

PREFORMA

Nichts ist wie es scheint

(manchmal aber doch)

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

Ist ein: EU-Projekt (Januar 2014 bis Dezember 2017)

... mit sehr verschiedenen Partnern:

Partner kommen aus...

- Kompetenzzentren
- Kulturerbe-Einrichtungen
- Spezialisten für Software-Evaluation

Kompetenzzentren ...

- PACKED EXPERTISECENTRUM DIGITAAL ERFGOED VZW, Belgien
- HOGSKOLAN I SKOVDE (University of Skovde), Schweden

Kulturerbe-Einrichtungen...

- RIKSARKIVET, Schweden
- 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, Katalonien
- EESTI VABARIIGI KULTUURMINISTEERIUM, Estland
- KUNGLIGA BIBLIOTEKET, Schweden

Spezialisten für Software-Evaluation und -tests ...

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

Koordination / media-partner...

- PROMOTER, Italien

PREservati**ON FORM**ats for culture information/e-archives

Worum geht es ?

1. Einige Information muss/soll für lange Zeit bewahrt werden
1. Solche Information ist zumeist in Dateien gespeichert
1. Diese Dateien liegen in Dateiformaten vor
1. Die Dateiformate sind – größtenteils – standardisiert
1. Nutzung der Dateien (Daten, Informationen, ...) geschieht mit Programmen, die auf Grundlage der Standards für Dateiformate geschaffen wurden
1. Dateien, die in vielen Jahren lesbar sein sollen müssen den Dateiformat-Standards entsprechen

(Falls nicht: Wir können nicht sicher sein, dass zukünftige Programme die Information, die in den Dateien enthalten sind, lesen und korrekt interpretieren können)

Nicht alles, dass sich TIF nennt, ist eine TIF-Datei ... und
Nicht alles, dass eine TIF-Datei ist, ist eine TIF 6.0-Datei ... und Nicht jede TIF
6.0 Datei folgt dem TIF 6.0-Baseline-Standard

Nicht alles, dass sich PDF nennt, ist eine PDF-Datei ... und
Nicht alles, dass eine PDF-Datei ist, ist eine PDF/A-Datei ... und Nicht jede
PDF/A Datei folgt dem PDF/A-1b-Standard

...

...

...

... What's in a name ? – Genaueres Hinsehen ist nötig

Um genauer hinzusehen brauchen wir Werkzeuge !

Die Entwicklung guter Werkzeuge zur Dateiformat-Validierung – das ist das Ziel von PREFORMA

Dabei wird die Entwicklung der Werkzeuge vom Preforma-Projekt koordiniert und kontrolliert.

Die Software-Entwicklung selbst wird von Firmen/Konsortieen durchgeführt, die das Preforma-Team ausgewählt hat.

Die Entwicklung guter Werkzeuge zur Dateiformat-Validierung – das ist das Ziel von PREFORMA

1. PREFORMA-Team definiert welche Formate zu validieren sind

Entscheidung fiel auf PDF, TIF, MKV/FFV1

2. PREFORMA-Team definiert die Anforderungen an die Werkzeuge

Werkzeuge müssen leicht nutzbar, skalierbar, mehrsprachenfähig, ...
und vor allem OPEN SOURCE sein

3. PREFORMA-Team organisiert Ausschreibung und wählt Entwickler

Die Entwickler der Werkzeuge sind ...

... aber ...

Die Entwicklung guter Werkzeuge zur Dateiformat-Validierung – das ist das Ziel von PREFORMA

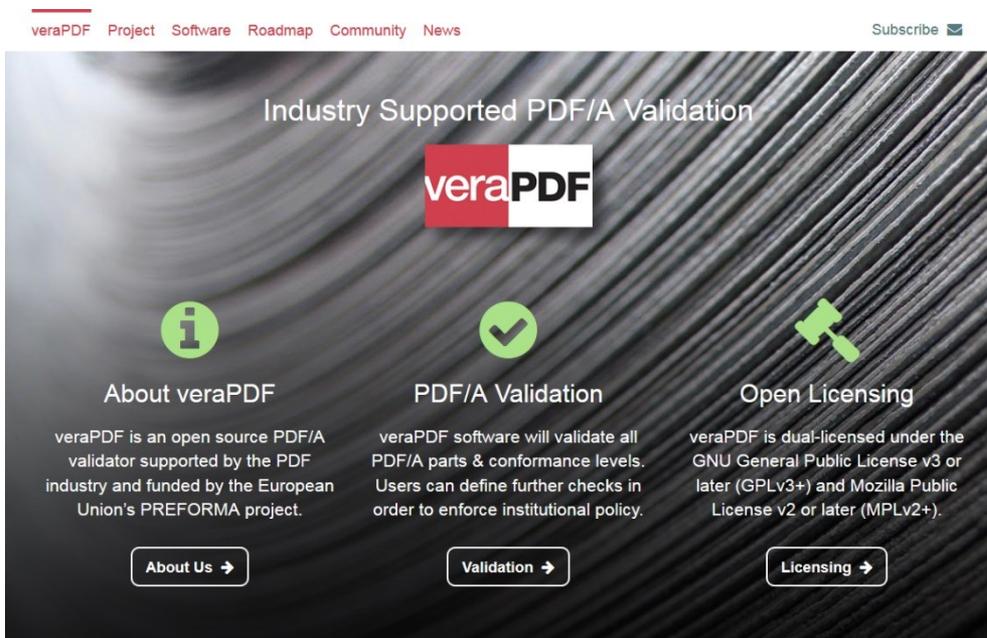
... selbstverständlich wurde vorher eine Auswertung vorhandener Validierungswerkzeuge vorgenommen:

