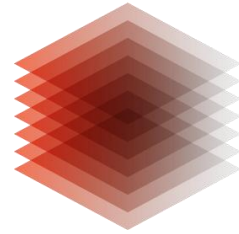

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK



TIB

How valid is your validation?

JHOVE as the go-to validator within
Rosetta

Michelle Lindlar
Sheffield, May 13th 2017
Rosetta Advisory Group Meeting

Agenda

Motivation

Traits of valid validation tool

**Benchmarking Approach
for TIFF
for JPEG**

**Synthetic Test File Approach
for PDF**

Implication on Rosetta and Outlook

Motivation - Validation vs. Identification

%PDF-1.4
%%EOF



Resource	Extens...	Size	Last ...	Ids	Format	Version	Mime ...	PID	Method	Hash
C:\Files\0...	pdf	15 bytes	27.03.1...		Acrobat...	1.4	applicati...	fmt/18	Signature	



Adobe Acrobat

There was an error opening this document. The file is damaged and could not be repaired.

OK



JHOVE

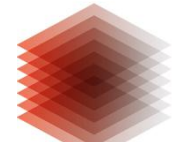
Documents

- C:\Files\08_Vortraege\2017\IPRES\testfiles\Body\ALLE\minir
- Module
- ReplInfo
 - URI: C:\Files\08_Vortraege\2017\IPRES
 - LastModified: Mon Mar 27 13:27:12 CE
 - Size: 15
 - Format: PDF
 - Status: Not well-formed
 - SignatureMatches
- Messages
- MimeType: application/pdf

	File Name	PID	Status	Problems				
1	minimal_test.pdf	FL111191		1 Tasks failed	Download	Replace	Recheck	More...

1 - 1 of 1 Records

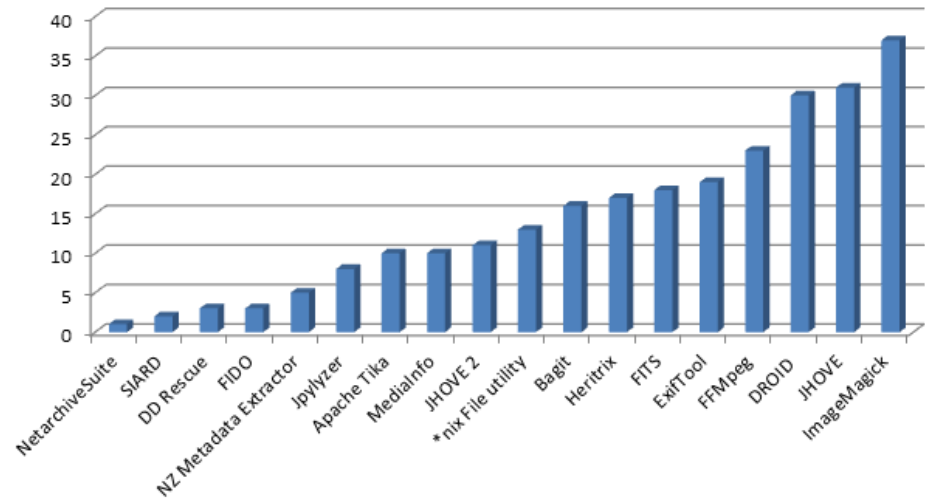
Fixity Check:Pass
Virus Check:Pass
File Format:Pass
Technical MD Extract:Fail - Error/s returned during metadata extraction (Missing startxref keyword or value,Failed to retrieve extractor properties) Agent: JHOVE , PDF-hul 1.7 , Plugin Version 3.0
Risk Analysis:Pass



Motivation – why JHOVE ?

