:
R&D Projects · Platform Architectures · Digital Preservation · IPR · Standards · Artificial Intelligence · Open Source Projects

University of Girona (Spain)
Agents Research Lab

Know-how:
R&D · Open Source · Digital Preservation

University of Basel (Switzerland)
Digital Humanities Lab

Know-how:
Cultural Heritage · Medical Image · TIFF · Image Formats · Digital Preservation

Dr. Miquel Montaner
CTO

Robert Sallo
R&D Manager

Xavi Tarrés
Project Manager

Antonio López
Senior Developer

Dr. Víctor Muñoz
R&D Manager

Prof. Dr. Josep Lluís de la Rosa
Full Professor and Researcher

Dr. Albert Trias
R&D Manager

Prof. Dr. Lukas Rosenthaler
Full Professor and Researcher

Dr. Peter Fornaro
Managing Director

Scope of the Project

DPF Manager
Open source software
Main features

DPF Manager

Standardization process
TI/A Standard Initiative
Research

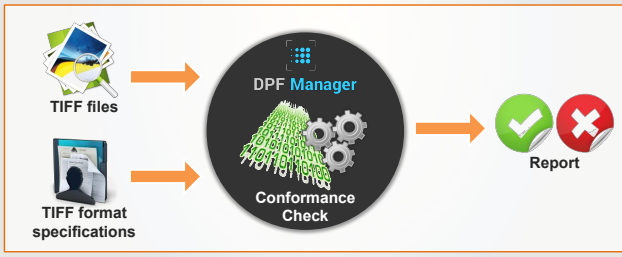
TI A

Memory Institutions
TIFF Experts
Developers

Universitat de Girona

What is DPF Manager?

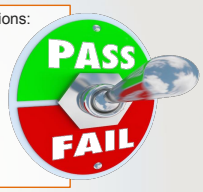
DPF Manager is the most advanced TIFF conformance checker for digital preservation



TIFF Specifications

DPF Manager is able to validate all these TIFF specifications:

- TIFF Baseline, Revision 6.0 Final
- Extended TIFF 6.0
- TIFF/IT *(ANSI IT8.8–1993)*
  - TIFF/IT-P1 *(ISO 12639:1998)*
  - TIFF/IT-P2 *(ISO 12639:2004)*
- TIFF/EP *(ISO 12234-2)*
- Customized specification



DANGER AHEAD

None of these specifications is prepared for digital preservation!!

TI A

TIFF For Archival Recommendations
The TI/A Standard Initiative

NEW

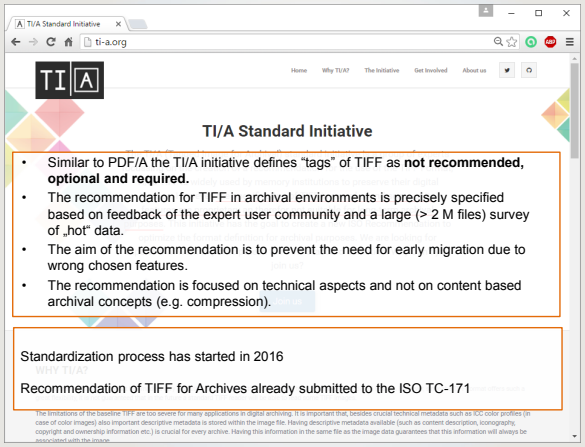
INDEX

Initial Presentation
TI/A Initiative
1. TI/A Standard initiative
2. Timeline
DPF Manager
Community

