PREFORMA

Preservation taken seriously

(at least … trying to take it that way)

The project …

PREservation **FORMA**ts for culture information/e-archives

Is an: EU-funded project (running since January 2014 until December 2017)

… with very different partners:

Partners comming from …

- Center of Expertise
- Cultural Heritage
- Software-Evaluation

PREFORMA

The partners …

Center of Expertise …

- PACKED EXPERTISECENTRUM DIGITAAL ERFGOED VZW, Belgium
- HOGSKOLAN I SKOVDE (University of Skovde), Sweden

Cultural Heritage …

- RIKSARKIVET, Sweden
- STICHTING NEDERLANDS INSTITUUT VOOR BEELD EN GELUID, Netherlands
- KONINKLIJK INSTITUUT VOOR HET KUNSTPATRIMONIUM, Belgium
- GREEK FILM CENTRE AE, Greece
- LOCAL GOVERNMENT MANAGEMENT AGENCY, Ireland
- STIFTUNG PREUSSISCHER KULTURBESITZ, Germany
- AYUNTAMIENTO DE GIRONA, Spain
- EESTI VABARIIGI KULTUURMINISTEERIUM, Estonia
- KUNGLIGA BIBLIOTEKET, Sweden

Software-Evaluation und – tests …

- UNIVERSITA DEGLI STUDI DI PADOVA, Italy
- FRAUNHOFER (Ilmenau), Germany

Coordination / media-partner…

- PROMOTER, Italy

PREFORMA

The problem …

PREservation **FORMA**ts for culture information/e-archives

What is it about ?

The problem …

1. Some information has to be preserved for a very long time

2. This information (often) is stored in files

3. These files use file-formats

4. These file-formats are (mostly) standardised

5. Usage of files (data, information, …) is done with programms that are build on the standards of file-formats

6. Files that should be readable in many years ahead have to follow the file-format standards

(If not: We cannot be sure that future programms can read/interpret the information contained in the files correctly)

The problem …

Not everything that calles itself a TIF-file is a TIF-file … and
Not everything that is a TIF-file, is a TIF6.0-file … and
Not everything that is a TIF6.0-file follows the TIF6.0-Baseline-Standard

Not everything that comes along as PDF is a PDF … and
Not everything that is a PDF is a PDF/A … and
Not everything that is a PDF/A is a PDF/A-1b

…
…
…

… What‘s in a name ? – A closer look is needed …

The problem …

To be able to „look more closely“ we need tools !

Development of good tools for file-format-validation – that is the objective of PREFORMA

The development of the tools is coordinated and controlled by Preforma-project.

The software development is done by companies/consortia that have been choosen by Preforma (and that get nearly all of the finances of the project)

The way …

The development of good tools for file-format-validation – objective of PREFORMA

1. PREFORMA-consortium had defined which formats should be validated

   Decision was for PDF, TIF, MKV/FFV1

2. PREFORMA-consortium had defined requirements for the tools

   Tools have to be easy-to-use, scalable, multilingual, … and (above all) OPEN SOURCE

3. PREFORMA-consortium organised a tender and choose companies/consortia

   The developers of the tools are …

… but …

The way…

The development of good tools for file-format-validation – objective of PREFORMA

… of course, there was an evaluation of existing file-format validation tools:

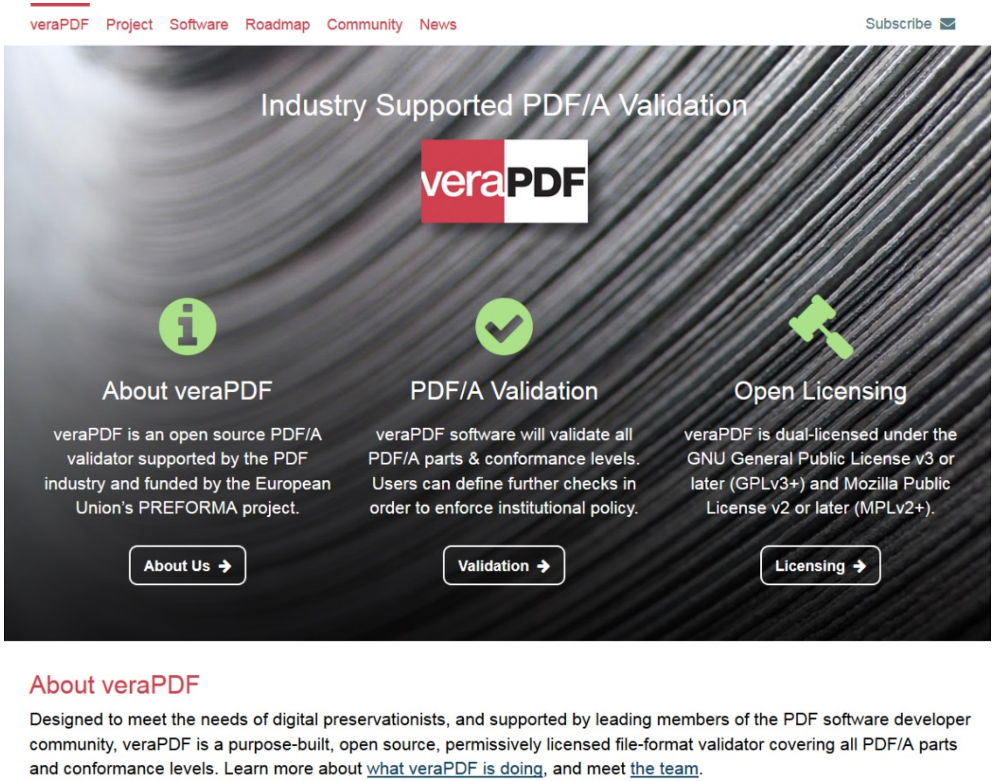
- There are very few good open-source validators
- It has been shown that the quality of the results of validation differs very much! While one validator says a given file is valid others decides that the same file is not valid
- It is often very hard to integrate existing validators into already established technical workflows for digital preservation

