ETH BIBLIOTHEK

Handling File Format Issues in Rosetta – and Beyond

Rosetta Advisory Group 8th Annual Meeting
Sheffield, June 13, 2017

Dr. Franziska Geisser, ETH Library, ETH Zurich

Winner of Best Poster at iPRES 2016

iPRES 2016
Bern // October 3–6, 2016

17 To Act or Not to Act

Mathias Töwe, Franziska Geisser, Dr. Roland Suri

Barbara Signori
iPRES 2016 General Chair

nestor

nestor
iPRES 2016 Best Poster Sponsor

ETH zürich

ETH BIBLIOTHEK

To Act or Not to Act – Handling File Format Identification and Validation Issues in Practice

Matthias Töwe, Franziska Geisser, Roland E. Suri
ETH-Bibliothek, ETH Zurich – www.library.ethz.ch/Digital-Curation, data-archive@library.ethz.ch

To act …

Identification issues with Text Files or Delusive Extensions

Research Data often comes in text files. But the supposedly "simple" text format poses problems with regard to exact format identification.

- Plain text: multiple matches in our system, but single hit in DROID

- Extension .dat: turns out to be CSV format!

➡ In our Rosetta-based preservation system, we have implemented customized rules that assign the correct PRONOM PUID.

TIFF Validation: Invalid DateTime Separator

A large number of TIFF files that we want to archive have wrong punctuation marks in the DateTime Tag:

JHOVE TIFF-hul throws an error message, file is not valid:

These TIFF files are well-formed and perfectly readable – BUT:

- DateTime format is not compatible with baseline TIFF specification
- We do not know what problems could arise in the future

➡ Therefore we check all TIFF files prior to ingest and correct the error with ExifTool in an automated process.

… or not to act

XML: Multiple Identification due to Missing XML Declaration

Thousands of XML files archived in our preservation system lack an XML declaration. According to the XML specification, the XML declaration is optional for v. 1.0. But in PRONOM, the internal signature for fmt/101 is exclusively based on the XML declaration! Therefore DROID is unable to identify these files.

➡ In our preservation system, we have implemented a rule that automatically identifies these XML files as fmt/101 for this known data producer.

XML Validation: Invalid METS Schema Location

In some older METS files, the syntax of the xsi:schemaLocation attribute is not correct.

In the meantime, the data producer has acknowledged the error and has delivered correct "delta" METS files, that will supersede the faulty "master" METS files once they will be exported from the preservation system.

➡ Thus we can happily archive the invalid "master" METS files.

PDF Validation: Unexplained JHOVE Errors

JHOVE PDF-hul Module reports lots of errors for PDF formats with unclear consequences for preservation.*) We continuously analyse these errors and to date found only two of them to be useful:

- **Invalid PDF trailer:** file may have been damaged during transmission
- **Failed to retrieve extractor properties:** file may be encrypted

*) See http://wiki.opf-labs.org/display/Documents/JHOVE+issues+and+error+messages by Yvonne Friese

Most of the errors we currently do not regard as a risk for preservation.

➡ Therefore we have configured a rule that ignores the "irrelevant" errors and allows the files to pass through.

XML Validation: Vendor-specific Attributes

Many of the METS files that we want to archive in our preservation system are invalid due to a vendor-specific attribute that is not specified in the METS schema.

The software vendor has been informed of the problem, but has taken no action.

➡ These attributes are not relevant for a reconstitution of the metadata in the source system. Thus we currently ignore this error.

… and revisit former (in)actions!

TIFF: Uncommon Compression Scheme and Colour Space

According to our digitization standards, TIFFs should be uncompressed and in RGB colour space. But a few have ISO JPEG compression and YCbCr colour space. They were not caught during ingest, because JHOVE didn't care. But according to a validation tool supported by an expert study*), these properties are "not advisable", and a combination of the two should be avoided. We plan to migrate these files.

*) See the KOST recommendations and preservation planning study (http://kost-ceco.ch/cms/index.php?preservation_tiff_de – german/french only) and the KOST Validation Tool (http://kost-ceco.ch/cms/index.php?id=240.495.0.0.1.0). KOST is the Swiss Coordinating Office for Long-term Archiving of Electronic Documents, a cooperative venture of Swiss archives.