- Es gibt nur sehr wenige gute open-source Validatoren
-
- Test zeigen, dass die Qualität der Validierung sehr sehr unterschiedlich ist! Während der eine Validator eine Datei als valide bezeichnet entscheidet ein anderer Validator, dass dem nicht so ist.
-
- Es ist oft sehr schwer existierende Validatoren in bereits bestehende technische Workflows für Langzeitarchivierung zu integrieren.

... jetzt zu den Entwicklern:

Werkzeuge zur Dateiformat-Validierung ...

PDF :: Wird entwickelt von 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 contains 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.
[About Us](#)
- PDF/A Validation**
veraPDF software will validate all PDF/A parts & conformance levels. Users can define further checks in order to enforce institutional policy.
[Validation](#)
- Open Licensing**
veraPDF is dual-licensed under the GNU General Public License v3 or later (GPLv3+) and Mozilla Public License v2 or later (MPLv2+).
[Licensing](#)

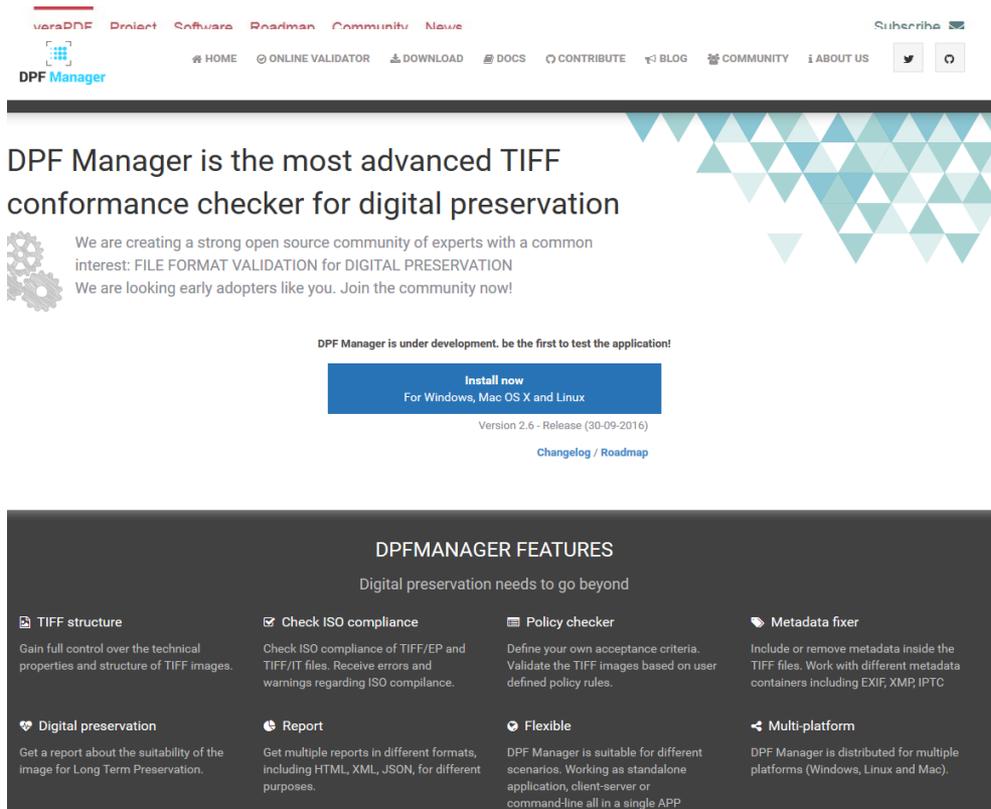
About veraPDF

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 Dateiformat-Validierung ...