… now the developers:

The tools…

Tools for file-format validation …

PDF :: Is developed by a consortium of Open Preservation Foundation (OPF) and PDF Association – supported by Digital Preservation Coalition



Industry Supported PDF/A Validation

About veraPDF

veraPDF is an open source PDF/A validator supported by the PDF industry and funded by the European Union's PREFORMA project.

PDF/A Validation

veraPDF software will validate all PDF/A parts & conformance levels. Users can define further checks in order to enforce institutional policy.

Open Licensing

veraPDF is dual-licensed under the GNU General Public License v3 or later (GPLv3+) and Mozilla Public License v2 or later (MPLv2+).

About veraPDF

Designed to meet the needs of digital preservationists, and supported by leading members of the PDF software developer community, veraPDF is a purpose-built, open source, permissively licensed file-format validator covering all PDF/A parts and conformance levels. Learn more about what veraPDF is doing, and meet the team.

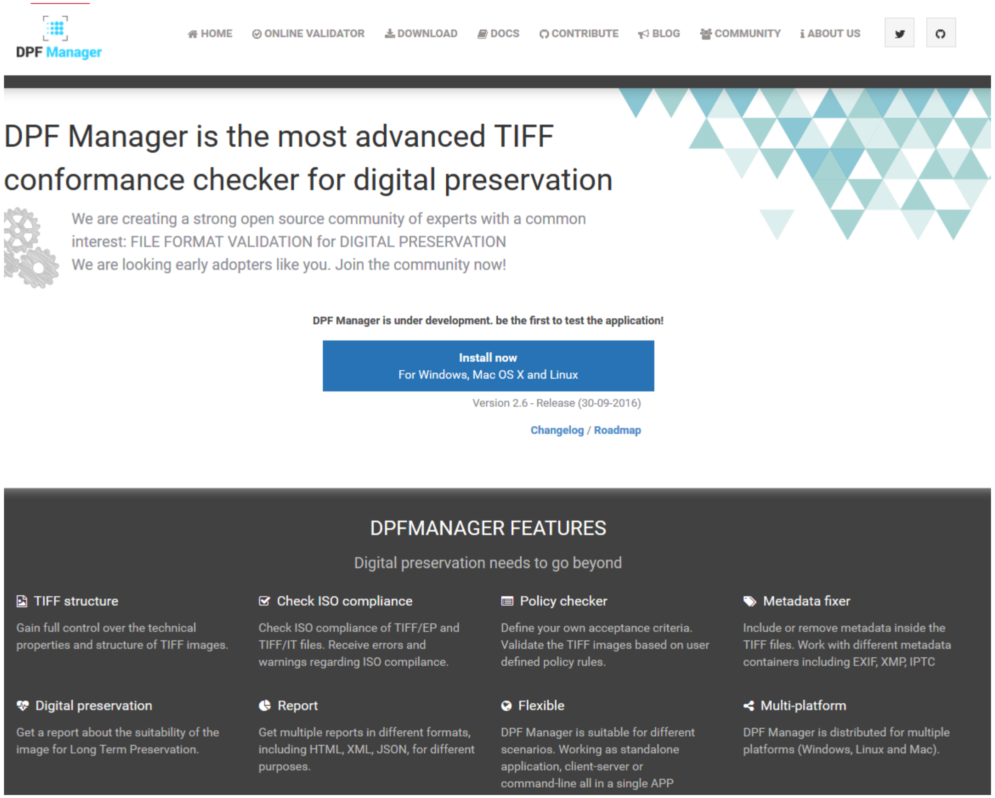
http://verapdf.org/home/

PREFORMA

The tools …

Tools for file-format validation …

TIF :: Is developed by easyinnova (Barcelona) and Digital Humanities Lab der Uni Basel



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

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

DPF Manager is under development. be the first to test the application!

