How Valid Is Your Validation? A Closer Look Behind The Curtain Of JHOVE

Michelle Lindlar (TIB Hannover)
Yvonne Tunnat (ZBW Kiel)

TIB LEIBNIZ-INFORMATIONSZENTRUM TECHNIK UND NATURWISSENSCHAFTEN UNIVERSITÄTSBIBLIOTHEK

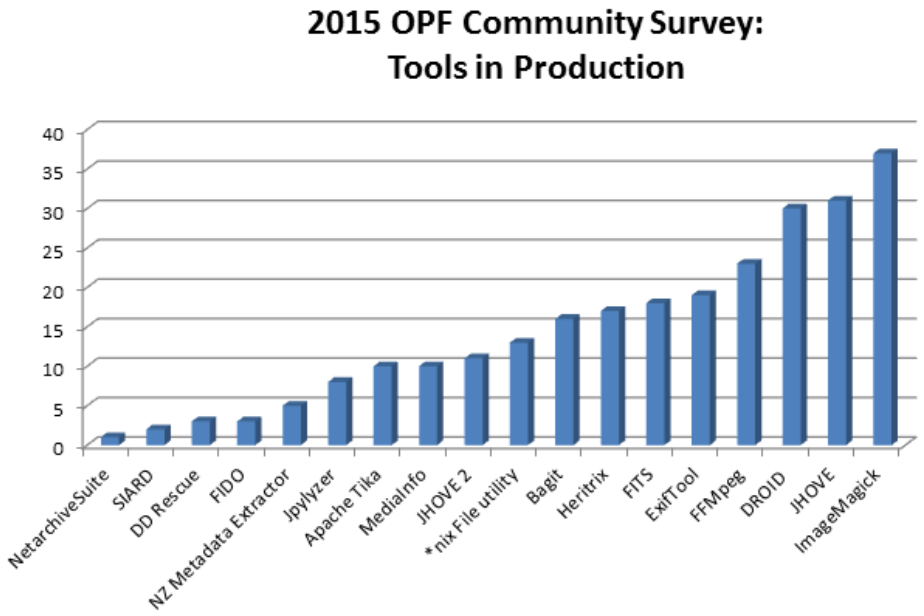
ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Agenda

- Motivation / Approach
- PDF-Module Evaluation
- TIFF-Module Evaluation
- JPEG-Module Evaluation
- Conclusion / Outlook

Motivation

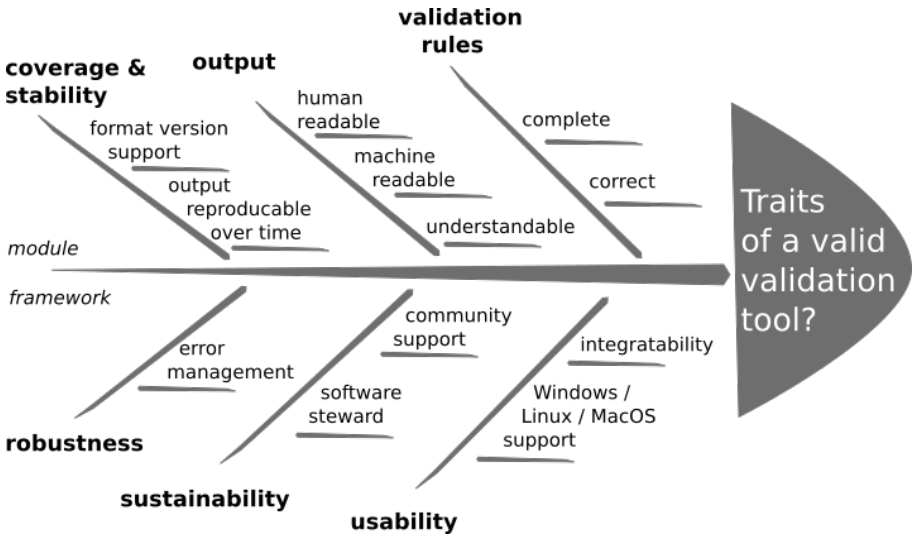
- prefer valid files in our digital archives
- rely on tools for validation
- JHOVE as the go-to validator of the digital preservation community
- but …. can we trust the result?
- … and how can we improve the tool / method?



73% of respondents use JHOVE in production

Approach

- Expectations in a validation tool
- Focusing on module traits



  - Coverage / Stability:
    what is / is not covered ?
  - Output:
    do we understand it ?
  - Validation rules:
    are they complete and correct ?

