



**PREFORMA**

Rules und Tools

## **PRE**servation **FORM**ats for culture information/e-archives

Ist zunächst ein EU-Projekt ...

- an dem sehr verschiedene Partner teilnehmen

Partner aus den Bereichen:

- Software-Technik
- Kulturerbe
- Software-Evaluation

### Software-Technik ...

PACKED EXPERTISECENTRUM DIGITAAL ERFGOED VZW, Belgien  
HOGSKOLAN I SKOVDE (Universität Skovde), Schweden

### Kulturerbe-Einrichtungen ...

STICHTING NEDERLANDS INSTITUUT VOOR BEELD EN GELUID, Niederlande  
KONINKLIJK INSTITUUT VOOR HET KUNSTPATRIMONIUM, Belgien  
GREEK FILM CENTRE AE, Griechenland  
LOCAL GOVERNMENT MANAGEMENT AGENCY, Irland  
STIFTUNG PREUSSISCHER KULTURBESITZ, Deutschland  
AYUNTAMIENTO DE GIRONA, Spanien  
EESTI VABARIIGI KULTUURMINISTEERIUM, Estland  
KUNGLIGA BIBLIOTEKET, Schweden

### Software-Evaluation und –Tests ...

UNIVERSITA DEGLI STUDI DI PADOVA, Italien  
FRAUNHOFER (Ilmenau), Deutschland

## **PRE**servati**ON FORM**ats for culture information/e-archives

Worum geht es ?

1. Es gibt Information, die man für lange Zeit bewahren muss
2. Diese Information ist (häufig) in Form von Dateien vorhanden
3. Diese Dateien haben bestimmte Formate
4. Diese Formate sind (zumeist) standardisiert
5. Die Datei-Nutzung geschieht mit Programmen, die auf diesen Standards aufbauen
6. Dateien, die noch in vielen Jahren genutzt werden sollen müssen diesen Format-Standards entsprechen

Nicht alles, was sich TIF nennt ist ein TIFF und  
Nicht alles, was ein TIF ist, ist ein TIF6.0 und  
Nicht alles, was ein TIF6.0 ist, ist ein TIF6.0 (Baseline)

Nicht alles, was als PDF daherkommt ist ein PDF und  
Nicht alles, was ein PDF ist, ist ein PDF/A und  
Nicht alles, was ein PDF/A ist, ist ein PDF/A-1b

...

...

...

... Namen sind Schall und Rauch – man muss genauer hinschauen ...

Um „genauer hinschauen“ zu können sind Werkzeuge nötig !

**Werkzeuge zur Format-Validierung entwickeln - das ist das Anliegen von PREFORMA**

Die Entwicklung der Werkzeuge wird von Preforma nur koordiniert und kontrolliert.

Die Software-Entwicklung übernehmen Firmen / Konsortien, die dafür von Preforma ausgewählt wurden (und fast alle Projektmittel erhalten)

## **Werkzeuge zur Format-Validierung entwickeln - das ist das Anliegen von PREFORMA**

1. PREFORMA-Konsortium hat definiert, welche Formate validierbar sein müssen

Entscheidung fiel auf PDF, TIF, MKV/FFV1

2. PREFORMA-Konsortium hat Anforderungen an die Werkzeuge formuliert

Werkzeuge müssen skalierbar sein, multilingual, ... und vor allem OPEN SOURCE

3. PREFORMA-Konsortium hat Ausschreibung geschaltet und Auswahl durchgeführt

Die Entwickler der Werkzeuge sind ...