Version 2.6 - Release (30-09-2016)

DPFMANAGER FEATURES

Digital preservation needs to go beyond

TIFF structure
Gain full control over the technical properties and structure of TIFF images.

Check ISO compliance
Check ISO compliance of TIFF/EP and TIFF/IT files. Receive errors and warnings regarding ISO compilance.

Policy checker
Define your own acceptance criteria. Validate the TIFF images based on user defined policy rules.

Metadata fixer
Include or remove metadata inside the TIFF files. Work with different metadata containers including EXIF, XMP, IPTC

Digital preservation
Get a report about the suitability of the image for Long Term Preservation.

Report
Get multiple reports in different formats, including HTML, XML, JSON, for different purposes.

Flexible
DPF Manager is suitable for different scenarios. Working as standalone application, client-server or command-line all in a single APP

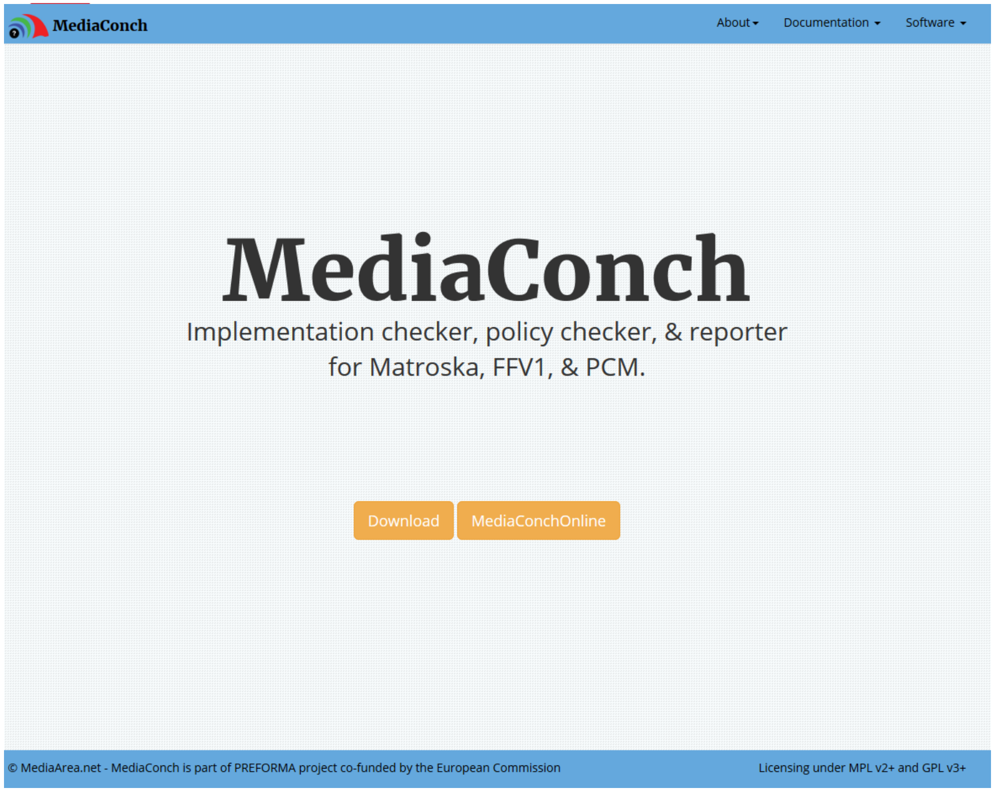
Multi-platform
DPF Manager is distributed for multiple platforms (Windows, Linux and Mac).

http://www.dpfmanager.org/

The tools …

Tools for file-format validation …

Matroska/FFV1 :: Is developed by mediaarea (developers of *mediainfo*) and supported by the developers of Matroska and of FFmpeg



MediaConch

Implementation checker, policy checker, & reporter for Matroska, FFV1, & PCM.

© MediaArea.net - MediaConch is part of PREFORMA project co-funded by the European Commission