Approach

- Evaluation methods: alternative tools / standard / error messages
- Scope: PDF-module, TIFF-module, JPEG-module

Overview analysed formats

	No. of pages in specification	No. of possible JHOVE errors
PDF	1310	152
JPG	481[7]	13
TIFF	121	68

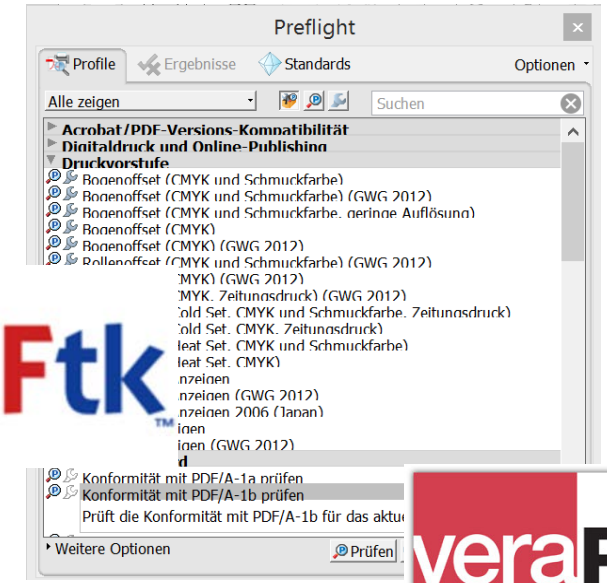
Alternative Tools

PDF Module

callas pdf

PDFtk

PDFBox



veraPDF

… but no validator for „plain“ PDF

Coverage & stability

PDF Module

- PDF versions 1.0-1.6
- Linearized and Tagged PDF (available since PDF 1.4)
- PDF/X-1 (ISO 15930-1:2001), PDF/X-1a (ISO 15930-4:2003), PDF/X-2 (ISO 15930-5:2003), PDF/X-3 (ISO 15930-6:2003)
- PDF/A-1 (ISO/DIS 19005-1)
- Not a profile / version validator, but mainly a structural / syntactical checker

Output

PDF Module

- 90 possible JHOVE validation errors (with ~ 160 rules)
- Not verbose enough:
  e.g. „Invalid structure attribute" – which structure attribute? Why is it invalid?
- Not precise enough:

  "No PDF header" – when header is present, but content is invalid, e.g. *%PDF-2.1*
- Not accessible enough:
  Offset as indicator for error location is not easy to use, object number would be more helpful

Validation Rules

PDF Module

- No tool alternatives = no experiments
- Successful tests against structure violations:
  - incomplete header / trailer
  - manipulated cross-reference table
  - invalid document catalogue
- Unsuccessful tests against page dictionary values – false positives (e.g. rotation values)
- Known false negatives (e.g. unbalanced page tree) – not in violation of standard

Alternative Tools

TIFF-Module

SLUB-digitalpreservation / checkit_tiff

PREFORMA

DPF Manager

LibTIFF

exiftool.exe

Coverage & stability

TIFF Module

JHOVE TIFF module supports:

- Three major version: 4.0, 5.0 & 6.0
- Standardized extensions (TIFF/IT, TIFF/EP, GeoTIFF 1.0)

Output

TIFF-Module

Premature EOF

No TIFF header

No TIFF magic number

TileWidth not defined

Undocumented TIFF tag

No IFD in file (IFD: Image File Directory)

More than 50 IFDs in chain, probably an infinite loop

Validation Rules

TIFF-Module

Validation-Quality-Check: TIFF image test suite

(166 files, 83 of them really damaged)

	JHOVE	ImageMagick	ExifTool	DPF Manager (Extended TIFF)	LibTiff
Valid (% of all 166 files)	17,5	11	34	9	13
Valid (% of 83 non-renderable files)	2	0	7	0	0

Validation Rules

TIFF-Module

Validation-Quality-Check: TIFF image test suite

- two false positives (false alarm)
- (most likely) two false negatives (not renderable in common viewers + most other tools report invalidity or errors)

- **Conclusion:** decent quality, high percentage of invalidity detection + not too picky

Alternative Tools

JPEG Module

Coverage & stability

JPEG Module

The JHOVE JPEG module supports:

- JPEG (ISO/IEC 10918-1:1994)
- JFIF 1.02 (JPEG File Interchange Format)
- Exif 2.0, 2.1 (JEIDA-49-1998) 2.1, and 2.2 (JEITA CP-3451)
- SPIFF (ISO/IEC 10918-3:1997)
- JTIP (ISO/IEC 10918-3:1997)
- JPEG-LS (ISO/IEC 14495)31