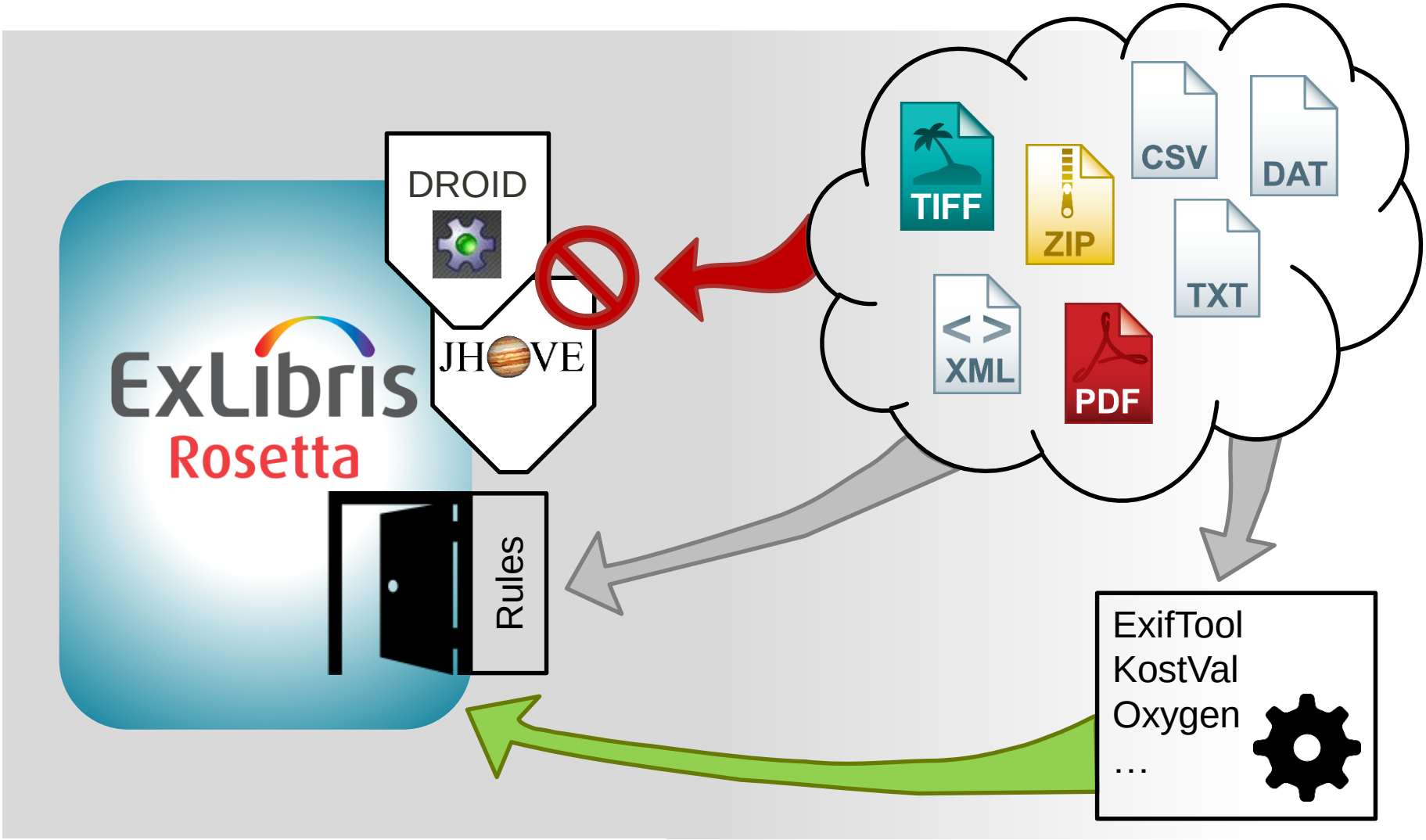
➡ Sometimes we have to postpone preservation actions in the hope of getting better tools one day, or we have to revise former actions in the light of new findings.

File format icons Copyright (c) 2008 Yamdex Technologies, S.L.

To act …

… or not to act

… and revisit former (in)actions!


DROID
JHOVE
ExLibris
Rosetta
Rules
TIFF
ZIP
CSV
DAT
TXT
XML
PDF
ExifTool
KostVal
Oxygen
…

«ETH Data Archive»

- Operated by ETH Library as an infrastructure for the medium to long-term preservation of digital data such as
  - Library materials (master files of digitized books and images)
  - Scholarly publications and research data produced at ETH Zurich
  - Administrative records and bequests from University Archives of ETH
  - Websites
- Operational since 2012
- May 2017: 105 TB, 104’000 IEs, 5.85 million files
- Mostly dark archive
- Mostly automated uploads

Format Identification and Validation Issues

- TIFF
- XML
- PDF
- Research Data

TIFF:
adhere to the spec

TIFF: invalid DateTime separator

JHOVE

agent	JHOVE , TIFF-hul 1.7 , Plugin Version 3.0
plugin name	TIFF-hul-1.10
status	-
format	TIFF
version	6.0
mime Type	image/tiff
profile	-
is Valid	false
is Well Formed	true
error Message	Invalid DateTime separator: 2011.03.22