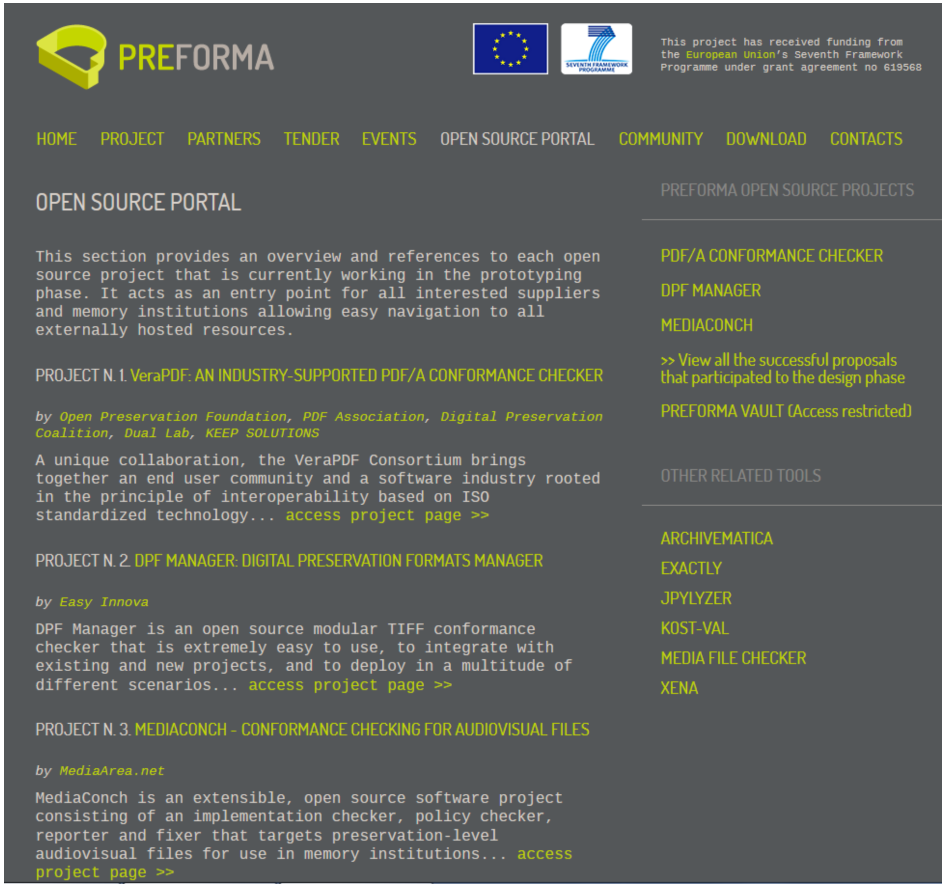
Licensing under MPL v2+ and GPL v3+

https://mediaarea.net/MediaConch/

The tools …

All consortia agreed to create OPEN-SOURCE software (GPLv3+)

More or less every month new „releases“ are published. The lastest versions of the three validators can be downloaded from the Preforma-Open Source Portal


PREFORMA
This project has received funding from the European Union’s Seventh Framework Programme under grant agreement no 619568
HOME
PROJECT
PARTNERS
TENDER
EVENTS
OPEN SOURCE PORTAL
COMMUNITY
DOWNLOAD
CONTACTS
PREFORMA OPEN SOURCE PROJECTS
PDF/A CONFORMANCE CHECKER
DPF MANAGER
MEDIACONCH
>> View all the successful proposals that participated to the design phase
PREFORMA VAULT (Access restricted)
OTHER RELATED TOOLS
ARCHIVEMATICA
EXACTLY
JPYLYZER
KOST-VAL
MEDIA FILE CHECKER
XENA

OPEN SOURCE PORTAL

This section provides an overview and references to each open source project that is currently working in the prototyping phase. It acts as an entry point for all interested suppliers and memory institutions allowing easy navigation to all externally hosted resources.

PROJECT N. 1. VeraPDF: AN INDUSTRY-SUPPORTED PDF/A CONFORMANCE CHECKER

by Open Preservation Foundation, PDF Association, Digital Preservation Coalition, Dual Lab, KEEP SOLUTIONS

A unique collaboration, the VeraPDF Consortium brings together an end user community and a software industry rooted in the principle of interoperability based on ISO standardized technology... access project page >>

PROJECT N. 2. DPF MANAGER: DIGITAL PRESERVATION FORMATS MANAGER

by Easy Innova

DPF Manager is an open source modular TIFF conformance checker that is extremely easy to use, to integrate with existing and new projects, and to deploy in a multitude of different scenarios... access project page >>

PROJECT N. 3. MEDIACONCH - CONFORMANCE CHECKING FOR AUDIOVISUAL FILES

by MediaArea.net

MediaConch is an extensible, open source software project consisting of an implementation checker, policy checker, reporter and fixer that targets preservation-level audiovisual files for use in memory institutions... access project page >>

http://www.preforma-project.eu/open-source-portal.html

The tools …

Testing the tools in their respective status of development was done perpetually. Everyone was invited to test the software and report errors or whishes …

An intense phase of tests coordinated and done by Preforma-partners started after the publication of „Release Candidates“ in december 2016. The test were done with big and small, valid and corrupted, real world files and synthetic test files, …

By december 2017 the development of the tools should be finished.

