PREFORMA

Rules und Tools

PREservation **FORMA**ts 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

Das Problem …

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

Worum geht es ?

Das Problem …

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

Das Problem …

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 …

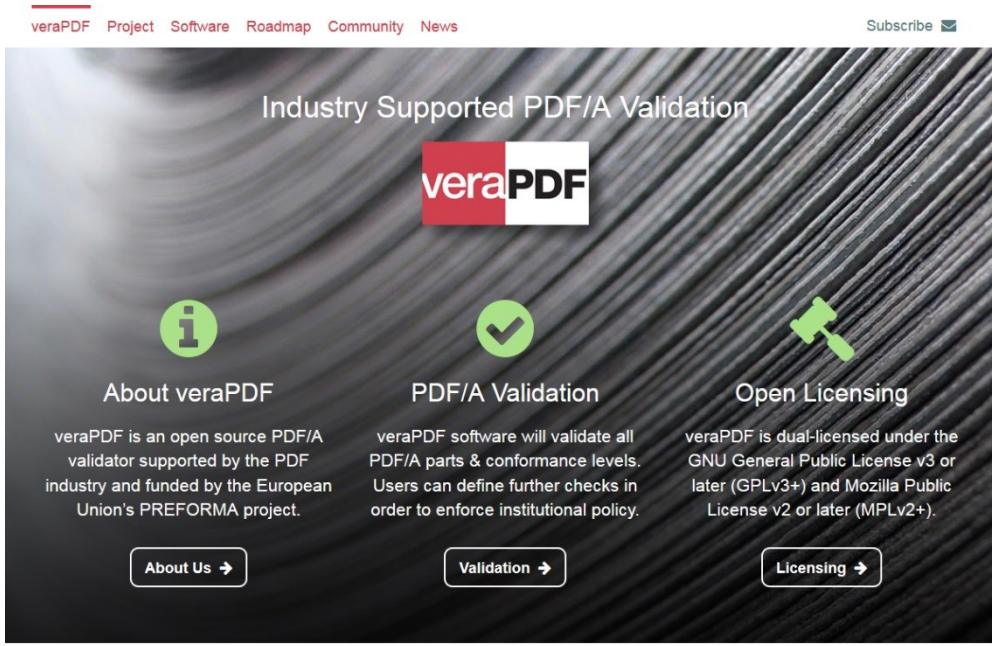
Werkzeuge zur Format-Validierung entwickeln - das ist das Anliegen von PREFORMA

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 Qualtität 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 …

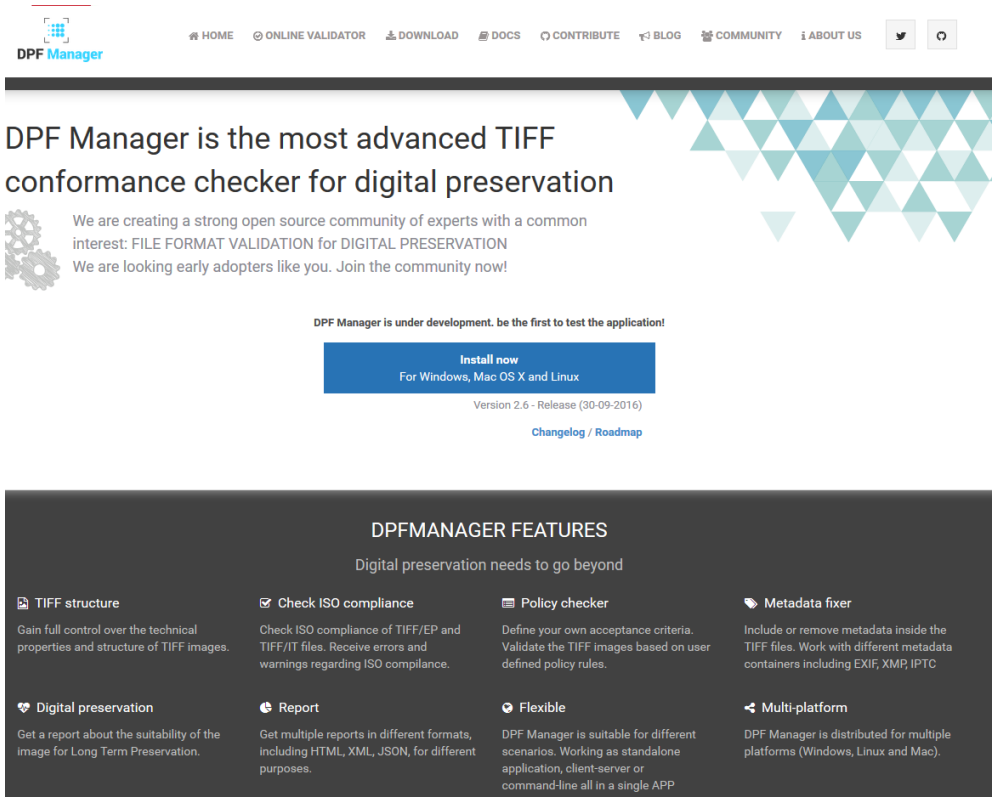
PDF :: Wird entwickelt von einem Konsortium aus Open Preservation Foundation (OPF) und PDF Association – unterstützt von Digital Preservation Coalition



