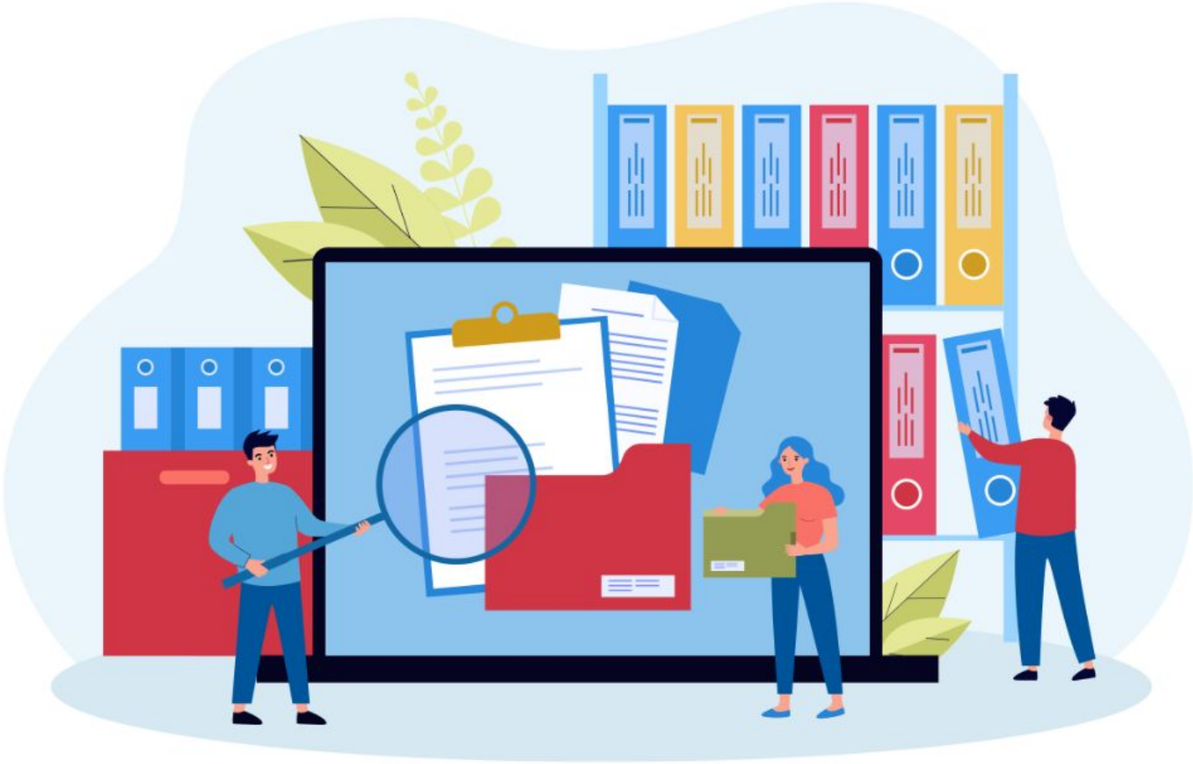
ORPALIS
imaging technologies



PDF
PDF Days Europe 2021

PDF/A Conversion and Validation Challenges

Two similar processes with many differences

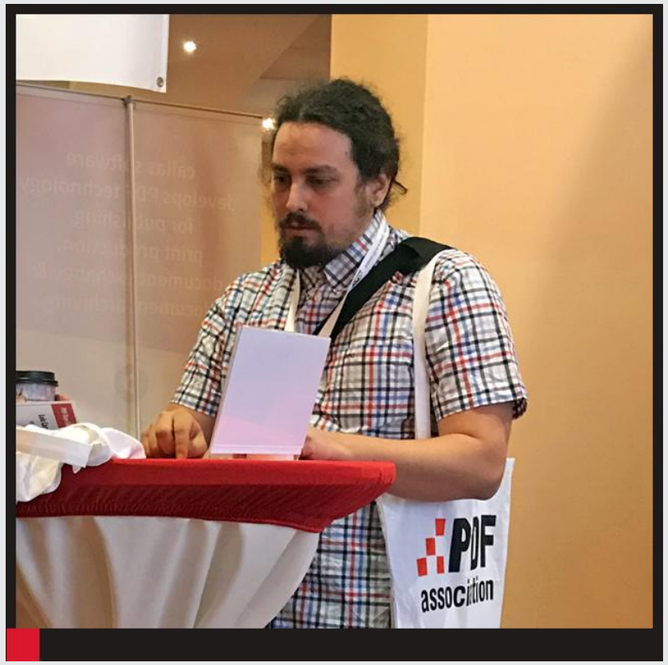
Content

- PDF/A facts
- PDF/A validation
- PDF to PDF/A conversion
- Technical similarities
- Differences
- Examples: Metadata
- Examples: Fonts
- Examples: Transparency
- Examples: File structure
- Examples: Color spaces
- Examples: Annotations
- Implementation limits
- Customers expectations
- The case of PDF/A 4