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

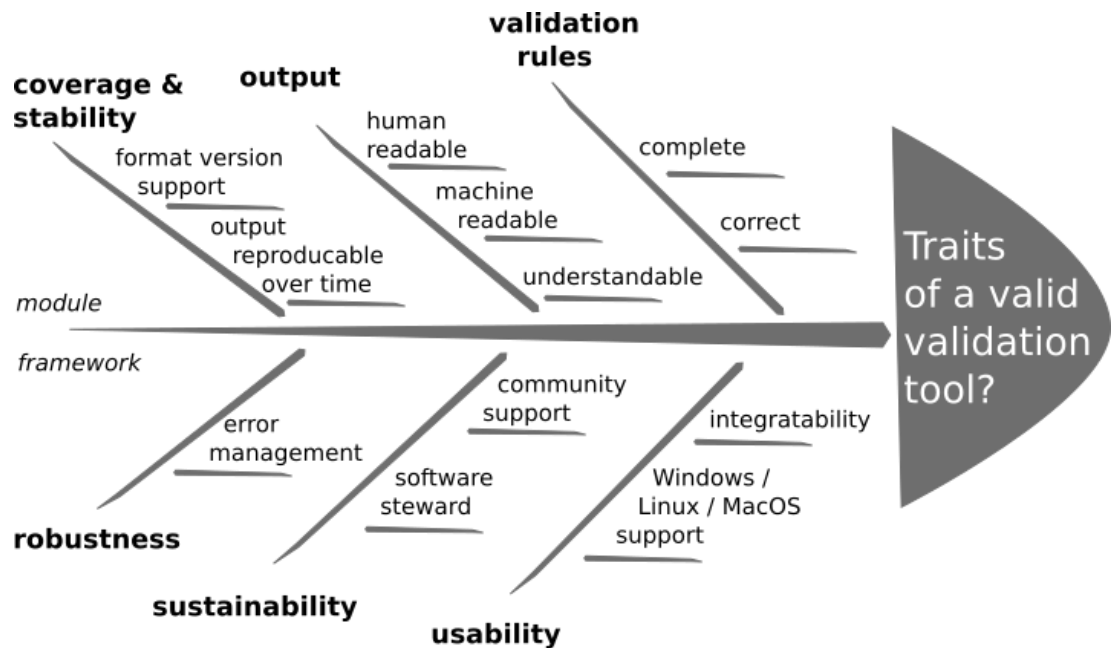
2015 OPF Community Survey:
Tools in Production



*73% of respondents (n=132)
use JHOVE
in production*

Approach – Traits of a valid validation tool

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



Coverage – are all versions covered ?

PDF module

- Not a profile / version validator, but mainly a structural / syntactical checker (also, recent profiles like PDF/A-2,3, PDF/E missing)
- Not covered: PDF 1.7

JPEG module

- Covers most JPEG format versions
- Not covered: JPEG2000 (but, that's a different format and there's a different module for it)

TIFF module

- Covers major versions and some standardized extensions
- Not covered: Extensions/Versions such as BigTIFF

Validation rules – completeness

	No. of pages in specification	No. of possible JHOVE Errors	Lines of code in JHOVE module
PDF	1 310	152	10 581
JPEG	481	13	895
TIFF	121	68	14 457

Checking completeness can be achieved via:

- deriving all shall / should clauses from the standards
- finding / creating test objects for each clause

Problems:

- pre-requisite: clear and formalized standard
- labor intensive task

Validation rules - correctness

Approach 1: Benchmarking

TIFF



JPEG



PDF



Approach 2: Synthetic file creation

Benchmark: TIFF – JHOVE vs. DPF Manager

Test corpus:

made up of Google Image Test Suite

(<https://code.google.com/archive/p/imagetestsuite/>)