Each of the three tools works with APIs that are geared to each other. This way it is possible to create a „Meta-Tool“ incorporating all validators developed (and others to come)

The tools …

It‘s quite an effort to develop such tools. Some examples …

PDF/A might contain images, annotations and signatures → Have to be validated too
PDF/A might contain font-definitions, scripts, forms etc. → Have to be validated too
PDF/A might appear as PDF/A-1a, PDF/A-1b, PDF/A-2a, PDF/A-2b, PDF/A-2u, PDF/A-3 → The respective specifications have to be taken into account

TIFF might be based on different color-space-definitions → Has to be validated
TIFF might appear as TIFF-EP, LibTIFF, BigTIFF, TIFF-IT, GeoTIFF, … → Has to be validated
TIFF has a large amount of Tags, TIFF-Tags might be missing, contain wrong information, contain right information in a wrong way, might be placed at a wrong place → Each Tag has to be validated

It‘s quite an effort to develop such tools. Some examples …

(Footnote from http://www.digitalpreservation.gov/formats/content/tiff_tags.shtml)
TIFF image classes are described in the 1992 TIFF 6.0 specification and may be summarized as follows:

- Class B. Baseline bilevel.
- Class G. Baseline grayscale.
- Class P. Baseline palette-color.
- Class R. Baseline RGB.
- Class Y. Extension YCbCr.

The TIFF/IT specification (ISO 12639, 2004) defines the following image categories:

- CT. Color continuous-tone picture.
- LW. Color line art.
- HC. High-resolution continuous-tone.
- MP. Monochrome continuous-tone picture.
- BP. Binary picture.
- BL. Binary line art.
- SD. Screened data image.
- FP. Final page.

The tools …

It‘s quite an effort to develop such tools. Some examples …

Matroska/FFV1 has the problem that these format-codec-combination is just on the way to become widely used and to become a standard

Matroska actually is in the process of formal standardisation with the IETF (The Internet Engineering Task Force)

(One can validate the compliance to a standard only if a standard is well documented and widely used …)

Tools and Rules …

Important: Validating if a standard is followed … that cannot be the only thing

Standards (if existing) are – as shown – in a way flexible, they might be interpreted very strict of (in parts) more freely

To enable cultural heritage institutions to use the tools – the institutions have to be able to influence the validation:

Examples:

- Some cultural heritage institutions might define PDF/A-3 as the format of choice for preservation of text (allowing container-elements in the PDF), another institution decides their format of choice for text is PDF/A-1b (no container-elements allowed)

- One museum thinks it very important that in their TIFF files for each time-entry also the time zone is stored (TIFF/EP), another museum considers this as not so important and wants to check against the baseline standard only

Tools and Rules …

Validating if a standard is followed … that cannot be the only thing

Rules …

- Cultural heritage institutions have to be able to check against their own policies (interpretations of the standards). This implies that the tools must offer the policies als option (or must be able to store them as options)
- It has to be made easy for cultural heritage institutions to define their „rules“ and implement them in the „tools“

Fixer …

- In some cases missing or wrongly used tags can be reconstructed automatically from values stored in other tags … this way (sometimes) the compliance with a standard (and policy) might be created automatically. The preforma tools have a basic „metadata fixer“ component

Tools and Rules …

Validating if a standard is followed … that cannot be the only thing

Reports …

- It is very important that the tools create easily understandable reports and analyses. Even non-IT-people should be able to understand what and where the problems are
- The reports should be principally available in the language of the user
- The reports have to be available in machine-readable language too, to be passed to other programs that are eventually able to do more corrections

Tools and Rules …

Validating if a standard is followed … that cannot be the only thing

Integration …

- The tools have to be available as single (offline) version, they also have to allow shared use in a LAN or via web
- The tools have to make the integration into existing workflows for digital preservation easy

Scalability…

- The tools must be able to check very small and very big files and also be able to validate small or very big groups of files (e.g. folders with 10000 images)