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



The screenshot shows the homepage of the DPF Manager website. At the top, there is a navigation bar with links for 'veraDPF', 'Project', 'Software', 'Roadmap', 'Community', and 'News'. Below this is a secondary navigation bar with 'HOME', 'ONLINE VALIDATOR', 'DOWNLOAD', 'DOCS', 'CONTRIBUTE', 'BLOG', 'COMMUNITY', and 'ABOUT US'. A 'Subscribe' button is also present. The main heading reads 'DPF Manager is the most advanced TIFF conformance checker for digital preservation'. Below this, a message 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 prominent blue button says 'Install now For Windows, Mac OS X and Linux', with 'Version 2.6 - Release (30-09-2016)' and links for 'Changelog / Roadmap' below it. The bottom section, titled 'DPFMANAGER FEATURES', lists: '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), and 'Multi-platform' (DPF Manager is distributed for multiple platforms (Windows, Linux and Mac)).

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

Werkzeuge zur Dateiformat-Validierung ...

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



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

Alle Entwickler schaffen OPEN-SOURCE software (GPLv3+)

Fast jeden Monat wurden neue „Releases“ veröffentlicht. Die jeweils letzte Version der drei Validatoren kann vom Preforma-Open Source Portal heruntergeladen werden.



The screenshot shows the PREFORMA Open Source Portal website. The header includes the PREFORMA logo, the European Union flag, and the 7th Framework Programme logo. The navigation menu includes HOME, PROJECT, PARTNERS, TENDER, EVENTS, OPEN SOURCE PORTAL, COMMUNITY, DOWNLOAD, and CONTACTS. The main content area is titled "OPEN SOURCE PORTAL" and contains a description of the section, followed by three project entries: Project N. 1 (VeraPDF), Project N. 2 (DPF Manager), and Project N. 3 (MediaConch). A sidebar on the right lists "PREFORMA OPEN SOURCE PROJECTS" and "OTHER RELATED TOOLS".

PREFORMA

HOME PROJECT PARTNERS TENDER EVENTS OPEN SOURCE PORTAL COMMUNITY DOWNLOAD CONTACTS

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 >>](#)

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
- JPPLYZER
- KOST-VAL
- MEDIA FILE CHECKER
- XENA

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

Jede neue Version der Werkzeuge wurde vom Preforma-Tool getestet. Jeder war eingeladen die Software zu testen, Fehler zu berichten und Wünsche zu äußern ...

Ab Dezember 2016, nach der Veröffentlichung von „Release Candidates“, begann eine Phase koordinierter und kontrollierter Tests durch die Preforma-Partner. Getestet wurden große, kleine, valide, korrupte, echt-weltliche und synthetische Dateien ...

Im Dezember 2017 wird die Entwicklung der Werkzeuge abgeschlossen sein.

Jedes der drei Werkzeuge arbeitet mit APIs die aufeinander abgestimmt sind. Auf diese Weise ist es möglich ein „Meta-Werkzeug“ zu erstellen, das alle drei Validatoren integrieren kann (und weitere, die noch zu entwickeln sind).

Es ist keine Kleinigkeit solche Werkzeuge zu entwickeln. Beispiele:

PDF/A kann Abbildungen, Anmerkungen und Signaturen enthalten.

Solche Teile müssen ebenfalls validiert werden

PDF/A kann Font-Definitionen, Skripte und Formulare enthalten

Solche Teile müssen ebenfalls validiert werden

PDF/A kann auftreten als PDF/A-1a, PDF/A-1b, PDF/A-2a, PDF/A-2b, PDF/A-2u, PDF/A-3

Die entsprechenden Spezifikationen müssen berücksichtigt werden

TIFF kann auf verschiedenen Farbraum-Definitionen basieren

Die Angaben müssen validiert werden

TIFF kann als TIFF-EP, LibTIFF, BigTIFF, TIFF-IT, GeoTIFF, ... erscheinen

Übereinstimmung mit Spezifikationen ist zu validieren

TIFF hat viele sog. TAGs, TIFF-Tags können fehlen, falsche Information enthalten, korrekte Information in falscher Weise enthalten, an falscher Stelle platziert sein

Jeder TAG muss validiert werden

Es ist keine Kleinigkeit solche Werkzeuge zu entwickeln. Beispiele:

(Fußnote aus 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.

Die TIFF/IT Spezifikation (ISO 12639, 2004) definiert die folgenden „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.

Es ist keine Kleinigkeit solche Werkzeuge zu entwickeln. Beispiele:

Matroska/FFV1 hat das Problem, dass diese Format-Codec-Kombination gerade erst dabei ist ein breit genutzter Standard zu werden

Matroska ist gerade im Prozess der formalen Standardisierung durch die IETF (The Internet Engineering Task Force)

(Man kann die Entsprechung zu einem Standard nur überprüfen, wenn der Standard gut dokumentiert ist ...)