Speakers

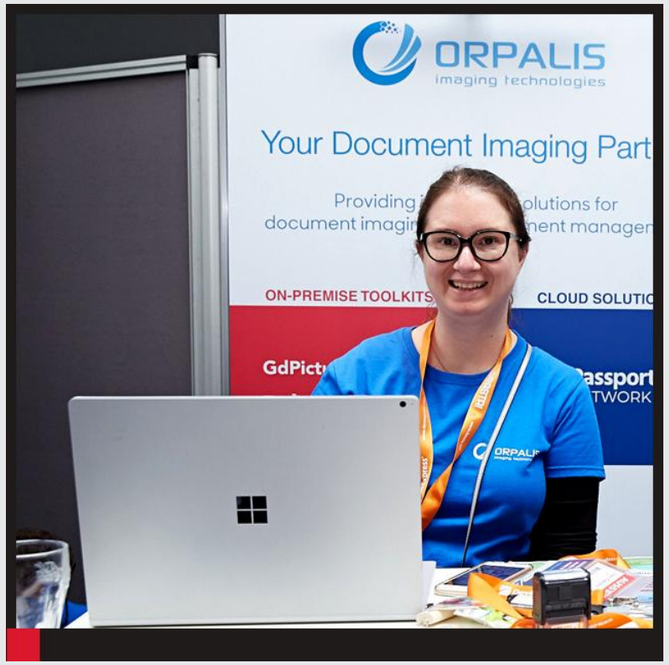
Matúš Pizúr

Senior developer and PDF specialist



Elodie Tellier

Managing Director and PDF Association Board Member



PDF/A Facts

- ISO standard for archiving and long-term preservation of electronic documents.
- Ensures same visual appearance of the document over the course of time.
- Platform & Software independent.

PDF/A-1 ISO 19005-1: 2005

PDF/A-2 ISO 19005-2: 2011

PDF/A-3 ISO 19005-3: 2012

PDF/A-4 ISO 19005-4: 2020

PDF/A Facts

Main goals:

- Contains everything needed to display
- Contains nothing that could negatively impact the display

- All fonts required for rendering must be embedded and meet specification requirements.
- Color information must be given in a platform-independent format using calibrated color profiles.
- Encryption and password-protected access are forbidden.
- All metadata streams present in the PDF shall conform to the XMP Specification.
- Launch, Sound, Movie and certain other actions are forbidden.

History

- PDF/A-1 - published in 2005 - based on PDF 1.4
- PDF/A-2 - published in 2011 - based on ISO 32000-1 (PDF 1.7)
- PDF/A-3 - published in 2012 - PDF/A-2 + extended embedded files support
- PDF/A-4 - published in 2020 - based on ISO 32000-2:2020 (PDF 2.0)

PDF/A Validation

- To claim PDF/A conformance is not enough.
- Validates if the document conforms to the specification.
- It’s a laborious process.

PDF/A Validation

- Provides automated tool to find problems before they cause damage.
- Does not fix the documents that fail the validation process.

PDF to PDF/A Conversion

- Many active documents in the world are not fit for archiving.
- Not always possible to optimize workflows to create PDF/A from the start.

PDF to PDF/A Conversion

- Converts any PDF document to PDF/A compliant document.
- Can be fully automated.
- Ensures PDF/A Validation will succeed.
- Can improve the final quality of the document.

Technical similarities

- Backbone - PDF Document load and parsing.
- Objects, content streams, file structure, and syntax analysis.
- Check if all the features conform to the specification.

Differences

PDF/A Validation is an exact process

- Precisely defined goals
- No ambiguity

PDF to PDF/A Conversion has options, it can be customized

- Rasterizing and postprocessing (MRC, OCR)
- Clone existing file with
- Modify / repair existing file

Examples - Metadata

Validation

- Needs to check not only information but also structure and syntax of the XMP metadata.

Conversion

- Validates and repairs or…
- Has option to “extract” what is possible and recreate 100% valid XMP metadata from scratch.

<?xpacket begin="" id="W5M0MpCehiHzreSzNTczkc9d"?>

<x:xmpmeta xmlns:x="adobe:ns:meta/" x:xmptk="Adobe XMP Core 4.2.1-c041 52.342996, 2008/05/07-20:48:00">

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">

<rdf:Description rdf:about="" xmlns:pdfaid = "http://www.aiim.org/pdfa/ns/id/">

<pdfaid:part>2</pdfaid:part>

<pdfaid:conformance>A</pdfaid:conformance>

</rdf:Description>

…

Examples - Fonts

- Validation
  - Needs to check all font requirements.
  - Logs results.
- Conversion
  - Same as validation and on top of it…
  - Comes up with solutions for correcting the problems.

Examples - Fonts

- Non-embedded fonts
  - “Standard 14 Fonts” still very popular
  - Finds suitable font system.
  - Option to substitute font.
- Inconsistent glyph width information
- Missing glyphs
- Encoding problems
  - Encoding Differences
  - Invalid CMaps

Examples - Transparency (PDF/A-1)

- Validation
  - Checks for any transparency usage in images, groups, xobjects, annotations.
  - Logs and done.
- Conversion
  - Problematic.
  - Softmask to stencil.
  - Rasterizes.
  - Loss of fidelity, and possibly of information.

