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

Matthias Töwe
ETH Zurich, ETH-Bibliothek
Rämistrasse 101
8092 Zurich, Switzerland
+41-(0)44 632 60 32
matthias.toewe@library.ethz.ch

Franziska Geisser
ETH Zurich, ETH-Bibliothek
Rämistrasse 101
8092 Zurich, Switzerland
+41-(0)44 632 35 96
franziska.geisser@library.ethz.ch

Roland E. Suri
ETH Zurich, ETH-Bibliothek
Weinbergstr. 74
8006 Zurich, Switzerland
+41-(0)44 632 39 19
roland.suri@library.ethz.ch

ABSTRACT
Format identification output needs to be assessed within an institutional context, also considering provenance information that is not contained in the data, but provided by data producers by other means. Sometimes, real issues in the data need to be distinguished from warnings. Ideally, this assessment should permit to decide where to invest effort in correcting issues, where to just document them, and where to postpone activities. The poster presents preliminary considerations at the ETH Data Archive of ETH-Bibliothek, the main library of ETH Zurich, on how to address file format identification and validation issues. The underlying issues are mostly independent of the specific tools and systems employed.

KEYWORDS
File format identification; Format validation; Technical metadata extraction; Ingest; Decision making; Preservation planning.

1. INTRODUCTION
To facilitate preservation actions in the future, digital archives rely on comprehensive technical information on file formats being available. Therefore, they try to derive as much information on the characteristics of digital objects as possible already upon or even before ingest. While the processes of format identification, validation or metadata extraction are understood in principle, a number of issues occur in everyday practice. They require an assessment of the specific case followed by a decision on how to proceed without compromising preservation options. Obviously, the broader the spectrum of file formats to be archived and the larger the number of files, the more are scalable efforts required.

One challenge is to understand what kind of issues can be encountered with different types of data. In addition, the tools in use might issue warnings which can also be related to their internal logic. An additional layer is metadata extraction which is also format related, but generally has less immediate effects than identification or validation issues. The practical implications of these issues differ between use cases, customers, types of material, and formats.

2. ETH DATA ARCHIVE
The ETH Data Archive is the institutional data archive of ETH Zurich, a research intensive technical university. We operate the application Rosetta [Ex Libris 2016] as digital preservation system, integrating DROID [The National Archives 2016a] (relying on PRONOM [The National Archives 2016b]) for file format identification and JHOVE [Open Preservation Foundation 2015] for format validation and metadata extraction.

Ingests to the ETH Data Archive comprise research data, administrative records and bequests to the University Archives, and born digital as well as digitized content from the library's online platforms and its digitization center. For research data alone, a broad range of use cases apply, from safeguarding data for a limited period of time (ten years at minimum) to publishing and preserving data in the long term. Several ingest workflows are available to cater for different requirements.

Handling all categories of this varied landscape of use cases adequately is a challenge in many respects. For handling format identification and validation issues, drawing criteria from those use cases' characteristics helps in gaining a better understanding of what actually matters most in each case. Preliminary results are presented in this poster.

3. ISSUES TO BE DISCUSSED
3.1 Format Identification
Ideally, format identification should yield reliable and unambiguous information on the format of a given file. In practice, a number of problems render the process much less straightforward. When it comes to large collections of heterogeneous files in a range of formats, which each may be subject to identification challenges, any effort on the individual files does not scale well. This is a situation we encounter with deposits of research data in particular, but also with bequests of mixed materials to our University Archives. As a result, more or less unsatisfactory decisions need to be taken to keep the volume of data manageable while not rendering potential identification or preservation measures in the future impossible.

3.1.1 Criteria
Example criteria to consider:

- 'Usability': can the file currently be used in the expected way with standard software?
- Tool errors: is an error known to be tool-related?
- Understanding: is the error actually understood?
- Seriousness: is an error concerning the significant properties of the format in question?
- Correctability: is there a straightforward or otherwise documented solution to the error?
- Risk of correcting: what risks are associated with correcting the error?

- Effort: what effort is required to correct the error in all files concerned?
- Authenticity: are there cases where a file's authenticity is more relevant than proper format identification?
- Provenance: is the data producer still available and willing to collaborate in the resolution of preservation issues at least with respect to future submissions?
- Intended preservation level: if bitstream preservation only is expected, the investment into resolving format identification issues might not be justified.
- Intended retention period: if data only needs to be retained for a maximum of ten years, incomplete file format identification might be acceptable.

Obviously, none of these criteria can easily be quantified or translated into simple rules. Even more unfortunately, some of these criteria can actually drive in opposite directions for the same set of files. Therefore, additional questions have evolved:

- Can we continue to handle format identification during ingest into the actual digital archive or will we need to perform it as a pre-ingest activity?
- In the latter case, how would we document in the digital archive measures which are taken prior to ingest to rectify identified problems?
- Under which conditions may we have to admit files with identification fmt/unknown into the archive?
- Should we envisage regular reruns of format identification? If so, how can they be done efficiently and effectively?
- Do we need local format definitions or can we exclusively rely on registries such as PRONOM [The National Archives 2016b] and add information there?
- Is the 'zero applications' risk addressed in any way?

As an indication of the practical and solution independent implications of these issues see e.g. [Mitcham 2015].

3.2 Format Validation and Characterization

File format validation and characterization through metadata extraction are related from a technical point of view. However, the implications of problems in either field can be quite different.

3.2.1 Format Validation

Format validation can fail when file properties are not in accord with its format's specification. However, it is not immediately clear if such deviations prevent current usability of a file or compromise the prospects for a file's long term preservability. If a file can be used readily today, this does not necessarily mean that the file is in itself 'valid enough', either. It rather means that the combination of the file with the application used today is working. This usually requires some generosity in the application's interpretation of the format specification. Obviously, it cannot be assumed that future tools which might have to rely on the documented specification will tolerate such issues. Therefore digital archives need to balance the efforts for making files valid vs. making files pass validation in spite of known issues.

3.2.2 Metadata Extraction

A failure to extract information on significant properties has no immediate consequences, and institutions need to balance the effort in correcting issues. This is even more the case, if embedded metadata or file properties are actually faulty and a correction would involve touching the file itself with a certain risk of unknowingly introducing other changes, too. Based on the criteria listed for format identification, we act therefore even more cautiously when it comes to fixing metadata extraction issues which require a manipulation of embedded metadata or other file properties.

4. ACKNOWLEDGMENTS

We thank all members of the Rosetta Format Library Working Group and in particular the colleagues from the National Library of New Zealand for their ongoing responsiveness and support.

5. REFERENCES

[1] Ex Libris. 2016. Rosetta. (2016). Retrieved July 4, 2016 from http://knowledge.exlibrisgroup.com/Rosetta

[2] Mitcham, Jenny. 2015. File identification ...let's talk about the workflows. (2015). Retrieved July 4, 2016 from http://digital-archiving.blogspot.ch/2015/11/file-identification-lets-talk-about.html

[3] Open Preservation Foundation. 2015. JHOVE - Open source file format identification , validation & characterisation. (2015). Retrieved July 4, 2016 from http://jhove.openpreservation.org/

[4] The National Archives. 2016a. Download DROID: file format identification tool. (2016). Retrieved July 4, 2016 from http://www.nationalarchives.gov.uk/information-management/projects-and-work/droid.htm

[5] The National Archives. 2016b. PRONOM - The Technical Registry. (2016). Retrieved July 4, 2016 from http://apps.nationalarchives.gov.uk/PRONOM/Default.aspx