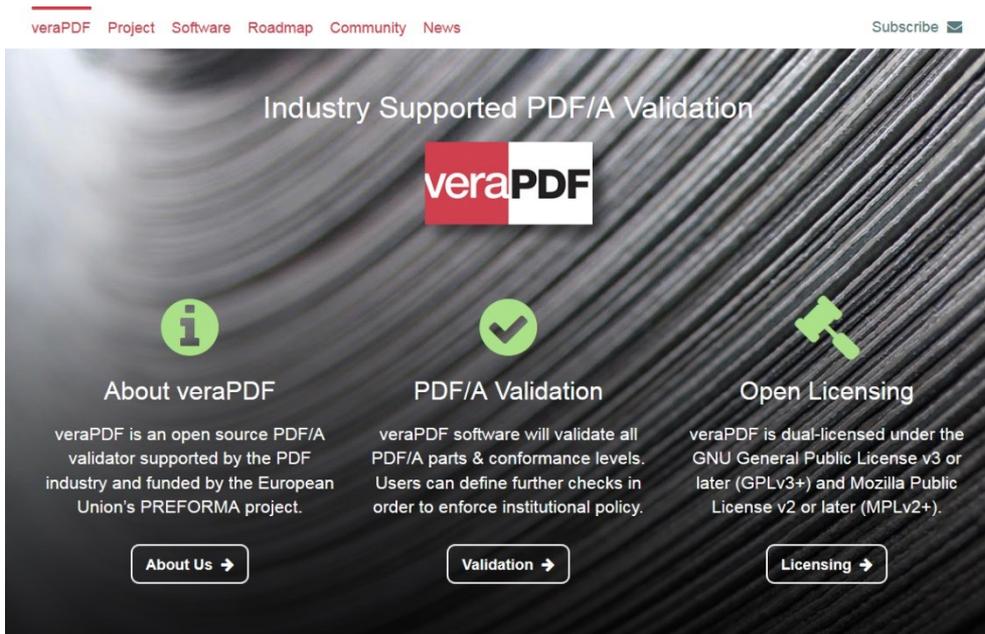
## **Werkzeuge zur Format-Validierung entwickeln - das ist das Anliegen von PREFORMA**

Es gab und gibt für diese Formate Validatoren, die getestet wurden. Sie haben ein paar Nachteile haben:

- Es gibt nur wenige opensource-Validatoren
- Es hat sich gezeigt, dass die Qualität des Validierens sehr unterschiedlich ist, während das eine Validator-Programm eine bestimmte Datei für valide erklärt, bestimmt ein anderes Validator-Programm die gleiche Datei als nicht-valide
- Die verfügbaren Validatoren lassen sich häufig nur sehr bedingt in bestehende technische Workflows integrieren

## Werkzeuge zur Format-Validierung entwickeln ...

**PDF** :: Wird entwickelt von einem Konsortium aus [Open Preservation Foundation \(OPF\)](#) und [PDF Association](#) – unterstützt von [Digital Preservation Coalition](#)



The screenshot shows the homepage of the veraPDF website. At the top, there is a navigation menu with links for 'veraPDF', 'Project', 'Software', 'Roadmap', 'Community', and 'News', along with a 'Subscribe' button. The main heading is 'Industry Supported PDF/A Validation'. Below this is the veraPDF logo. Three main sections are highlighted with green icons: 'About veraPDF' (information icon), 'PDF/A Validation' (checkmark icon), and 'Open Licensing' (gavel icon). Each section has a brief description and a button with a right-pointing arrow.