DateTime

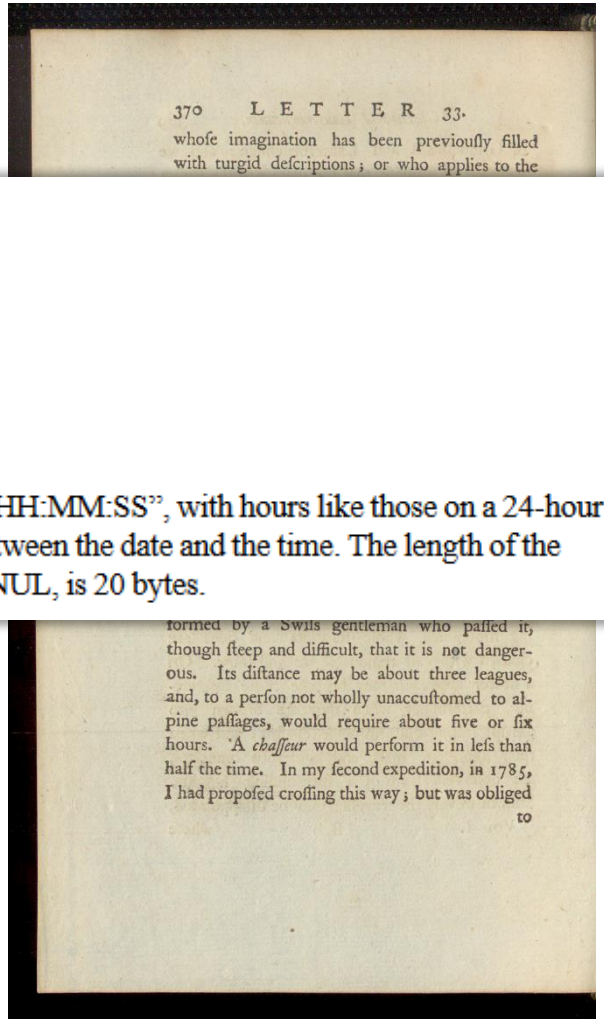
Date and time of image creation.

Tag = 306 (132.H)

Type = ASCII

N = 20

The format is: "YYYY:MM:DD HH:MM:SS", with hours like those on a 24-hour clock, and one space character between the date and the time. The length of the string, including the terminating NUL, is 20 bytes.



Corrected with ExifTool

TIFF: compression and wrong colour space

JHOVE

is Valid	true
is Well Formed	true

35	Type=nisoImage.compressionScheme, Value=ISO JPEG
36	Type=nisoImage.dateTimeCreated, Value=2016-02-04T14:38:33
37	Type=nisoImage.imageLength, Value=5143
38	Type=nisoImage.xSamplingFrequency, Value=600/1
39	Type=nisoImage.yCbCrCoeffiecients, Value=[299/1000, 587/1000, 114/1000]
40	Type=nisoImage.scanningSoftware, Value=ImageGear Version: 16.00.000
41	Type=nisoImage.stripByteCounts, Value=[15881969]
42	Type=nisoImage.stripOffsets, Value=[669464]
43	Type=nisoImage.colourSpace, Value=YCbCr

KOST-Val

Format (PDF/A, TIFF, JP2, SIARD, JPEG):

Invalid:

Validation: TIFF -> C:\temp\14712938.tif	
C) compression	Compression: JPEG is not allowed.
D) color space	PhotometricInterpretation: YCbCr is not allowed.

Convert and replace

XML:
anything goes?

XML: invalid attributes

```
- <mets:fileSec>
  - <mets:fileGrp USE="master image">
    - <mets:file ID="IMG6484620" MIMETYPE="image/tiff" vls:file-omitted="true" CREATED="2013-11-14T18:11:50Z" SIZE="16542200">
        <mets:FLocat LOCTYPE="URL" xlink:href="image/6484620.tif"/>
      </mets:file>
    - <mets:file ID="IMG6484621" MIMETYPE="image/tiff" vls:file-omitted="true" CREATED="2013-11-14T18:11:53Z" SIZE="150417866">
        <mets:FLocat LOCTYPE="URL" xlink:href="image/6484621.tif"/>
      </mets:file>
    </mets:fileGrp>
  </mets:fileSec>
```

error Message	cvc-complex-type.3.2.2: Attribute 'vls:file-omitted' is not allowed to appear in element 'mets:file'.,cvc-complex-type.3.2.2: Attribute 'vls:file-omitted' is not allowed to appear in element 'mets:file'.