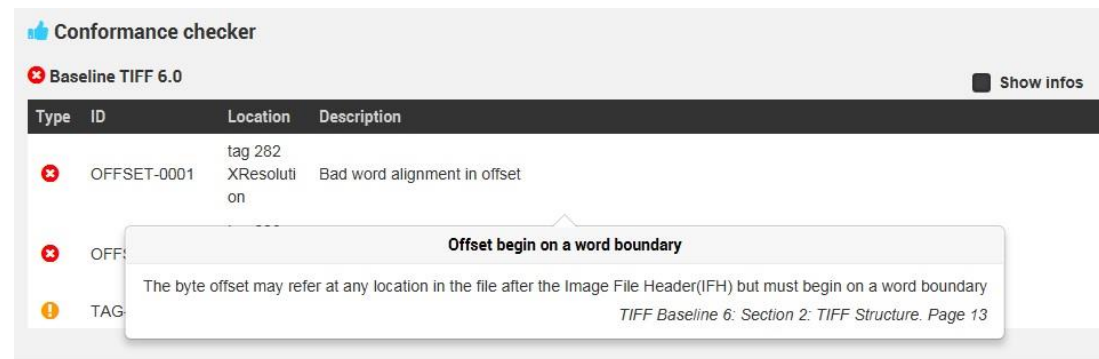
Results for Google Image Test Suite JHOVE vs.



non-renderable files only:

81 files → agree

2 files → disagree (both declared as well-formed and valid by JHOVE)



Conformance checker

Baseline TIFF 6.0 Show infos

Type	ID	Location	Description
✖	OFFSET-0001	tag 282 XResoluti on	Bad word alignment in offset
✖	OFFS		Offset begin on a word boundary
!	TAG		The byte offset may refer at any location in the file after the Image File Header(IFH) but must begin on a word boundary

TIFF Baseline 6: Section 2: TIFF Structure. Page 13

Benchmark: JPEG – JHOVE vs. Bad Peggy

Test corpus:

made up of Google Image Test Suite

(<https://code.google.com/archive/p/imagetestsuite/>)

and collected broken examples (mostly from colleagues)

Results for Google Image Test Suite JHOVE vs. Bad Peggy

89 files → agree

8 files → disagree

(7 of those missed
by JHOVE, i.e. declared
as well-formed and valid)



image170.JPG



image171.JPG



image172.JPG

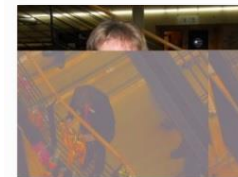


image183.JPG



image185.JPG

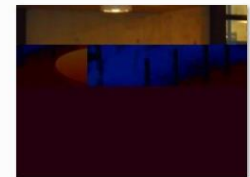


image188.JPG

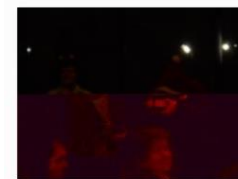
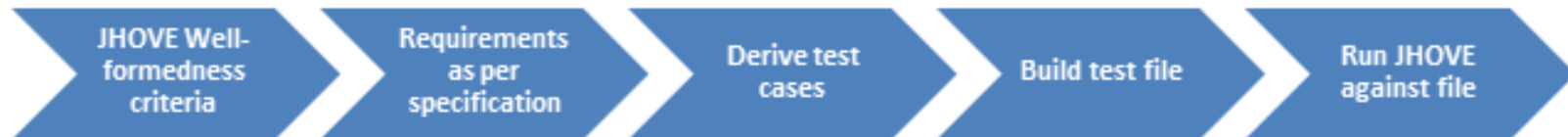