Wichtig: Feststellen, ob eine Datei einem Standard entspricht ... das kann nicht alles sein !

Standards (sofern vorhanden) sind – wie gezeigt – eigentlich flexibel, sie können sehr strikt interpretiert werden oder stellenweise recht frei verstanden werden

Um Kulturerbe-Einrichtungen in die Lage zu versetzen, die Werkzeuge zu nutzen müssen diese Einfluss auf die Überprüfung nehmen können

Beispiele:

- Manche Einrichtungen mögen sich für PDF/A-3 als das Format der Wahl für die Bewahrung von Text entscheiden (weil es Container-Elemente erlaubt), andere Einrichtungen entscheiden, dass für diesen Zweck PDF/A-1b (Container-Elemente verboten) zu verwenden ist
- Ein Museum hält es für sehr wichtig, dass in ihren TIFF-Dateien zu jedem Zeiteintrag auch die Zeitzone festgehalten wird (TIFF/EP), ein anderes Museum hält das nicht für wichtig und möchte nur gegen den „baseline standard“ prüfen

Feststellen, ob eine Datei einem Standard entspricht ... das kann nicht alles sein !

Regeln ...

- Kulturerbe-Einrichtungen müssen in der Lage sein gegen ihre eigenen „Policies“ (Interpretationen der Standards) zu prüfen. Deshalb müssen die Werkzeuge solche „Policies“ als Option anbieten (oder müssen in der Lage sein, diese als Option zu speichern)
-
- Generell: Es muss einfach sein für Kulturerbe-Einrichtungen ihre „Regeln“ in das Werkzeug zu integrieren.

Reparateur ...

- In einigen Fällen können die Inhalte von fehlenden oder falsch genutzten TIFF-TAGs aus den Werten anderer TAGs abgeleitet werden ... auf diese Weise kann (bisweilen) die Übereinstimmung mit dem Standard automatisch hergestellt werden. Die Preforma-Werkzeuge haben einen einfachen „metadata fixer“ integriert.

Feststellen, ob eine Datei einem Standard entspricht ... das kann nicht alles sein !

Berichte ...

- Es ist sehr wichtig, dass die Werkzeuge leicht verständliche Berichte und Analysen erstellen. Selbst Nicht-IT-Menschen sollen in der Lage sein zu verstehen worin Probleme bestehen
- Die Berichte sollten prinzipiell in der Sprache der Nutzer verfügbar sein
-
- Die Berichte müssen auch in maschinenlesbarer Form vorliegen, um an andere Programme weitergegeben zu werden, die möglicherweise in der Lage sind weitergehende Korrekturen vorzunehmen

Feststellen, ob eine Datei einem Standard entspricht ... das kann nicht alles sein !

Integration ...

- Die Werkzeuge müssen als Einzelplatz-Version verfügbar sein, sie müssen aber auch gemeinsame Nutzung in Netzwerken erlauben.
-
- Die Werkzeuge müssen leicht in vorhandene Workflows für Langzeitarchivierung eingebaut werden können.

Skalierbarkeit ...

- Die Werkzeuge müssen in der Lage sein sehr kleine wie sehr große Dateien zu prüfen und ebenso kleine oder große Gruppen von Dateien (z.B. Ordner mit 10.000 TIFF-Dateien)

Stand der Entwicklung...

- Die Entwicklung ist fast abgeschlossen. Gründliche Tests haben gezeigt, dass die Werkzeuge gut arbeiten.
- Jeder ist frei, die Entwicklung fortzusetzen, indem er z.B. Validatoren für weitere Formate entwickelt oder die vorhandenen Validatoren noch weiter verbessert.
-
- Die Software ist mehrsprachenfähig aber die Übersetzungen fehlen noch (waren nicht Teil des Projekts)

... take it, use it, improve it, share it ...

!

Zuletzt ein Beispiel aus dem Online-Validator für TIFF-Dateien. Diese Version des TIFF-Werkzeugs hat keine Möglichkeiten eigene Regeln (s.o.) zu definieren ...



DPF Manager

CONFORMANCE CHECKER

File



Configuration

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

Check files



Eine TIFF-Datei ...



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

Der Bericht ...

IFD Tags

Expert mode Default values

File structure

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

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.

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
ℹ️	TAG-281-0005	IFD1	MaxSampleValue Tag is not defined. Then 2*(BitsPerSample) - 1 value is assumed
ℹ️	TAG-280-0005	IFD1	MinSampleValue Tag is not defined. Then 0 value is assumed
ℹ️	TAG-254-0009	IFD1	NewSubfileType Tag is not defined. Then a full resolution, single image, no transparency is assumed
ℹ️	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

Erklärung könnte einfacher sein!

Alle Werkzeuge unter:

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

... take it, use it, improve it, share it ...



Herzlichen Dank !