JHOVE

```
- <mix:ScannerCapture>
  - <mix:ScannerModel>
      <mix:scanningModelName>Microbox GmbH - Book2Net</mix:scanningModelName>
    </mix:ScannerModel>
```

Ignore MD Extraction Error

error Message	cvc-complex-type.2.4.a: Invalid content was found starting with element 'mix:scanningModelName'. One of '{"http://www.loc.gov/mix/v20":scannerModelName; "http://www.loc.gov/mix/v20":scannerModelNumber; "http://www.loc.gov/mix/v20":scannerModelSerialNo}' is expected.,cvc-complex-type.2.4.a: Invalid content was found starting with element 'mix:scanningModelName'. One of '{"http://www.loc.gov/mix/v20":scannerModelName; "http://www.loc.gov/mix/v20":scannerModelNumber; "http://www.loc.gov/mix/v20":scannerModelSerialNo}' is expected.,cvc-complex-type.2.4.a: Invalid content was found starting with element 'mix:scanningModelName'. One of '{"http://www.loc.gov/mix/v20":scannerModelName; "http://www.loc.gov

JHOVE

ETH zürich

XML: when the tool is stricter than the spec (1)

DROID

File Format:**Fail** - Multiple formats found for file (vs_Error.5)

Format ID	Format Mime Type	Format Name
fmt/101	text/xml	fmt/101
fmt/120	text/xml	fmt/120
fmt/121	text/xml	fmt/121
fmt/896	application/vnd.recordare.musi	fmt/896
unknown		

```
WARNING: no match; possibilities
based on extension are fmt/101,
fmt/121, fmt/120, fmt/896, fmt/1011
```

<oXygen/> XML Editor

Document is valid.

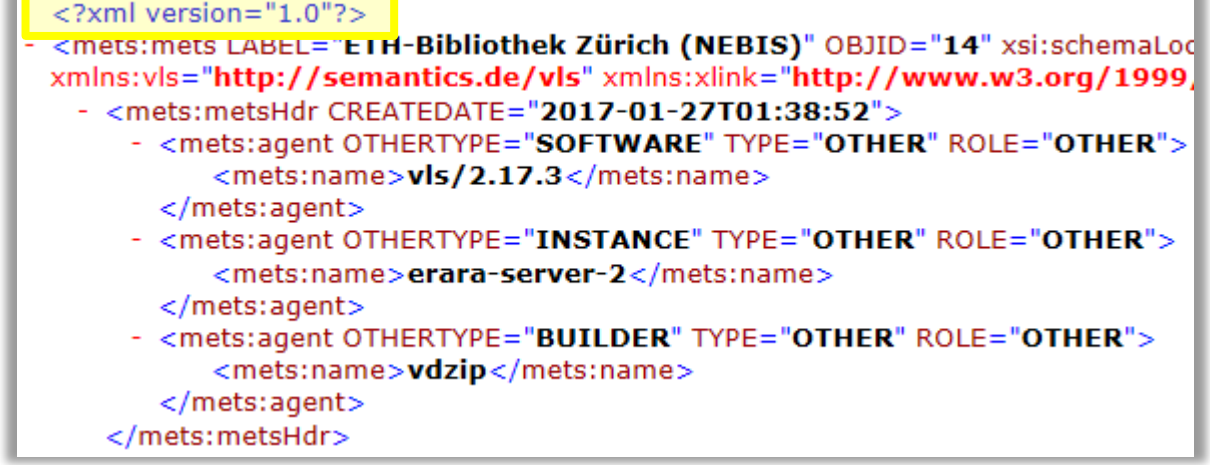
Signatures

The National Archives

External signatures	File extension: xml	
Internal signatures	**Name**	XML 1.0
	Description	XML declaration and version. Signature updated November 2011 to allow an optional 3 byte undisclosed beginning of file offset to allow for UTF-8 byte order mark (BOM) that may precede the XML declaration ([<?xml version='])
	Byte sequences	**Position type** Absolute from BOF
		Offset 0

XML: when the tool is stricter than the spec (2)

```
<?xml version="1.0"?>
- <mets:mets LABEL="ETH-Bibliothek Zürich (NEBIS)" OBJID="14" xsi:schemaLoc
  xmlns:vls="http://semantics.de/vls" xmlns:xlink="http://www.w3.org/1999/
    - <mets:metsHdr CREATEDATE="2017-01-27T01:38:52">
        - <mets:agent OTHERTYPE="SOFTWARE" TYPE="OTHER" ROLE="OTHER">
            <mets:name>vls/2.17.3</mets:name>
          </mets:agent>
        - <mets:agent OTHERTYPE="INSTANCE" TYPE="OTHER" ROLE="OTHER">
            <mets:name>erara-server-2</mets:name>
          </mets:agent>
        - <mets:agent OTHERTYPE="BUILDER" TYPE="OTHER" ROLE="OTHER">
            <mets:name>vdzip</mets:name>
          </mets:agent>
      </mets:metsHdr>
```