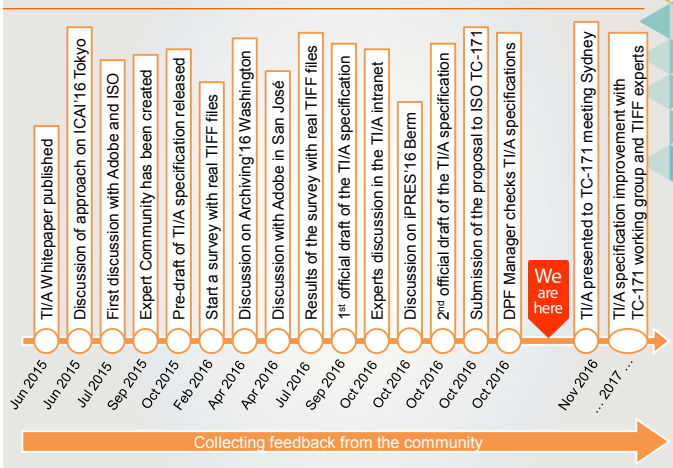
Dr. PETER FORNARO
Managing Director at University Basel
peter.fornaro@unibas.ch

UNIVERSITY OF BASEL
www.unibas.ch

TI/A Standard Initiative



TI/A Initiative Timeline



INDEX

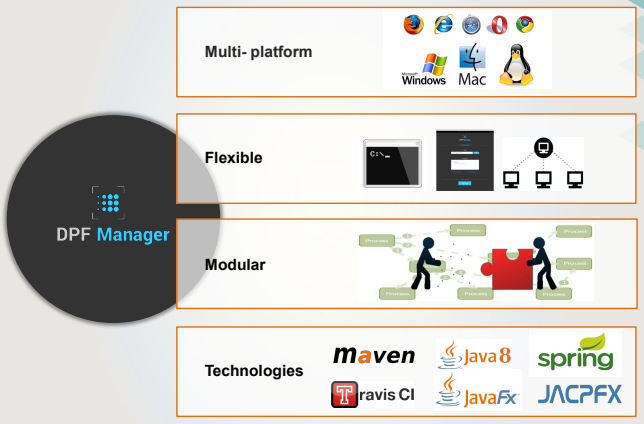
- Initial Presentation
- T/A Initiative
- DPF Manager
  1. DPF Manager Functionalities
  2. DPF Manager Technical Features
  3. Use Scenarios
  4. OAIS Integration
  5. DPF Manager current version
  6. What's Next?
  7. Open Source Project
  8. DPF Manager Website
- Community

Dr. VÍCTOR MUÑOZ
R&D Manager at Easy Innova
victormunoz@easyinnova.com

XAVIER TARRÉS
Project Manager at Easy Innova
xavitarres@easyinnova.com

EASY INNOVA
www.easyinnova.com

DPF Manager Technical Features



Use Scenarios

Standalone

1

Graphic Interface

```
C:\>dpfmanager.exe -
files='archive*.tif' -
config=config.xml
-reports_folder='tmp/reports -
fixed_files_folder='tmp/files
```

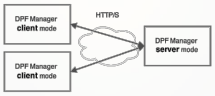
Command line

Client - Server

2

```
client:\>dpfmanager.exe
-server=80.45.32.45
-port=80
```

Client




```
server:\>dpfmanager.exe
-mode=server
```

Server

Use Scenarios

Web application

3

Browser Client


```
server:\>dpfmanager.exe
-mode=server
```

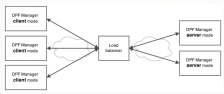
Server

High availability

4

```
client:\>dpfmanager.exe
-server=80.45.32.45
-port=80
```

Client




```
server:\>dpfmanager.exe
-mode=server
```

Server

OAIS integration

Integration with OAIS

Document lifecycle
Conformance checking at:

- Creation
- Transfer
- Migration
- Digitalization

Reference framework
QA, AIP generation

OAIS integration

METS REPORT

The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation.