veraPDF Project Software Roadmap Community News Subscribe

### 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 Us](#) [Validation](#) [Licensing](#)

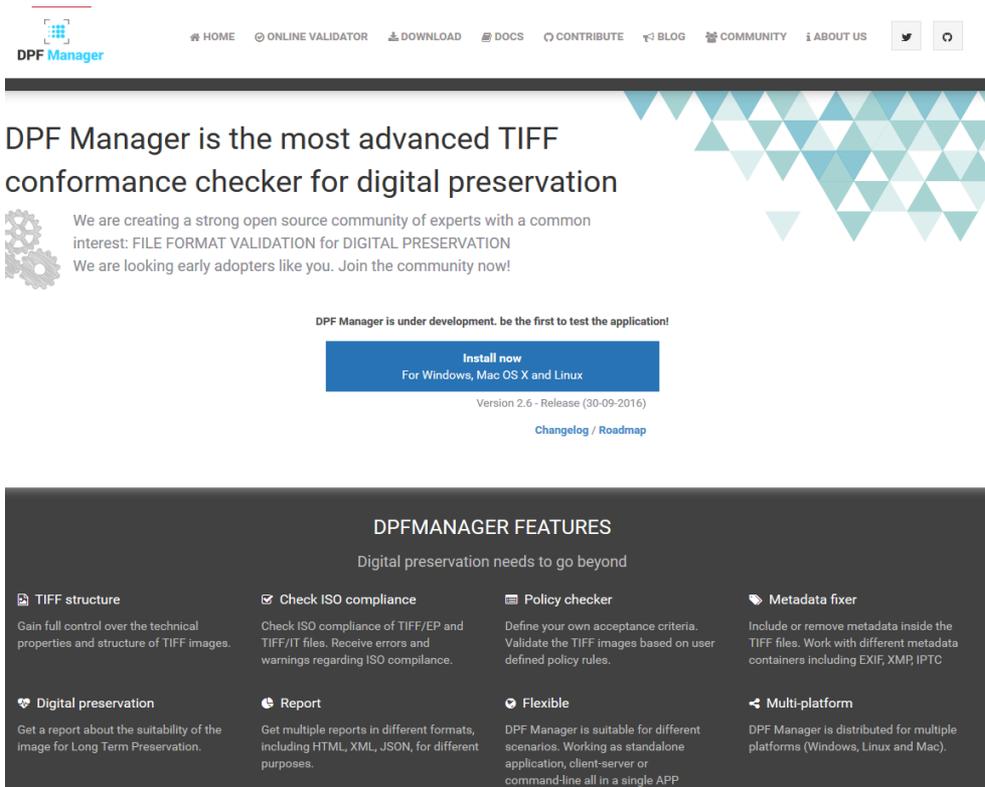
### About veraPDF

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/>

## Werkzeuge zur Format-Validierung entwickeln ...

**TIF** :: Wird entwickelt von [easyinnova](#) (Barcelona) und [Digital Humanities Lab](#) der Uni Basel



The screenshot shows the homepage of the DPF Manager project. At the top, there is a navigation bar with links for HOME, ONLINE VALIDATOR, DOWNLOAD, DOCS, CONTRIBUTE, BLOG, COMMUNITY, and ABOUT US, along with social media icons for Twitter and GitHub. The main heading reads "DPF Manager is the most advanced TIFF conformance checker for digital preservation". Below this, a paragraph states: "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!". A blue button labeled "Install now" is prominently displayed, with subtext "For Windows, Mac OS X and Linux" and "Version 2.6 - Release (30-09-2016)". A link for "Changelog / Roadmap" is also present. The lower section, titled "DPFMANAGER FEATURES", lists several capabilities: TIFF structure control, ISO compliance checking, a policy checker, metadata fixing, digital preservation reporting, flexibility in deployment, and multi-platform support.

DPF Manager

HOME ONLINE VALIDATOR DOWNLOAD DOCS CONTRIBUTE BLOG COMMUNITY ABOUT US

### 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!

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

Version 2.6 - Release (30-09-2016)

[Changelog / Roadmap](#)

#### DPFMANAGER FEATURES

Digital preservation needs to go beyond

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

<http://www.dpfmanager.org/>

## Werkzeuge zur Format-Validierung entwickeln ...

**Matroska/FFV1** :: Wird entwickelt von [mediaarea](#) (Entwickler von mediainfo) und unterstützt von den Entwicklern von Matroska und von FFmpeg



<https://mediaarea.net/MediaConch/>

Die Konsortien haben sich verpflichtet OPEN SOURCE-Software zu erstellen (GPLv3+)

Ca. monatlich werden neue „Releases“ veröffentlicht. Diese sind über die jeweiligen (gerade genannten) (Teil-)Projekt-Seiten oder vom Preforma-Open Source Portal