image195.JPG

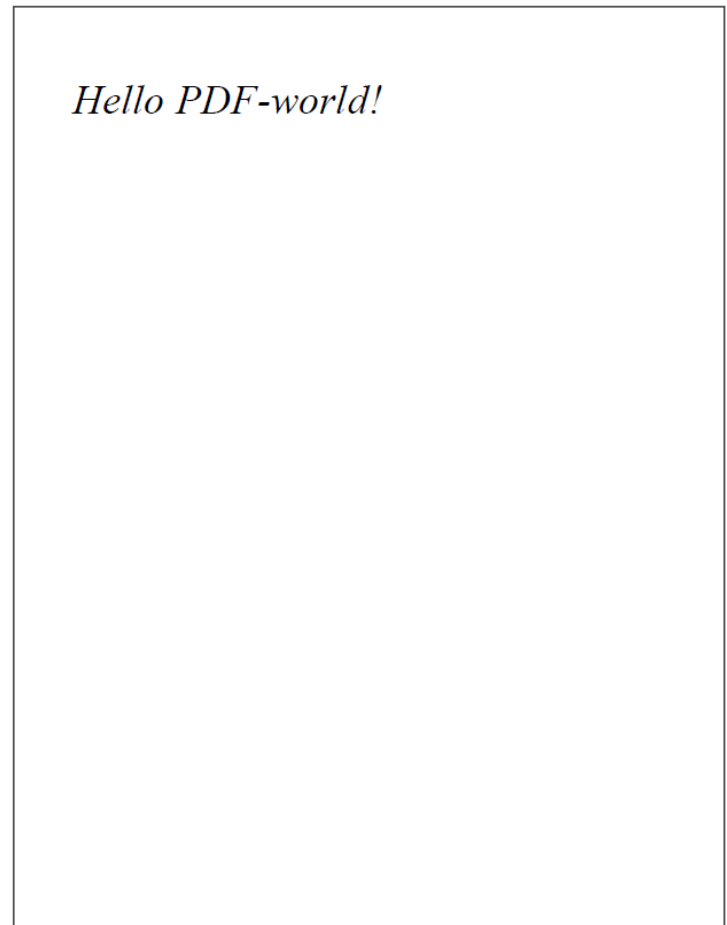
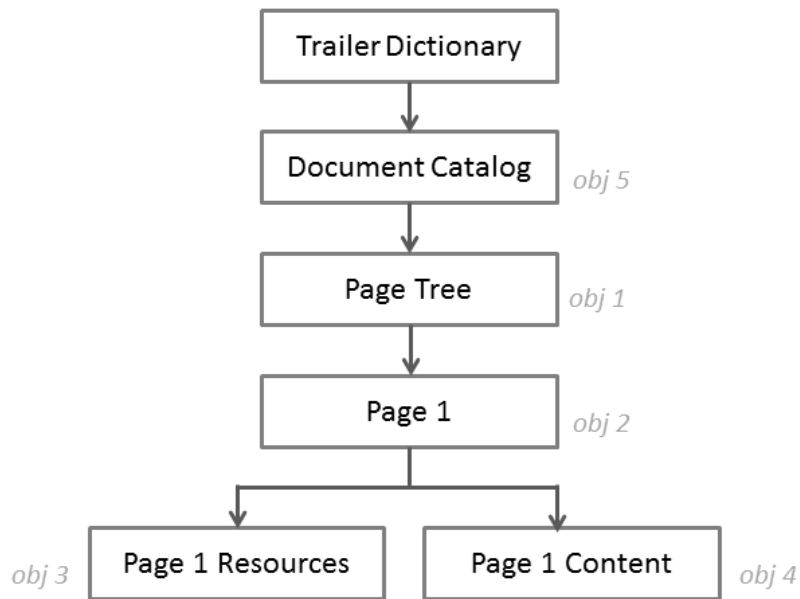
Validating validation via synthetic test files



“In general, a file is well-formed if

- it has a **header** : %PDF-*m.n*,
- a **body** consisting of well-formed objects;
- a **cross-reference** table;
- and a **trailer** defining the
 - cross-reference table size,
 - and an indirect reference to the document catalog dictionary,
 - and ending with: %%EOF”

Building synthetic test files – the „parent file“



From JHOVE condition to test file

“a **body** consisting of well-formed objects”



Table 28 – Entries in the catalog dictionary

Key	Type	Value
Type	name	<i>(Required)</i> The type of PDF object that this dictionary describes; shall be Catalog for the catalog dictionary.