PREFORMA

So far …

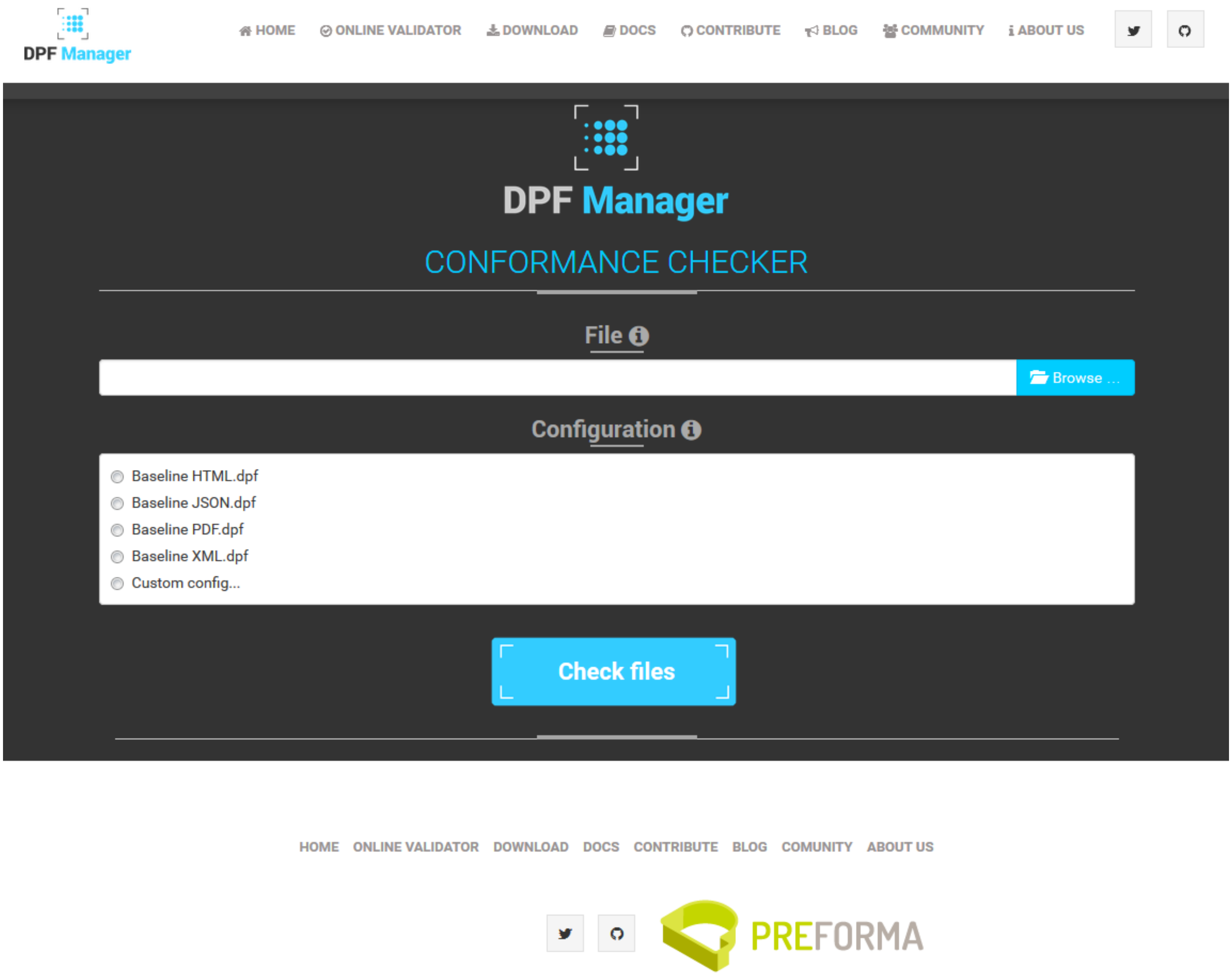
Status of development…

- The development is nearly finished. Tests revealed that at the end of the preforma project the tools are all working well.
- Everyone is free to continue the development, either by creating validators for other formats or by enhancing or improving the now existing validators.
- The software is created to enable multilinguality but the translations are not done yet (were not part of the project)

… take it, use it, improve it, share it …

!

Finally an example from the online-validator for TIF-files. The tested version of this form of the Tif-tool does not have the possibilities for defining / setting own „rules“ (explained above) …