The screenshot shows the Preforma Open Source Portal website. At the top left is the Preforma logo. To its right are the European Union flag and the logo for the European Union's Seventh Framework Programme. Below these is a text box stating: "This project has received funding from the European Union's Seventh Framework Programme under grant agreement no 619568". A navigation menu includes links for HOME, PROJECT, PARTNERS, TENDER, EVENTS, OPEN SOURCE PORTAL, COMMUNITY, DOWNLOAD, and CONTACTS. The main content area is titled "OPEN SOURCE PORTAL" and contains an introductory paragraph: "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." Below this are three project entries: "PROJECT N.1. VeraPDF: AN INDUSTRY-SUPPORTED PDF/A CONFORMANCE CHECKER" by Open Preservation Foundation, PDF Association, Digital Preservation Coalition, Dual Lab, KEEP SOLUTIONS; "PROJECT N.2. DPF MANAGER: DIGITAL PRESERVATION FORMATS MANAGER" by Easy Innova; and "PROJECT N.3. MEDIACONCH - CONFORMANCE CHECKING FOR AUDIOVISUAL FILES" by MediaArea.net. A right-hand sidebar lists "PREFORMA OPEN SOURCE PROJECTS" including PDF/A CONFORMANCE CHECKER, DPF MANAGER, MEDIACONCH, and a link to view all successful proposals. Below this is "PREFORMA VAULT (Access restricted)" and "OTHER RELATED TOOLS" including ARCHIVEMATICA, EXACTLY, JPYLYZER, KOST-VAL, MEDIA FILE CHECKER, and XENA.

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

Tests der entwickelten Tools werden ständig durchgeführt. Jeder ist jederzeit eingeladen, die Software zu testen und Fehler etc. zu melden.

Eine intensive Testphase setzt nach Abschluss Entwicklung von „Release Candidates“ ab Dezember 2016 ein.

Im Dezember 2017 soll die Software-Entwicklung abgeschlossen sein.

Jedes der drei Tools arbeitet mit einer aufeinander abgestimmten API. Es wird so möglich sein hieraus ein „Meta-Tool“ zu entwickeln und Validatoren für andere Formate einzubinden.

Die Entwicklung der Tools hört sich leichter an, als sie ist. Ein paar Beispiele ...

PDF/A kann Bilder, Anmerkungen und Signaturen enthalten → Müssen ebenfalls validiert werden

PDF/A kann Schriftarten-Definitionen enthalten, ausführbare Skripte, Formulare etc. → Müssen validiert werden

PDF/A kann als PDF/A-1a, PDF/A-1b, PDF/A-2a, PDF/A-2b, PDF/A-2u, PDF/A-3 vorliegen → Die jeweiligen Spezifikationen sind zu berücksichtigen

TIFF kann auf verschiedenen Farbraumdefinitionen basieren → Muss validiert werden

TIFF kann als TIFF-EP, LibTIFF, BigTIFF, TIFF-IT, GeoTIFF, ... daher kommen → Muss validiert werden

TIFF hat eine große Menge TAGs, die alle einzeln oder in Gruppen falsch sein können → Tags müssen validiert werden

TIFF-Tags können fehlen, falsche Informationen enthalten, richtige Informationen in falscher Weise enthalten, können an falscher Stelle stehen → Muss validiert werden

Die Entwicklung der Tools hört sich leichter an, als sie ist. Ein paar Beispiele ...

(Fußnote von [http://www.digitalpreservation.gov/formats/content/tiff\\_tags.shtml](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.

Die Entwicklung der Tools hört sich leichter an, als sie ist. Ein paar Beispiele ...

Matroska/FFV1 hat vor allem das Problem, dass diese Format-/Codec-Kombination erst auf dem Weg ist, sich zu verbreitern und somit ein Quasi-Standard zu werden.

Matroska ist gerade im Standardisierungsprozess bei der IETF (The Internet Engineering Task Force (IETF®))

Man kann die Einhaltung von Standards aber erst prüfen, wenn der Standard sauber dokumentiert ist und sich durchgesetzt hat.

Prüfen, ob ein Standard im Prinzip eingehalten wurde ... das kann nicht alles sein

Die Standards (soweit vorhanden) sind – wie gezeigt – einigermaßen flexibel, sie können sehr strikt ausgelegt werden, sie können aber an manchen Stellen auch freier interpretiert werden.

Damit Kulturerbe-Einrichtungen die Tools verwenden können müssen Sie einen Einfluss darauf haben, wogegen die Validierung erfolgt.

Beispiele:

- Manche Kulturerbe-Einrichtung wird beispielsweise festlegen PDF/A-3 als Basisformat zu wählen (erlaubt Formulare), eine andere wird entscheiden, dass PDF/A-1b für sie das Basisformat für die Langzeitarchivierung sein soll.
- Ein Museum findet es wichtig dass zu den in ihren TIFF-Dateien gespeicherten Zeitangaben auch die Zeitzone angegeben wird (TIFF/EP), ein anderes Museum wird dieses nicht so wichtig finden und nur gegen den Baseline Standard validieren wollen.