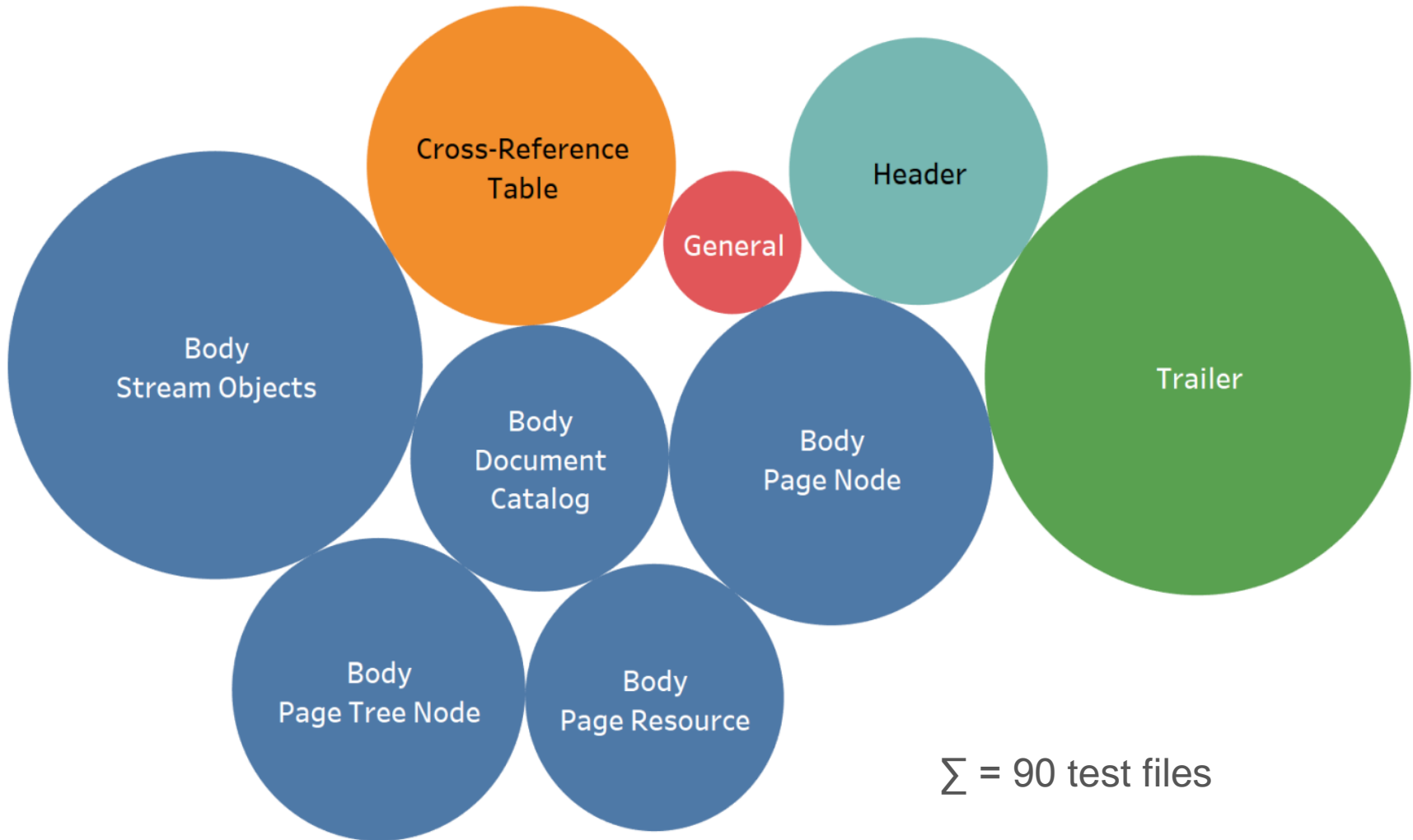


```
5 0 obj
<<
/Pages 1 0 R
/Type /Catalog
>>
```

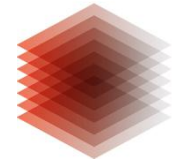
T02-01_005_document-catalog-type-key-missing.pdf