- Mets Header
- File Section
- Structural Map
- Administrative Metadata
- Descriptive Metadata

Open Source Project

DPF Manager

Open source licenses

MPL v2+

GPL v3+

Open source project

http://dpfmanager.org/

https://github.com/EasyinnovaSL/DPFManager

http://preforma-project.eu/dpf-manager.html

DPF Manager current version

VERSION 3.0 – Released on October 31th, 2016

Interfaces
- Command line interface
- Graphical user interface
- Client/Server
- Online validator

Implementation checker
- Baseline version 6.0 (Baseline 6.0 and Extended 6.0)
- TIFF/EP
- TIFF/IT (Profile 1 and 2)
- TI/A conformance checker

Policy checker
- Create Policy rules from GUI (with critical, error, warning and info levels)
- Custom standards implementations

Metadata fixer
- Add, edit or remove metadata
- Metadata fixer (Make baseline compliant fix, Autofix metadata incoherencies)

Report
- Human readable (HTML, PDF)
- Machine readable (XML, JSON)
- Achievable data formats (METS, PREMIS, NISO, Dublin Cores)

DPF Manager current version

VERSION 3.0 – Released on October 31th, 2016

File inspection analysis
- File structure
- Tags list
- IPTC fields
- EXIF tags
- ICC Profile
- XMP data (including history)
- File identification
- Metadata incoherencies
- Expert mode option

Other features
- Periodical checks
- Wizard to create and save configuration files
- Multi-threading
- Load balancing
- Console and tasks widget
- Logging system
- Show ISO references
- Conformance checkers interoperability
- Maven artifact for easy integration in other projects
- Multi language

What's Next?

In the next VERSIONS…

https://github.com/EasyinnovaSL/DPFManager/milestones

Implementation checker
- DNG
- TIFF-F (RFC 2306)
- TIFF-FX (RFC 3949)

Policy checker
- New policies to apply

Metadata fixer
- Detect and report file provenance

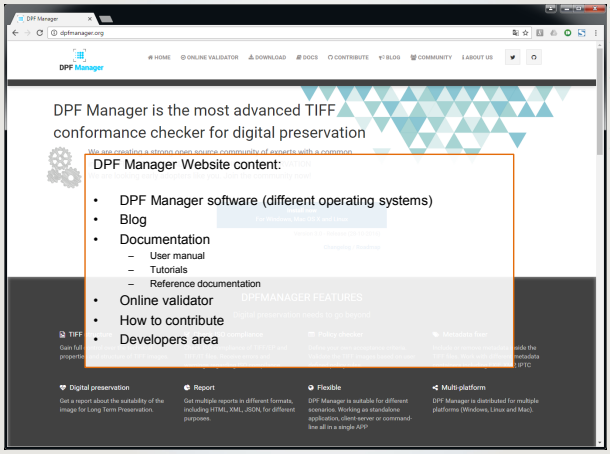
OAIS integrations
- Archivematica integration.

Reporting
- Global statistical analysis of TIFF checked

Contribute proposing new features

https://github.com/EasyinnovaSL/DPFManager/issues

DPF Manager Website



INDEX

Initial Presentation
TI/A Initiative
DPF Manager
Community
1. Join us!
2. Follow us!

Prof. JOSEP LLUÍS de la ROSA
Full Professor at University of Girona
peplluis@eia.udg.edu

UNIVERSITY OF GIRONA
www.udg.edu

Join us!

visit
www.dpfmanager.org

Memory Institutions
DPF Manager
Install now
For Windows, Mac OS X and Linux

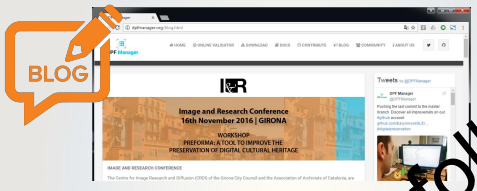
visit
www.ti-a.org

TIFF experts
TI A
Get involved!

visit
http://www.dpfmanager.org/community.html
https://github.com/EasyinnovaSL/DPFManager

Developers
github
New pull request

Follow us!