Prüfen, ob ein Standard im Prinzip eingehalten wurde ... das kann nicht alles sein

### Rules ...

- Kulturerbe-Einrichtungen sollen also eigene Auslegungen (Policies) der Standards prüfen dürfen. D.h. in den Tools müssen die Regeln als Wahlmöglichkeiten (oder Eingabemöglichkeiten) hinterlegt werden können (oder hinterlegt worden sein).
- Kulturerbe-Einrichtungen müssen ihre „Rules“ definieren und diese den „Tools“ auf einfache Weise mitteilen können.

### Fixer ...

- In manchen Fällen kann aus einer Kombination anderer Tags oder durch Rückfragen seitens der Tools eine Konformität zu den gewählten Standards automatisch hergestellt werden. Die Tools werden einen sogenannten „Metadaten-Fixer“ enthalten.

Prüfen, ob ein Standard im Prinzip eingehalten wurde ... das kann nicht alles sein

### Reports ...

- Vor allem aber müssen die Tools allgemein verständliche Berichte ihrer Analysen geben. Auch Nicht-IT-Menschen sollen in der Lage sein, das ermittelte Problem zu verstehen.
- Die Berichtsroutinen sollen grundsätzlich Berichte in mehreren Sprachen liefern.
- Die Berichte müssen aber auch in maschinenlesbarer Form gegeben werden damit sie ggf. von (zu entwickelnden) Zusatzmodulen interpretiert (und umgesetzt) werden können (z.B. weiterführende automatische Korrektur).

Prüfen, ob ein Standard im Prinzip eingehalten wurde ... das kann nicht alles sein

### Integrierbarkeit ...

- Die Tools müssen als Einzelplatzversion, in einem LAN oder über das Internet einsetzbar sein.
- Die Tools müssen als Softwarepartikel in bestehende technische LZA-Workflows eingebunden werden können (Dateien werden automatisch weitergereicht).
- Die Tools müssen aber auch als „One Purpose-Version“ arbeiten können (Dateien oder Ordner müssen bei Programmaufruf einzeln verarbeitet werden können).

### Skalierbarkeit ...

- Die Tools müssen in der Lage sein kleine wie große Dateien zu verarbeiten, sie müssen auch in der Lage kleine wie große Mengen von Dateien zu verarbeiten.

Prüfen, ob ein Standard im Prinzip eingehalten wurde ... das kann nicht alles sein

### Stand der Entwicklung ...

- Die Entwicklung der Tools schreitet voran – Tests lassen sich durchführen. Allerdings funktioniert häufiger die eine oder andere Version (Windows, Linux, ...online, offline, ...) der Tools auch einmal nicht. Das ist normal, denn wir sind ja noch in der Entwicklung.
- Die Erfahrung zeigt, dass die Kollegen aus den Konsortien sehr kooperativ sind und sich bemühen entdeckte Fehler auszugleichen und dass Vorschläge für Verbesserungen ernst genommen werden.

Ein paar Beispiele aus dem Online-Validator-Tool für TIFF-Dateien. Diese Form des Tool hat noch keine Möglichkeit eigene „Rules“ zu geben ...