Format Id. Rule: fmt/101

```
1 <mets:mets xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:mets="http://www.loc.gov/METS/"
2     <mets:metsHdr CREATEDATE="2017-01-27T01:38:52"><mets:agent ROLE="OTHER" TYPE="OTHER" OTHERTYPE="SO
3                 (Revision 1.106 2014/12/19)</recordOrigin></recordInfo><extension><vl:id>11583946</vl:
4                 (Revision 1.106 2014/12/19)</recordOrigin></recordInfo></mods></mets:xmlData></mets:md
5                 (Revision 1.106 2014/12/19)</recordOrigin></recordInfo></mods></mets:xmlData></mets:md
```

2.8 Prolog and Document Type Declaration

[Definition: XML documents SHOULD begin with an **XML declaration** which specifies the version of XML being used.]

```
<?xml version="1.0"?>
<greeting>Hello, world!</greeting>
```

XML: wrong syntax of METS Schema Location (1)

JHOVE

is Valid	false
is Well Formed	true
error Message	cvc-elt.1.a: Cannot find the declaration of element 'mets:mets'.

Wrong schema location:

```
xsi:schemaLocation="http://www.loc.gov/standards/mets/version16/mets.xsd"
```

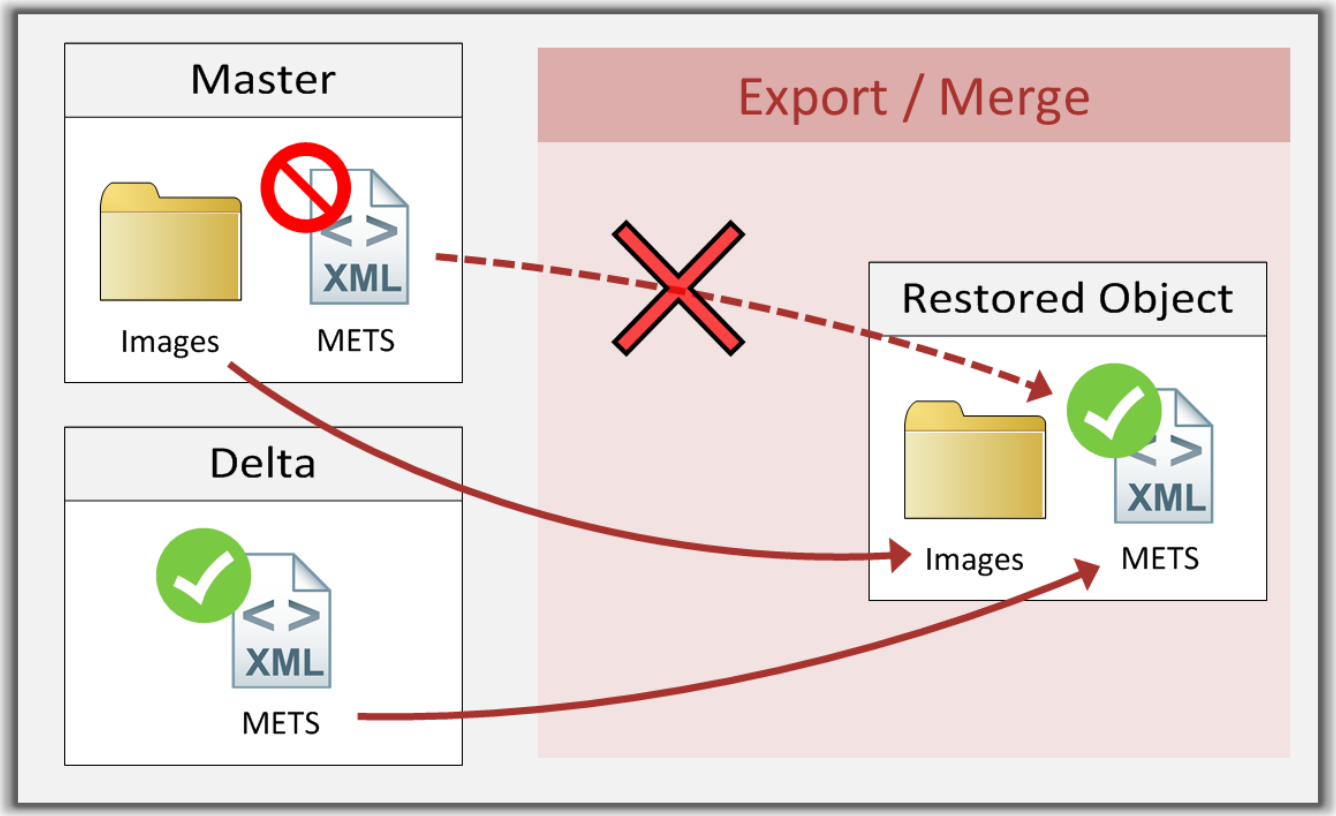
Correct (with namespace):

```
xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/version16/mets.xsd"
```