DPF Manager @DPFManager
TWEETS 630
FOLLOWERS 305

TI/A Initiative @TI_A_Standard
TWEETS 865
FOLLOWING 1,134
FOLLOWERS 414

1/3 of VIP follow us

MediaConch

Implementation and policy checking
on FFV1, Matroska, LPCM, and more

Jérôme Martinez, MediaArea

Experience Workshop - November 2016

PREFORMA

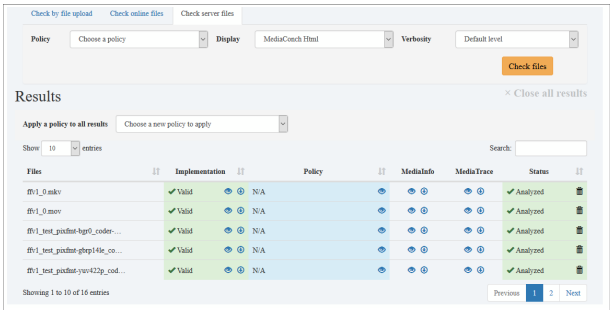
What is MediaConch?

MediaConch is a conformance checker

- Implementation checker
- Policy checker
- Reporter
- Fixer

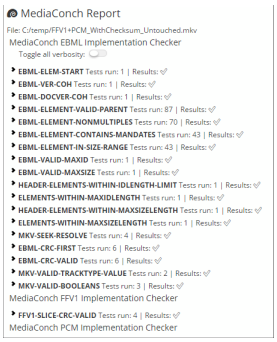
What is MediaConch?

Implementation and Policy reporter

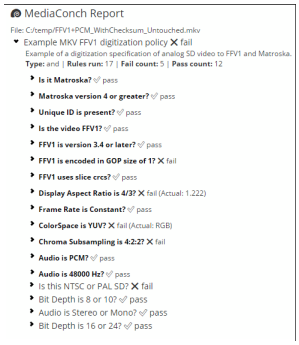


What is MediaConch?

Implementation report:

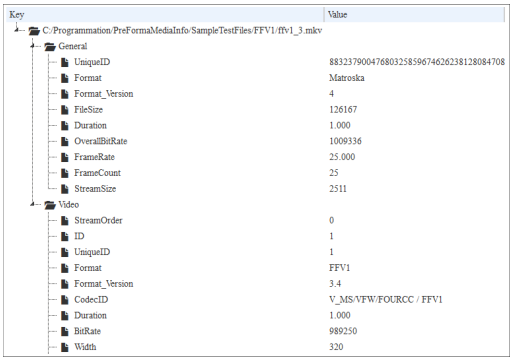


Policy report:



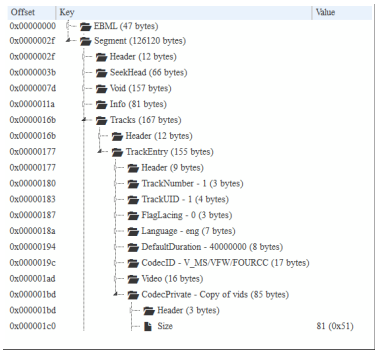
What is MediaConch?

General information about your files



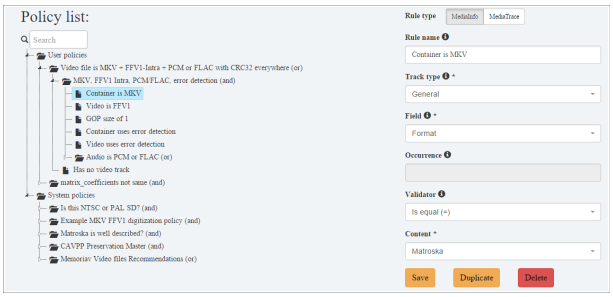
What is MediaConch?

Inspect your files



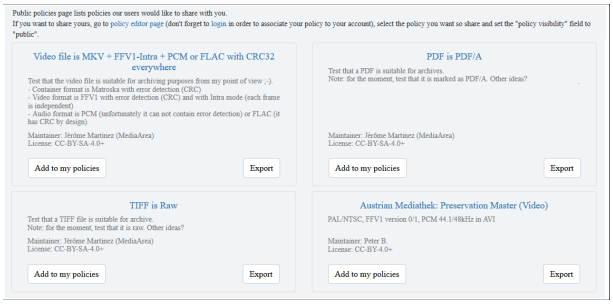
What is MediaConch?