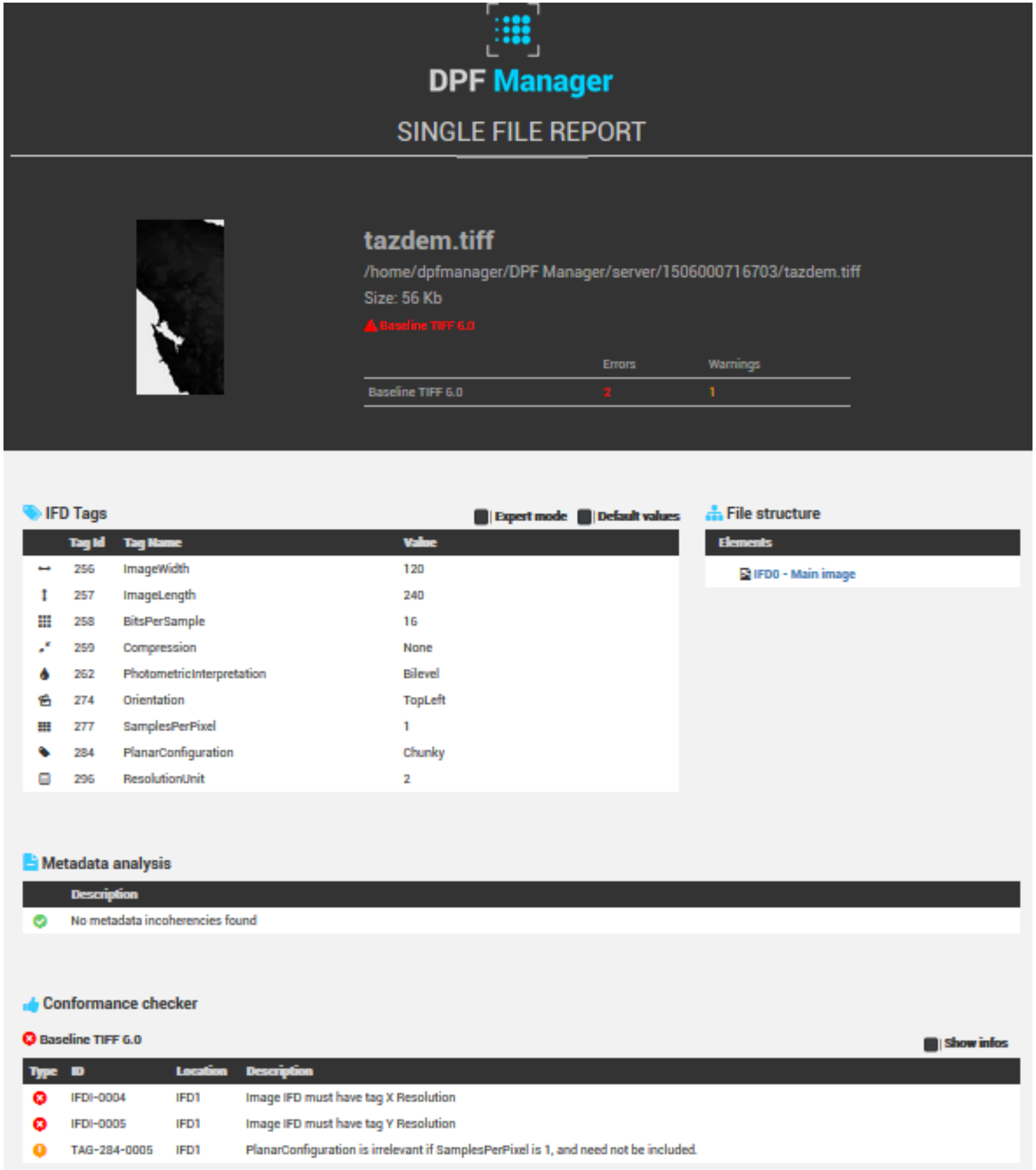


○ Baseline HTML.dpf
○ Baseline JSON.dpf
○ Baseline PDF.dpf
○ Baseline XML.dpf
○ Custom config...

The (limited) Online-validator (as of 2017-09-21)

A tiff-file …

http://dev.openlayers.org/releases/OpenLayers-2.13.1/examples/data/?



DPF Manager

SINGLE FILE REPORT

tazdem.tiff

/home/dpfmanager/DPF Manager/server/1506000716703/tazdem.tiff

Size: 56 Kb

⚠ Baseline TIFF 6.0

	Errors	Warnings
Baseline TIFF 6.0	2	1

IFD Tags □ Expert mode □ Default values

Tag Id	Tag Name	Value
256	ImageWidth	120
257	ImageLength	240
258	BitsPerSample	16
259	Compression	None
262	PhotometricInterpretation	Bilevel
274	Orientation	TopLeft
277	SamplesPerPixel	1
284	PlanarConfiguration	Chunky
296	ResolutionUnit	2