Ignore MD Extraction Error, because …

XML: wrong syntax of METS Schema Location (2)

… wrong METS file will be replaced by a correct version



XML: irrational JHOVE «not well-formed» errors

JHOVE

is Valid	false
is Well Formed	false
error Message	White spaces are required between publicId and systemId.,Failed to retrieve extractor properties

<oXygen/> XML Editor

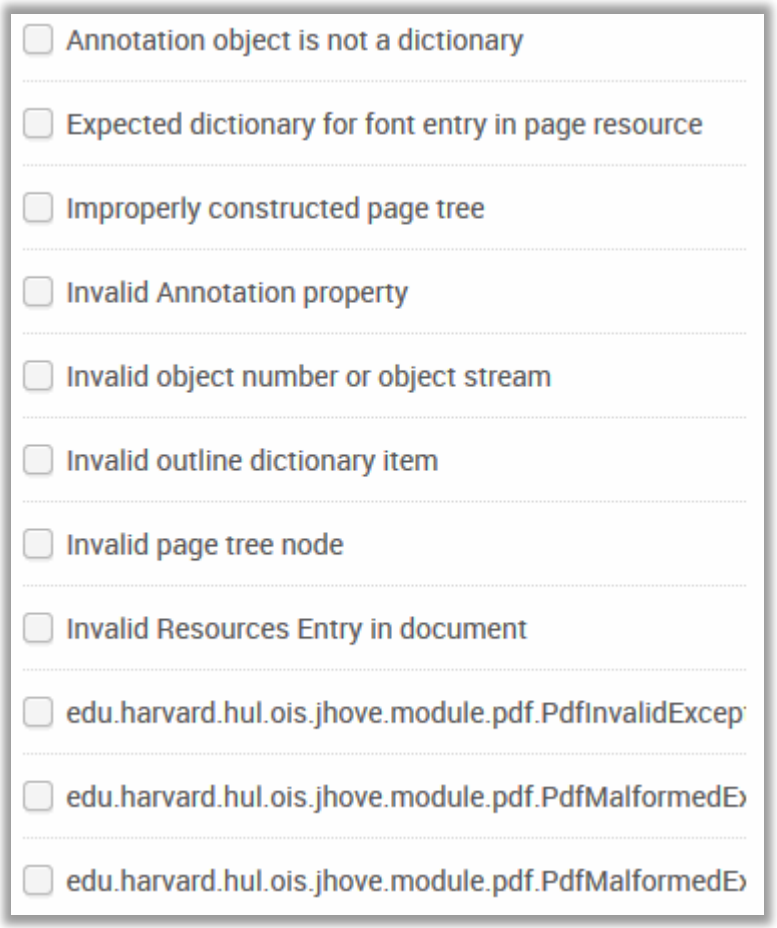
Document is valid.

- Schema URL with Redirect (HTTP status codes 301, 302):
  http://www.e-pics.ethz.ch/index/rosetta/schema/epics_rosetta_schema.xsd
  → **https://... !**
- HTTP status codes 403, 404 …

PDF:
JHOVE, are you serious?

PDF: various JHOVE errors (1)

«PDF Ignore Rule»



PDF: various JHOVE errors (2)

- We only take these errors seriously:
  - **Compression method is invalid or unknown to JHOVE**: file may be encrypted
  - **Failed to retrieve extractor properties**: file may be encrypted
  - **Invalid PDF trailer**: file may have been damaged during transmission

Producer-specific rules do not work in an Add/Update Representation process!

Research Data: maximum tolerance

Research Data: myriads of unknow formats

- NetCDF-4 (binary): Extension .nc, but hdf5 binary signature. PRONOM so far only knows NetCDF-3.
- .fcs: Flow Cytometry Standard File (binary)
- .md: Markdown Documentation File (text)
- .opj: OrCAD Project File (text)
- …

Research Data: what kind of text is it? (1)

- .txt: always x-fmt/111?