T02-01_006_document-catalog-wrong-type-key.pdf

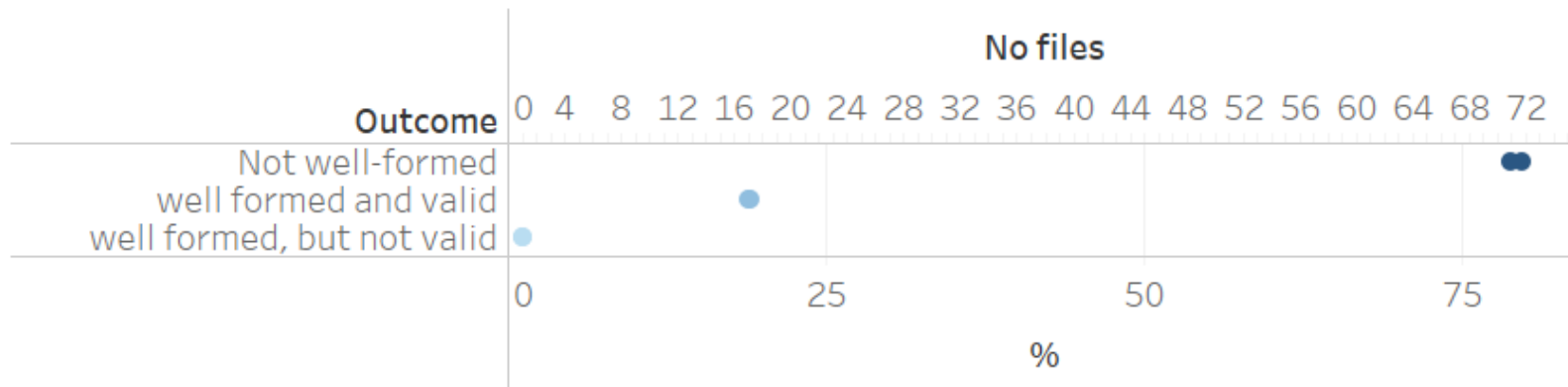
Test corpus - Content



$\Sigma = 90$ test files



Test set results



Good news:

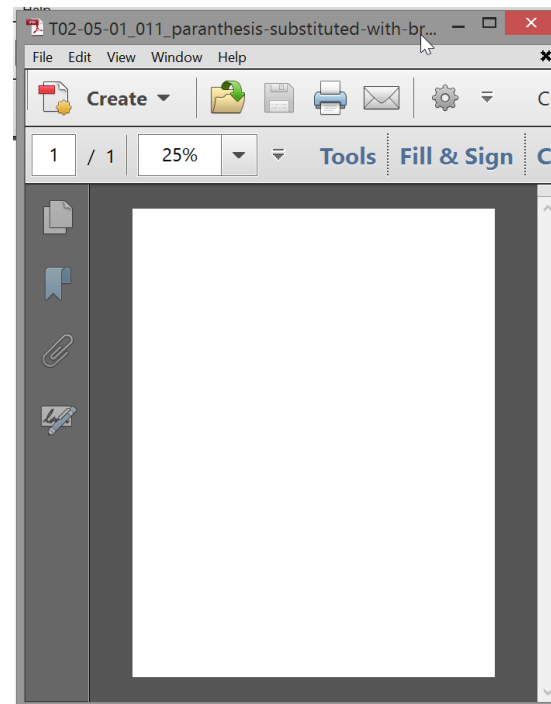
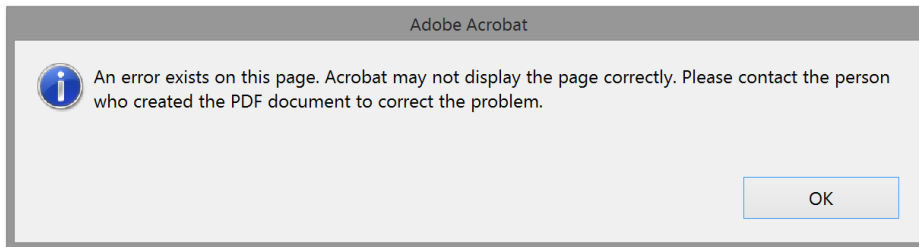
- Majority of testcases (71 files / 80%) were validated correctly

Bad news:

- 18 files were not validated correctly, 17 of those (=19.1%) were considered well-formed and valid

