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

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

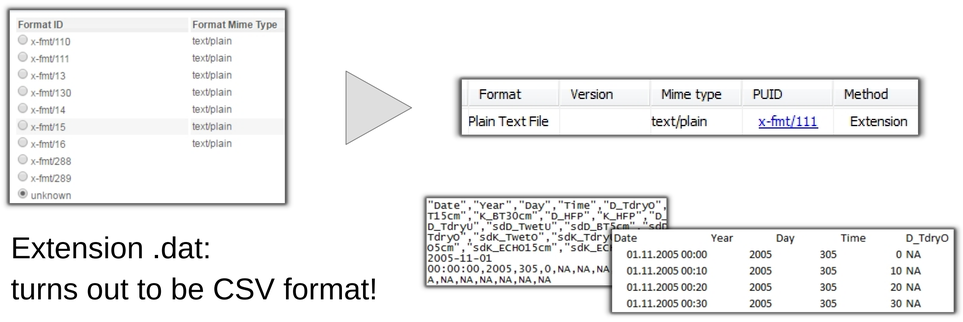
Matthias Töwe, Franziska Geisser, Roland E. Suri
ETH-Bibliothek, ETH Zurich – www.library.ethz.ch/Digital-Curation, data-archive@library.ethz.ch

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



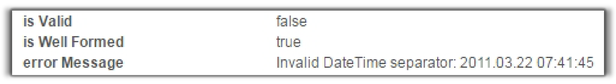
- Extension .dat:
  turns out to be CSV format!

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:

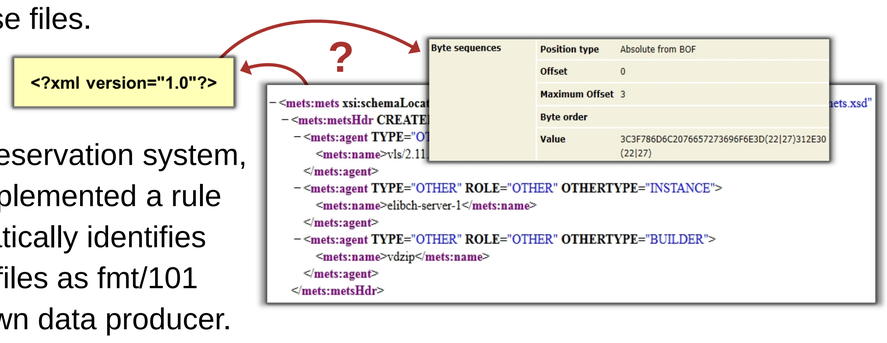
- DateTime format is not compatible with baseline TIFF specification
- We do not know what problems could arise in the future

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.

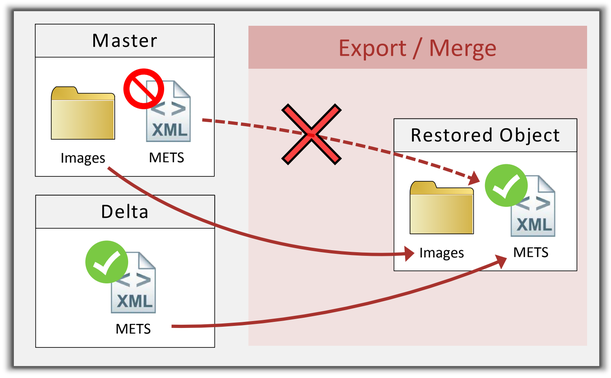


In our preservation system, we have implemented a rule that automatically identifies these XML files as fmt/101 for this known data producer.

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

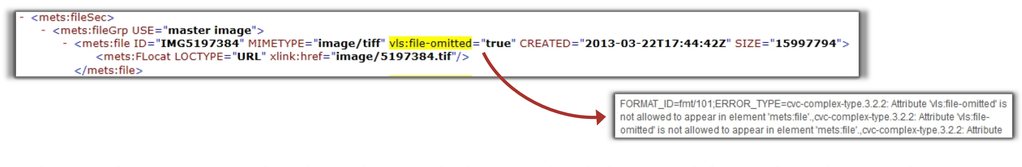
*) See http://wiki.opf-labs.org/display/Documents/JHOVE+issues+and+error+messages by Yvonne Friese

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=240,425,0,0,1,0). 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.