# DPF Manager

## CONFORMANCE CHECKER

### File

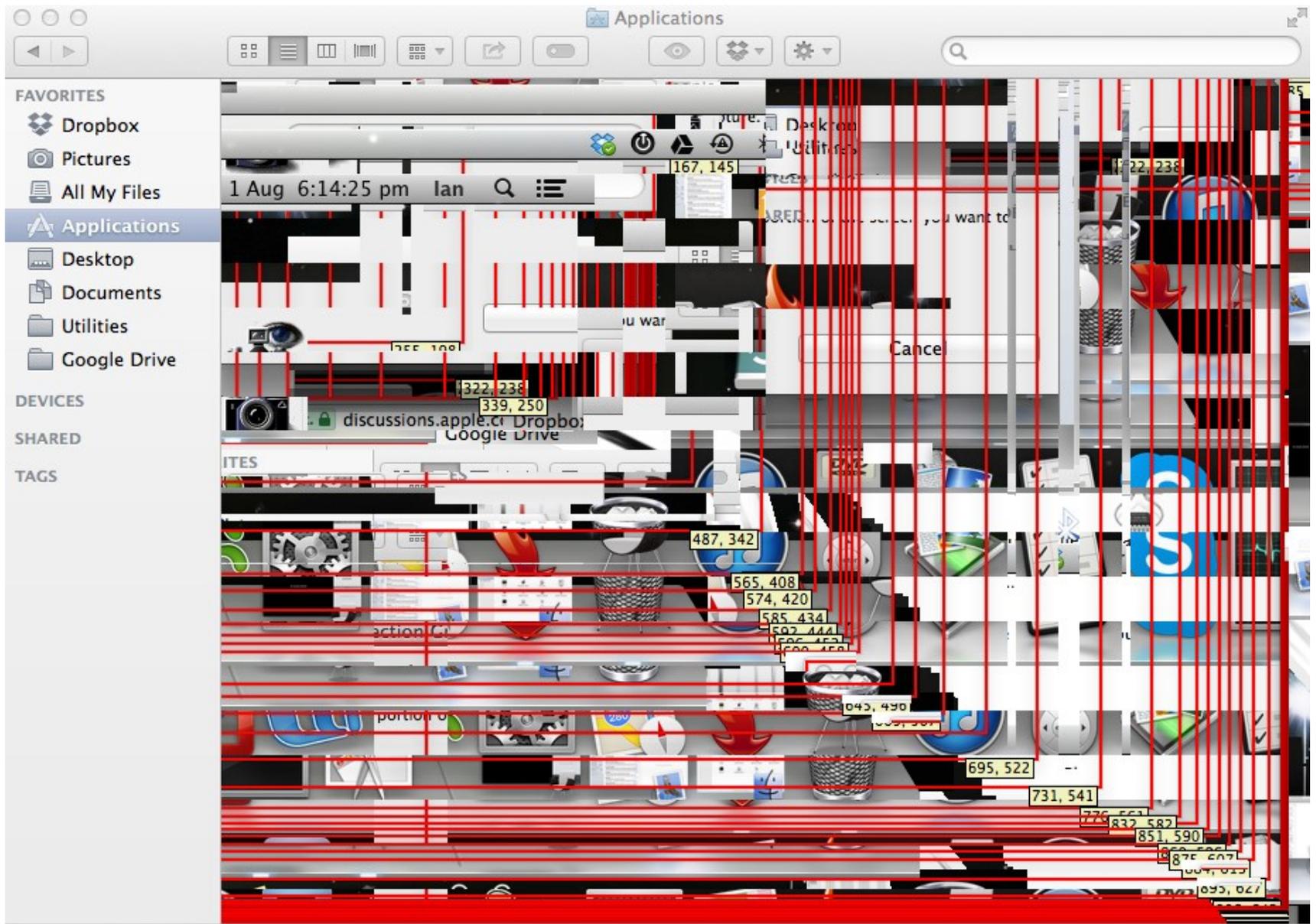
[Browse ...](#)

### Configuration

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

Check files



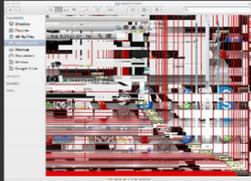


<https://discussions.apple.com/servlet/JiveServlet/showImage/2-26388528-457284/Corrupted+Applications.tiff>



# DPF Manager

## SINGLE FILE REPORT



### Corrupted\_Applications.tiff

/home/dpfmanager/DPF Manager/server/1475836365175/Corrupted\_Applications.tiff

▲ This file does NOT conform to conformance checker

	Errors	Warnings
Baseline	2	0

#### Tags

Expert mode

Tag Id	Tag Name	Value
256	ImageWidth	925
257	ImageLength	683
258	BitsPerSample	[8,8,8]
259	Compression	LZW
262	PhotometricInterpretation	RGB
274	Orientation	TopLeft
277	SamplesPerPixel	4
284	PlanarConfiguration	Chunky