Policy editor



What is MediaConch?

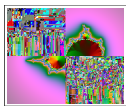
Public policies



What is MediaConch?

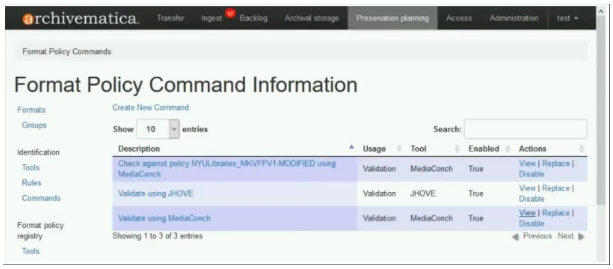
Fixer

- Segment sizes in Matroska
- Matroska “bit flip” correction
- FFV1 “bit flip” correction



Integration

Archivematica is an integrated suite of open-source software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model



MediaConch interfaces

- Graphical interface
- Web interface
- Command line
- Server (REST API)
- (Work in progress) a library (.dll/.so/.dylib)

MediaConch output formats

- XML (native format)
- Text
- HTML
- (Work in progress) PDF
- Tweakable! (with XSL)

Open source

- GPLv3+ and MPLv2+
- Relies on MediaInfo (metadata extraction tool)
- Use well-known open source libraries: Qt, sqlite, libevent, libxml2, libxslt, libexslt...

Supported formats

- Priorities for the implementation checker
  - Matroska
  - FFV1
  - PCM
- Can accept any format supported by MediaInfo for the policy checker
  - MXF + JP2k
  - QuickTime/MOV
  - Audio files (WAV, BWF, AIFF...)
  - ...

Supported formats

Can be expanded

- By plugins
  - Support of PDF checker: VeraPDF plugin
  - Support of TIFF checker: DPF Manager plugin
  - You use another checker? Let us know
- By internal development
  - More tests on your preferred format is possible
  - It depends on you!

Versatile

Several input formats are accepted

- FFV1 from MOV or AVI
- Matroska with other video formats
- (Work in progress) Extraction of a PDF or TIFF attachement from a Matroska container and analyze with a plugin (e.g. VeraPDF and DPF Manager)
- ...

Versatile

Input can be from:

- Files (local/network)
- FTP/FTPS/SFTP
- HTTP/HTTPS
- Amazon S3

Versatile

Binaries are provided for:

- Windows
- Mac

  Homebrew users:"brew install mediaconch", that's all!
- Linux (Ubuntu, Debian, Fedora, OpenSUSE...)

  Since Ubuntu 16.04 and Debian Testing/9 users:

  "apt-get install mediaconch" or in Ubuntu Store, that's all!

  (it is in the official distros repository)
- Embedded devices? Doable

  (we tested it on a Raspberry Pi )
- Can be ported on other distros (BSD...)

Standardization

- Matroska is widely used but not (yet) standardized
- FFV1 is gaining increasing usage in preservation contexts but is not (yet) standardized

CELLAR: IETF workgroup

- Open standards group
- Goal to IETF-standardize Matroska/FFV1/FLAC
- A lot of progress, especially with Matroska/EBML specs
- https://datatracker.ietf.org/wg/cellar/charter/

FFV1 performance

- NOA tested on SD 8-bit content:
  - i7-2600 (4 cores+HT, 3.4-3.8 GHz)
  - 3-4x real time
  - 4-5x decoding speed increase compared to JP2k
- VIAA is testing on SD 10-bit content (FFmpeg 3.2):
  - E5-2698V3 (16 cores+HT, 2.3-3.6 GHz)
  - 0.7x real time/thread, 11-12x real time/all cores+HT
  - 3-4x decoding speed increase compared to JP2k
  - Better compression ratio by 8-10% compared to JP2k