... in case you're still thinking „so what?“

- 2 test cases were considered well-formed and valid but couldn't be rendered by Adobe or other PDF rendering software
- And then there were well-formed cases like these:



Implications on Validation in Rosetta

False positives (i.e., „not well-formed“, when it is)

→ detectable by investigating files caught in validation stack

1 - 1 of 1 Records

<input type="checkbox"/>	File Name	PID <small>▲</small>	Status	Problems				
1	<input type="checkbox"/> minimal_test.pdf	FL111191		Tasks failed ▲	Download	Replace	Recheck	More... ▼

Fixity Check:**Pass**
Virus Check:**Pass**
File Format:**Pass**
Technical MD Extract:**Fail** - Error/s returned during metadata extraction (Missing startxref keyword or value,Failed to retrieve extractor properties) **Agent:**
JHOVE , PDF-hul 1.7 , Plugin Version 3.0
Risk Analysis:**Pass**

False negatives (i.e. „well-formed“, when it isn't) go straight to Permanent

→ how to detect?

Need multiple tools to find the truth

→ Currently work done pre-ingest

From „so what“ to „so now, what“

A call to arms for validation/JHOVE

- No one said digital preservation was easy.
- This is especially true for file formats.
- We, as a community need to take responsibility for the (community owned) **processes** and **tools** we use.
- Question tool output ! Get involved !

... and a suggestion for Rosetta:

- Allow for multiple tools to be run against each other in the validation stack (currently work being done pre-ingest)
- Clearer distinction between validation and technical metadata extraction in plugin type and error mapping

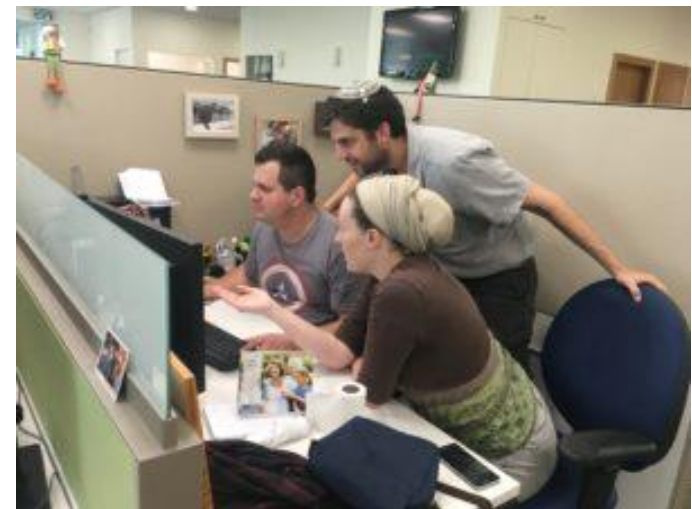
How can you help?

- Become a [JHOVE software supporter or OPF member](#)
- Make a donation to support JHOVE development:



- Contribute to improving the software or documentation

see <http://openpreservation.org/technology/products/jhove/>



Further information

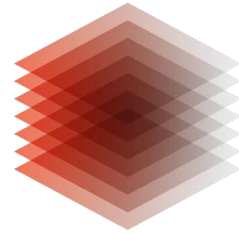
Lindlar, Tunnat: „How valid is your validation? A closer look behind the curtain of JHOVE“ IDCC 2017 paper

Lindlar, Tunnat, Wilson: „A Test-Set for Well-Formedness Validation in JHOVE – The Good, the Bad and the Ugly“ iPRES 2017 paper (forthcoming)

Yvonne Tunnat: „TIFF format validation: easy-peasy ?“ OPF blog <http://openpreservation.org/blog/2017/01/17/tiff-format-validation-easy-peasy/>

Yvonne Tunnat: „Error detection of JPEG files with JHOVE and Bad Peggy – so who’s the real Sherlock Holmes here?“ OPF blog <http://openpreservation.org/blog/2016/11/29/jpegvalidation/>

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK



TIB

Questions? Comments!

contact

M. Lindlar

Twitter @mickylindlar

T 0511 762-19826, michelle.lindlar@tib.eu