#### Baseline conformance

Type	Location	Description
Error	IFD1	X Resolution tag must exist on node tags
Error	IFD1	Y Resolution tag must exist on node tags



#### File structure

IFD - Main image
ICC
Description: sRGB IEC61966-2.1
Version: 2.1
Class: Display

#### Baseline conformance

Type	Location	Description
Error	IFD1	X Resolution tag must exist on node tags
Error	IFD1	Y Resolution tag must exist on node tags

Back



# DPF Manager

## SINGLE FILE REPORT



**021\_101p.tif**  
/home/dpfmanager/DPF Manager/server/1475846656042/021\_101p.tif  
**▲ This file does NOT conform to conformance checker**

	Errors	Warnings
Baseline	13	0

<http://www.davince.com/showcase>

Tags  Expert mode

Tag Id	Tag Name	Value
256	ImageWidth	2528
257	ImageLength	3296
258	BitsPerSample	1
259	Compression	CCITT
262	PhotometricInterpretation	Bilevel
277	SamplesPerPixel	1
282	XResolution	300/1
283	YResolution	300/1

File structure

- IFD - Main image

### Baseline conformance

Type	Location	Description
Error	header	Bad alignment in offset on node header
Error	IFD1	Bad alignment in offset on node tag 254 NewSubfileType
Error	IFD1	Bad alignment in offset on node tag 256 ImageWidth
Error	IFD1	Bad alignment in offset on node tag 257 ImageLength
Error	IFD1	Bad alignment in offset on node tag 258 BitsPerSample
Error	IFD1	Bad alignment in offset on node tag 259 Compression
Error	IFD1	Bad alignment in offset on node tag 273 StripOffsets
Error	IFD1	Bad alignment in offset on node tag 277 SamplesPerPixel
Error	IFD1	Bad alignment in offset on node tag 278 RowsPerStrip



### Baseline conformance

Type	Location	Description
Error	header	Bad alignment in offset on node header
Error	IFD1	Bad alignment in offset on node tag 254 NewSubfileType
Error	IFD1	Bad alignment in offset on node tag 256 ImageWidth
Error	IFD1	Bad alignment in offset on node tag 257 ImageLength
Error	IFD1	Bad alignment in offset on node tag 258 BitsPerSample
Error	IFD1	Bad alignment in offset on node tag 259 Compression
Error	IFD1	Bad alignment in offset on node tag 262 PhotometricInterpretation
Error	IFD1	Bad alignment in offset on node tag 273 StripOffsets
Error	IFD1	Bad alignment in offset on node tag 277 SamplesPerPixel
Error	IFD1	Bad alignment in offset on node tag 278 RowsPerStrip

Kann noch verständlicher werden, oder?

Wer die Software testen und mit den Schöpfern sprechen mag, der ist sehr willkommen !

23.11.2016 Kunstgewerbemuseum, Kulturforum, Berlin

The image is a promotional poster for a workshop. At the top left, there is a circular logo with a stylized 'P' and the text 'TABLE ber 1946'. To its right is a detailed drawing of a dirigible. The main logo, 'PREFORMA', features a yellow 3D ribbon forming a square with a diagonal cut, followed by the word 'PREFORMA' in a bold, sans-serif font. Below this, the text 'IMPROVING LONG-TERM DIGITAL PRESERVATION EXPERIENCE WORKSHOP' is centered. A yellow rectangular box contains the date '23 NOVEMBER 2016'. Below the date, the location 'Kulturforum Matthäikirchplatz, 10785 Berlin' is listed. The background is a collage of historical and cultural items, including a book cover with 'ALDERDOMS BRODER CARL M. BELLMAN.', a newspaper clipping with 'Röster i Radio · T RIV', and a photograph of a woman wearing a yellow headband. The overall color palette is muted, with yellow and white text providing contrast.

Bis dahin lautet die Empfehlung an alle:

Die Validatoren downloaden, installieren, testen. Fehler melden und Vorschläge machen

Vielen Dank

Stefan Rohde-Enslin, Institut für Museumsforschung (SMB-PK), [s.rohde-enslin@smb.spk-berlin.de](mailto:s.rohde-enslin@smb.spk-berlin.de)