DROID

Format	Version	Mime type	PUID	Method
Plain Text File		text/plain	x-fmt/111	Extension

Manual Format ID or Rule

File Format:**Fail** - Multiple formats found for file (vs_Error.5)

Format ID	Format Mime Type	Format Name
x-fmt/110	text/plain	x-fmt/110
x-fmt/111	text/plain	x-fmt/111
x-fmt/13	text/plain	x-fmt/13
x-fmt/130	text/plain	x-fmt/130
x-fmt/14	text/plain	x-fmt/14
x-fmt/15	text/plain	x-fmt/15
x-fmt/16	text/plain	x-fmt/16
x-fmt/288		x-fmt/288
x-fmt/289		x-fmt/289
unknown		

- But what about the other txt formats?
- UTF-8 or ASCII?

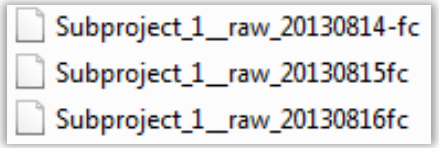
```
#Blattfl??che
m.leafarea<-matrix(TagBlattfl$Leaf_Area_ges,ncol=3
dim(m.efflux)
tail(m.efflux)
v.leafarea0<-(m.leafarea[,1])
```

Research Data: what kind of text is it? (2)

- .R, .c, .cpp, .h, .m, …
  ☹ How to search for these extensions?

Unable to identify format of file (vs_Error.3)

- no extension



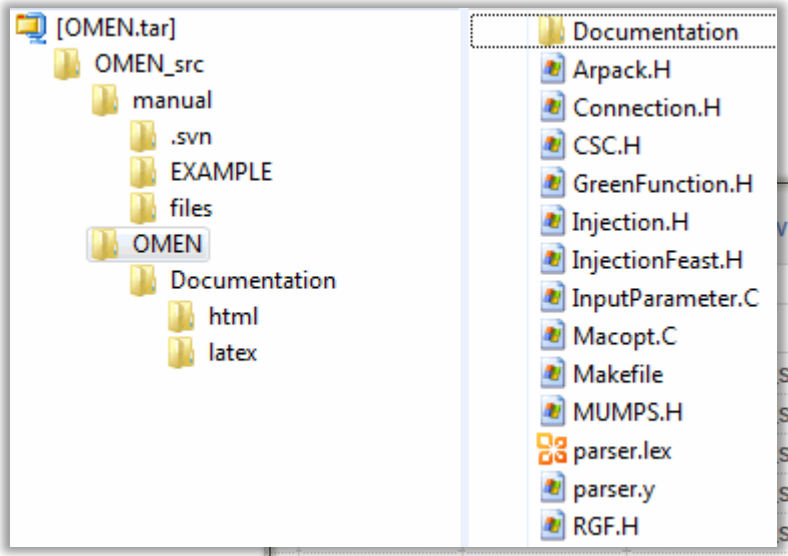
- .dat: to be identified as .csv (x-fmt/18)

- .csv: identification error in Rosetta

Format identification by file extension (vs_Error.9)

Manual Format ID or Rule

Research Data: put all your stuff in a ZIP file (1)



- Bytestream Extraction in Rosetta

vices | File Summary | BitStreams

				Extension	Mime Type	Format Nai	Format Description
			shared/...	svn-base		fmt/unknov	unknown format
			shared/...	svn-base		fmt/280	LaTeX (Master docume...
			shared/...	blg		fmt/unknov	unknown format
			shared/...	tex		fmt/281	LaTeX (Subdocument)
			shared/...	ps	application/po	x-fmt/406	PostScript
6	BS32430089	Chapter1.tex.b	/operational_shared/...	bak		fmt/281	LaTeX (Subdocument)
7	BS32430098	BS_square.eps	/operational_shared/...	eps	application/po	fmt/124	Encapsulated PostScr...
8	BS32430099	nano_trans.ep	/operational_shared/...	eps	application/po	fmt/124	Encapsulated PostScr...
9	BS32430107	circle_data.ep	/operational_shared/...	eps	application/po	fmt/124	Encapsulated PostScr...
1	BS32430111	spectral.eps	/operational_shared/...	eps	application/po	fmt/124	Encapsulated PostScr...
1	BS32430126	OMEN_manua	/operational_shared/...	out		fmt/unknov	unknown format
1	BS32430128	Chapter3.tex	/operational_shared/...	tex		fmt/281	LaTeX (Subdocument)
2	BS32430185	plot_internal_c	/operational_shared/...	m		fmt/unknov	unknown format
2	BS32430193	GFSolver.H	/operational_shared/...	H		fmt/unknov	unknown format
2	BS32430196	Macopt.C	/operational_shared/...	C		fmt/unknov	unknown format
2	BS32430214	LinearSolver.H	/operational_shared/...	H		fmt/unknov	unknown format
2	BS32430231	doxygen.png	/operational_shared/...	png	image/png	fmt/11	Portable Network Gra...