Output

JPEG Module

Self-explanatory

- Unexpected end of file
- Invalid JPEG header
- …

cryptic

- Marker not valid in context
- DTT segment without previous DTI
- …

Valiation Rules

JPEG Module

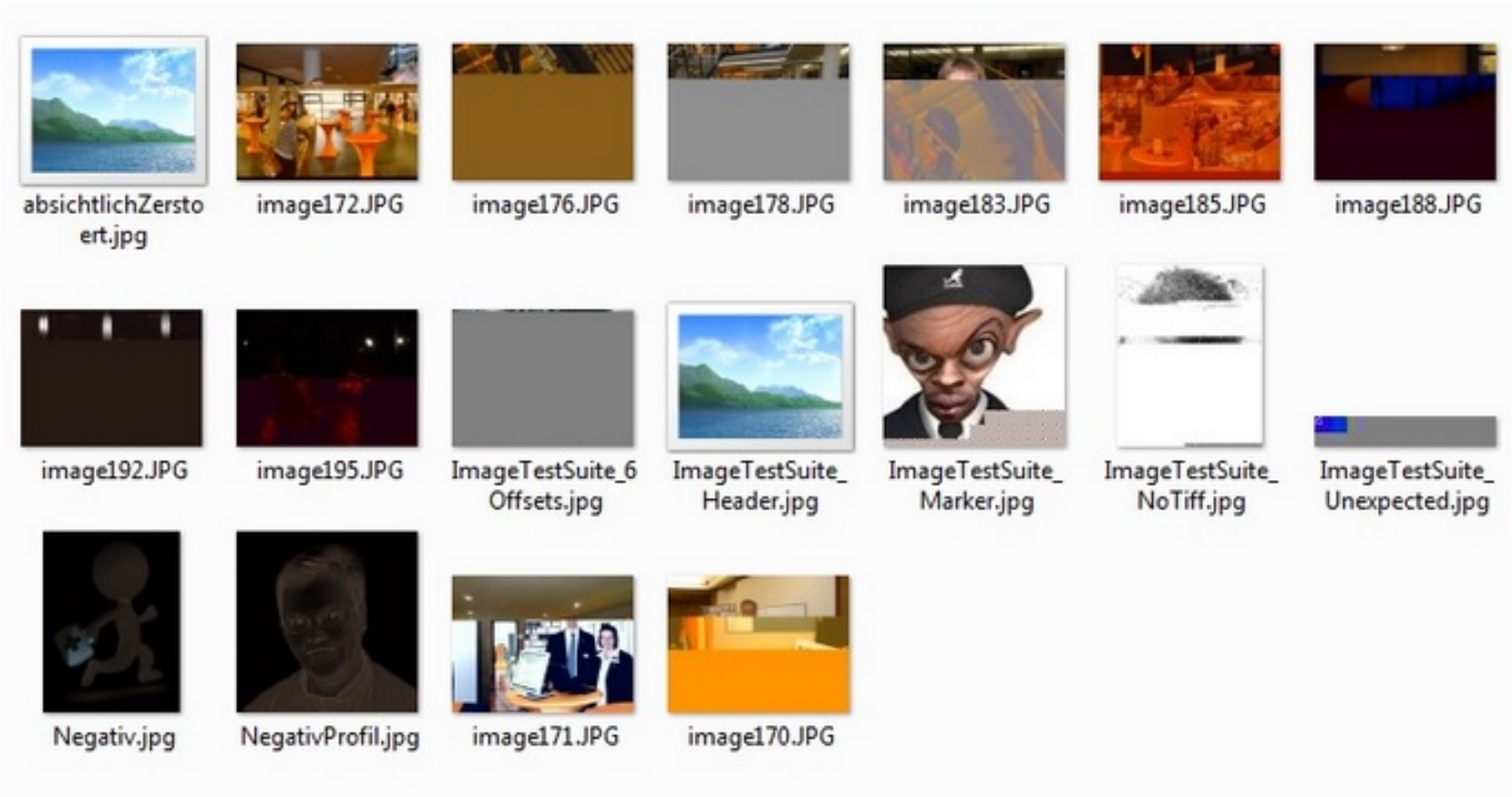
Validation-Quality-Check: JPEG image test suite

(98 files, 39 not renderable)

	JHOVE	ImageMagick	ExifTool	Bad Peggy
Valid (% of all 98 files)	11	6	36	4
Valid (% of 39 non-renderable files)	17	0	43	0

Valiation Rules

JPEG Module



Conclusion

So, how valid is your validation?

- PDF-Module: not ideal, but currently no alternative; update needed, e. g. support of PDF 1.7
- TIFF-Module: decent
- JPEG-Module: not the best choice yet

Community support needed

To improve JHOVE validation quality:

- Checking existing rules against file standards
- Develop new validation rules

Starting actions:

- JHOVE Hack day
- OPF Document Interest Group
- OPF Software Supporter /

To come: „Donate for JHOVE“

TIB LEIBNIZ-INFORMATIONSZENTRUM TECHNIK UND NATURWISSENSCHAFTEN UNIVERSITÄTSBIBLIOTHEK

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Questions? Comments?

Michelle Lindlar,
Leibniz Information Centre for Science and Technology (TIB), Hannover
michelle.lindlar@tib.eu
Twitter: @MickyLindlar

Yvonne Tunnat,
Leibniz Information Centre for Economics (ZBW), Kiel
y.tunnat@zbw.eu
Twitter: @YvonneFriese