File structure

Elements
IFD0 - Main image

Metadata analysis

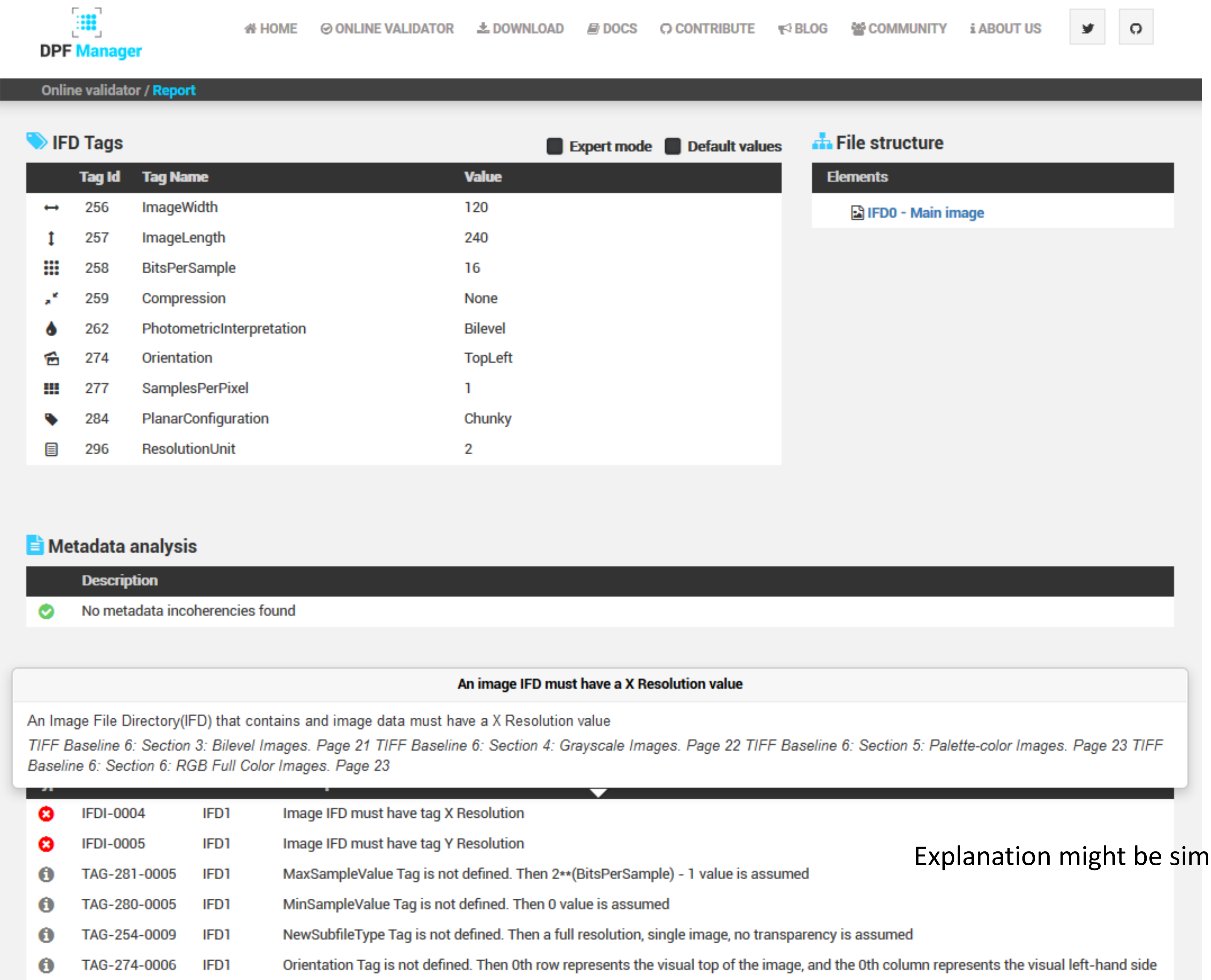
	Description
✓	No metadata incoherencies found

Conformance checker

✗ Baseline TIFF 6.0 □ Show infos

Type	ID	Location	Description
✗	IFDI-0004	IFD1	Image IFD must have tag X Resolution
✗	IFDI-0005	IFD1	Image IFD must have tag Y Resolution
!	TAG-284-0005	IFD1	PlanarConfiguration is irrelevant if SamplesPerPixel is 1, and need not be included.

The report …



Online validator / Report

IFD Tags

■ Expert mode ■ Default values

	Tag Id	Tag Name	Value
↔	256	ImageWidth	120
↕	257	ImageLength	240
	258	BitsPerSample	16
	259	Compression	None
	262	PhotometricInterpretation	Bilevel
	274	Orientation	TopLeft
	277	SamplesPerPixel	1
	284	PlanarConfiguration	Chunky
	296	ResolutionUnit	2

File structure

Elements
IFD0 - Main image

Metadata analysis

	Description
✓	No metadata incoherencies found

An image IFD must have a X Resolution value

An Image File Directory(IFD) that contains and image data must have a X Resolution value

TIFF Baseline 6: Section 3: Bilevel Images. Page 21 TIFF Baseline 6: Section 4: Grayscale Images. Page 22 TIFF Baseline 6: Section 5: Palette-color Images. Page 23 TIFF Baseline 6: Section 6: RGB Full Color Images. Page 23

☒	IFDI-0004	IFD1	Image IFD must have tag X Resolution
☒	IFDI-0005	IFD1	Image IFD must have tag Y Resolution
ℹ	TAG-281-0005	IFD1	MaxSampleValue Tag is not defined. Then 2**(BitsPerSample) - 1 value is assumed
ℹ	TAG-280-0005	IFD1	MinSampleValue Tag is not defined. Then 0 value is assumed
ℹ	TAG-254-0009	IFD1	NewSubfileType Tag is not defined. Then a full resolution, single image, no transparency is assumed
ℹ	TAG-274-0006	IFD1	Orientation Tag is not defined. Then 0th row represents the visual top of the image, and the 0th column represents the visual left-hand side

Explanation might be simpler!

All tools at: www.preforma-project.eu/open-source-portal.html

… take it, use it, improve it, share it …

PREFORMA

Final conference: Tallinn, 11-12. October 2017

Thank you very much !

Stefan Rohde-Enslin | Institut für Museumsforschung (SMB-PK) | s.rohde-enslin@smb.spk-berlin.de