Research Data: put all your stuff in a ZIP file (2)

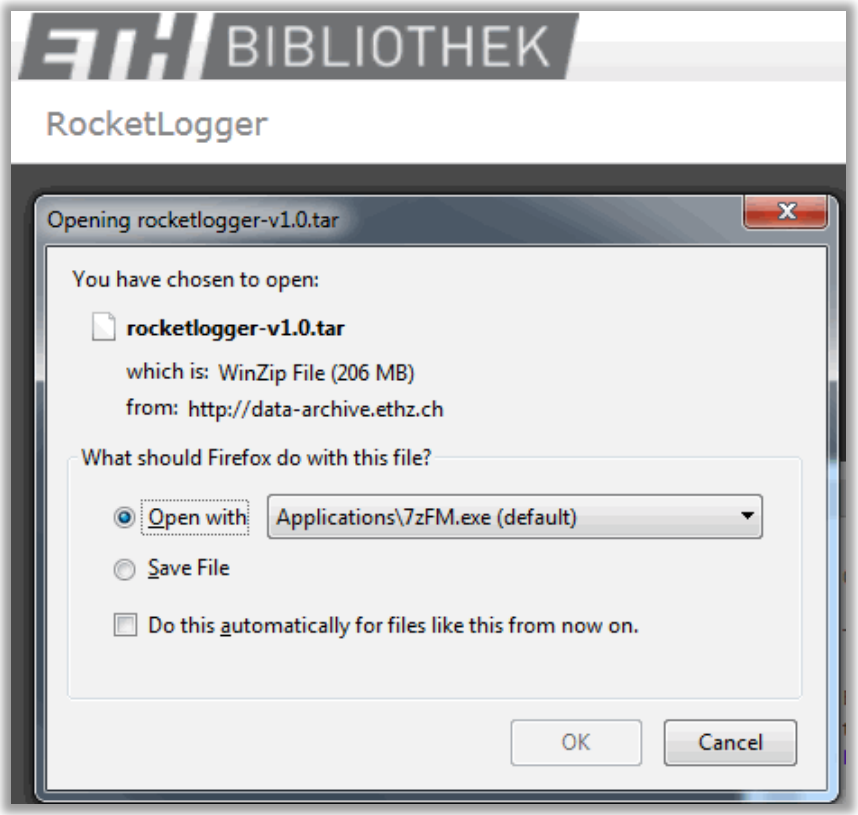
- Download single file …

Table of Contents

CH-ETH.2013.Ostermaier.META

Forschungspraktikum_Ostermaier
Read me
Subproject_1_Harvest_Leaf_Area_Neu
Subproject_1_R_Blattfläche_Watmung_Fiona
Subproject_1_R_Bofeuchte_Fiona
Subproject_1_R_Temp_Fiona
Subproject_1_Teilprojekt_1_Bodenatmung_FionasBohnen_Ne
Subproject_1_Teilprojekt_1_Pflanzen
Subproject_1__raw_20130814-fc
Subproject_1__raw_20130815fc
Subproject_1__raw_20130816-fc
Subproject_1__raw_20130816fc
Subproject_1__raw_20130817-fc-2n48
Subproject_1__raw_20130819-fc
Subproject_1__raw_20130819fc
Subproject_1__raw_20130820-fc
Subproject_1__raw_20130821-fc
Subproject_1__raw_20130822fc
Subproject_1__raw_20130903-fc
Subproject_1__raw_20130904-fc

- … or whole data package?



Conclusion

- Be prepared for anything!
- Pre-ingest analysis helps reduce efforts in Rosetta
- Pragmatic ad-hoc solutions vs. profound decisions from a truly long-term preservation perspective
- Rosetta Validation Stack Rules: a useful and powerful instrument – but handle with care!
- Leave the door open for future actions: define risks for preservation analysis, re-run validation, …
- What we need: more format signatures in Pronom, additional Validator/Extractor Plugins in Rosetta, …

Thank you for your attention!

- Any questions?
- … or answers?

http://www.library.ethz.ch/Digital-Curation

Dr. Franziska Geisser
Digital Curation Office
ETH-Bibliothek
Raemistrasse 101
8092 Zurich
+41 44 632 35 96
franziska.geisser@library.ethz.ch