http://verapdf.org/home/

Werkzeuge zur Format-Validierung entwickeln …

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



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 Tools …

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

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

Die Tools …

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 Tools …

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, … daherkommen → 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)
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 Tools …

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.

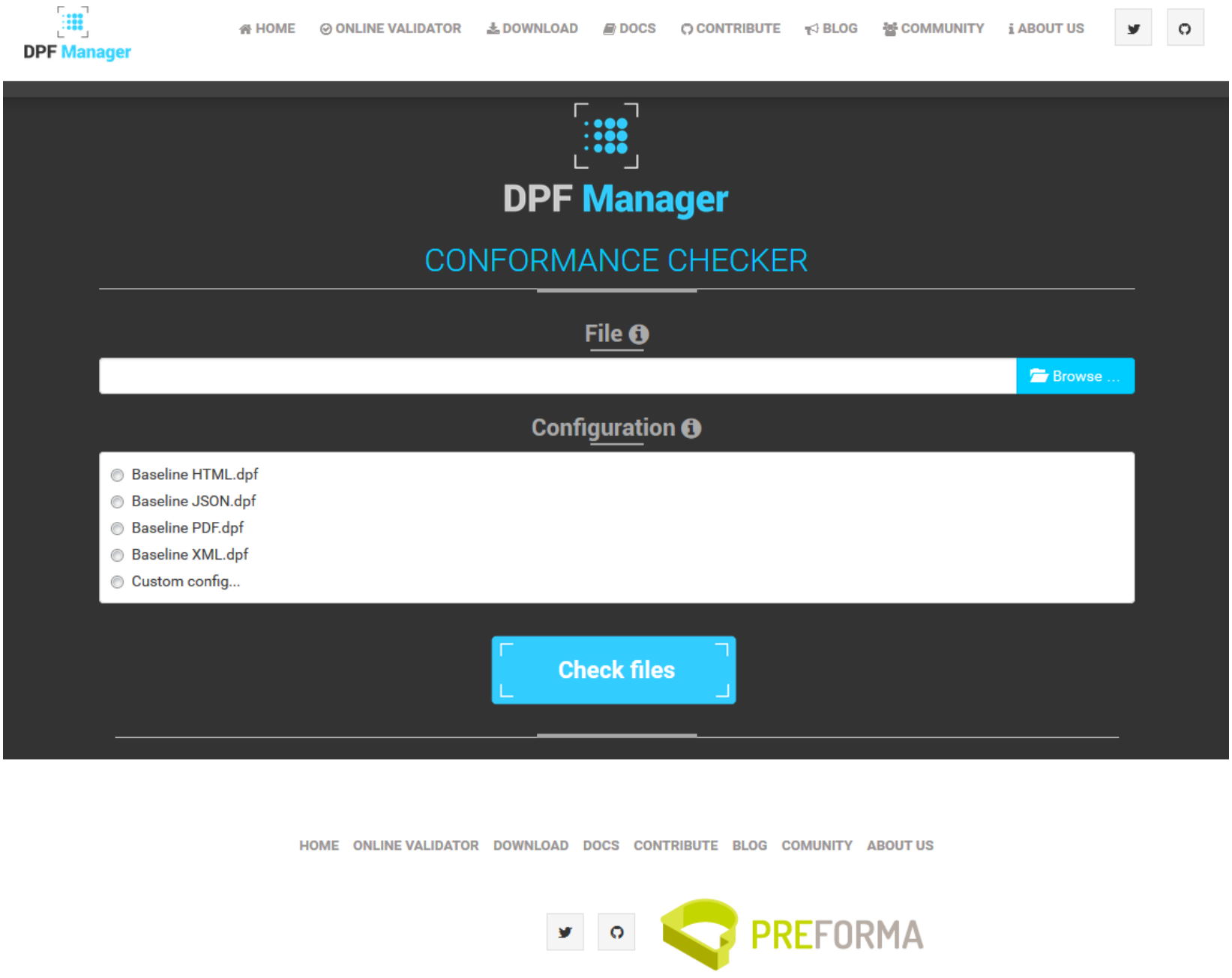
So far …

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

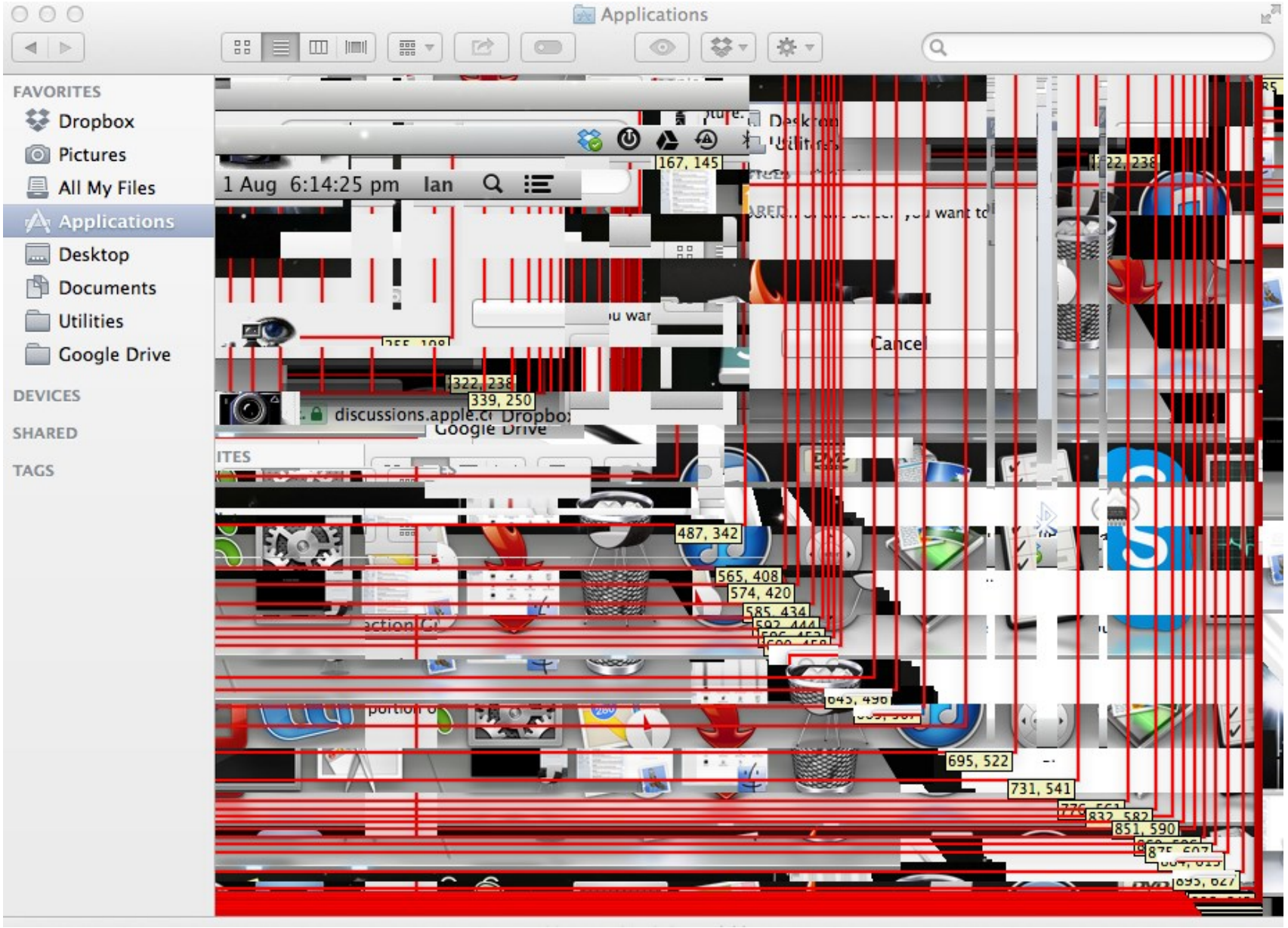
Stand der Entwicklung …

- 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 …



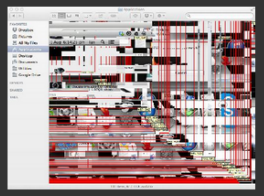
Der (sehr eingeschränkte) Online-Validator (Stand 07.10.2016)



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	[8,8,8,8]
259	Compression	LZW
262	PhotometricInterpretation	RGB
274	Orientation	TopLeft
277	SamplesPerPixel	4
284	PlanarConfiguration	Chunky

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

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

HOME | ONLINE VALIDATOR | DOWNLOAD | DOCS | CONTRIBUTE | BLOG | COMMUNITY | ABOUT US

Online validator / Report

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 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

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

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

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