Examples - File structure

- Validation
  - Checks correct file structure.
  - Header, object definitions, stream definitions, XREF table.
- Conversion
  - Tries to repair or...
  - Doesn’t worry about it and generates the pdf from scratch.

```
%PDF-1.7
%ÄÄÄÄ
1 0 obj
<</Type /Catalog /Pages 3 0 R /Outlines
29 0 R /Metadata 30 0 R>>
endobj
3 0 obj
...
```

Examples - Color spaces

- Validation
  - Checks correct usage of device dependant CSs.
  - Output Intent.
  - Makes sure the calibrated CSs are correctly defined.
- Conversion
  - Needs to do all the same checks and handle problems if possible or...
  - Uses own output intent and transforms colors.
  - Possible loss of information.

Examples - Annotations

- Validation
  - Checks types, correct object definitions, links, etc.
- Conversion
  - Flattens or deletes unsupported annotations.
  - Repairs or re-generates appearances.

Examples - Implementation limits

- Validation
  - Checks all values (numeric, string, streams etc.).
  - Logs data.
- Conversion
  - Same as validation, checks all values.
  - *How to handle invalid values?*
  - Long names, path with points in beyond the limits, invalid bounding boxes, etc.

What did you expect?

- Validation
  - The expectations are mostly the same for all users.
  - Reliable and quick.
- Conversion
  - The expectations might change based on use case.
  - Manual vs automated workflow.
  - Specific output requirements:
    - File size, fidelity, fonts used etc.

PDF/A-4

- PDF/A-4 brings new challenges.
- Still not enough real-world documents.
- Still not widely supported by validators / converters.

- It’s time to use it!

Register for the Solution Day!

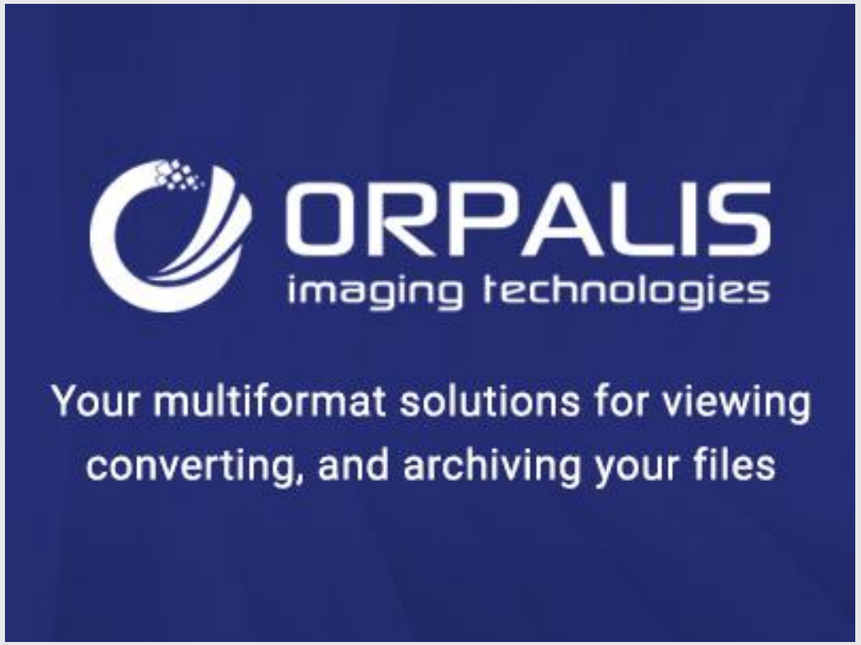
PDF Days Europe 2021

Oct. 19 at 1700 CEST / 1100 ET / 0800 PT

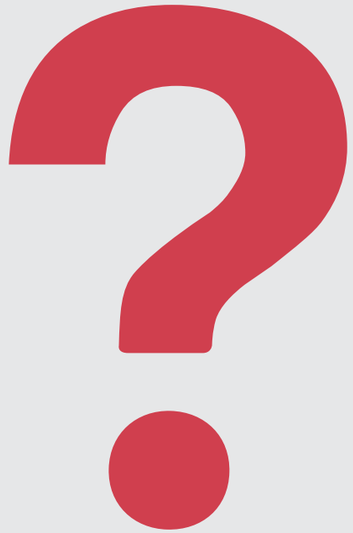
Build **all-in-one** PDF solutions with our toolkits for desktop, Web, and Cloud development

We also provide enterprise software,
low code/no code,
and productivity applications to batch process and automate your PDF workflows.

Do more with your electronic documents!



https://register.gotowebinar.com/register/6485730337622241804



QUESTIONS

Keep in touch with us! m.pizur@orpalis.com / e.tellier@orpalis.com

www.orpalis.com