FFV1 performance

- This is an average, results varies depending on the content of files
  - From 0.4x to 2.4x (average 0.7x) real time/thread (encoding/decoding)
  - From 0.7x to 16x (average 3.5x) the speed of JP2k (FFmpeg)
- Not convinced?
  - Test on your own files
  - MediaArea will provide test scripts
  - We can perform tests for you

Worldwide

- 2 project leaders
  - Jerôme Martinez (Digital Media Analysis Specialist, France)
  - Dave Rice (Archivist, USA)
- Presentations worldwide
  - IASA, France
  - FIAT/IFTA, Austria
  - FOSDEM, Belgium
  - AMIA, USA
  - Code4Lib, USA
  - JTS, Singapore
  - (3-6 October 2016) IPRES, Switzerland
  - (25-29 September 2016) IASA, USA

Matroska research corpus

- We analyze all Matroska files from archive.org
- Interface with some statistics of Matroska elements usage (e.g. files with CRC-32 elements...)
  https://mediaarea.net/MediaConchCorpus/

What's next?

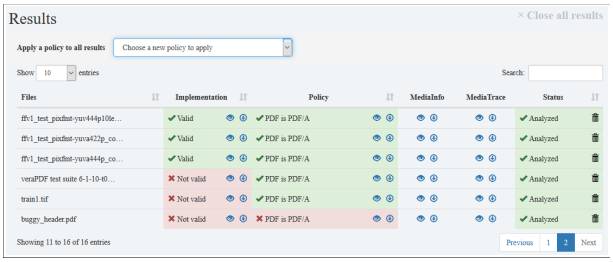
- Continue to improve handling of huge collections
- Continue to improve user interface
- Support of embedded attachments
- Statistics
- Finish standardization of Matroska and FFV1
- More conformance tests
- More fixing cases

And after PREFORMA sponsorship?

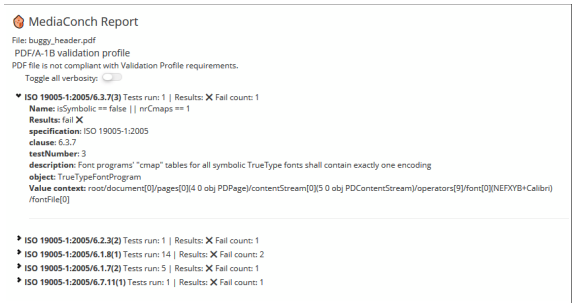
It depends on you!

- This is open source
- Driven by user requests
- Everyone can develop or sponsor a development
- Potential features:
  - Support of tests for your prefered format (MOV? MXF? JP2k? WAV?)
  - Support of other checkers (BWF MetaEdit? QCTools?)
  - Integration in your workflow
  - ...

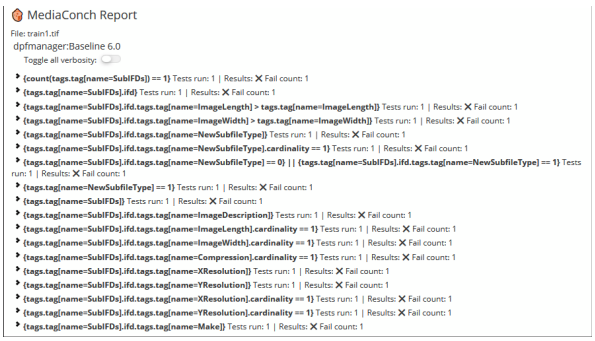
Example (Plugins)



Example (Plugins)



Example (Plugins)



Stay in touch

MediaArea: https://mediaarea.net, @MediaArea_net

MediaConch: https://mediaarea.net/MediaConch, @MediaConch

Jérôme Martinez: jerome@mediaarea.net

Slides: https://mediaarea.net/Events