Open Source Tools for Validation in the Digital Archive Workflow

By Merle Friedrichsen from the German National Library of Science and Technology (view original post)

On 9 / 10 November I was able to attend the ?No Time to Wait 2?, a conference on open source software in (film) archives held at the Austrian Film Museum in Vienna. In contrast to the first ?No Time to Wait? conference the scope was widened to audiovisual preservation, open formats, and standardization of formats. As part of the program, I had the chance to present on ?Open Source Tools in the Digital Archive Workflow? from our perspective.

Requirements for Open Source Tools for Validation

When we receive a file for our collection we will test if it fulfills the requests of the standard (e.g. does the PDF-file conform to the PDF-Standards?). There are different scenarios in which we use open source software for validation, and from each of these different requirements occur.

One requirement is that the **tool is easy to implement into existing software and workflows**, as we want to use the validation software within our archive framework. If we have implemented the tool, we can validate the files upon ingest into out archive. In another business case, we require a **simple graphical user interface** (GUI). With such a tool the acquisitions team can test the files as soon as they are received, and maybe even ask for another (valid) file ? if this is possible. When we are working together with other institutions or as a service provider, we typically receive a large amount of files which we validate before we ingest them into out archive. Due to this is we require a **command line interface** (CLI) in order to automate the validation of a large amount of files recursively through different folders.

Of course the validation software should fulfill its purpose: **checking against the standard** of the format. But apart from the standard, an institution might have other requests ? for example a specific resolution in TIFF files for a digitization project or a rule against embedded objects other than images and AV in PDF/A files. In order to easily enforce these, the tool should be able to **check the files against a custom policy**. As we want to store the report (success or failure) of the validation as preservation metadata, the **report should be available in XML or a different structured format**, to make it easy to integrate or process the output further. Any metadata schema will be appreciated, as it makes it more readable and (if applicable) the mapping to any other metadata schema would be easier (or already existing). The **performance of the tool** is another requirement.

In an ideal world, we would only receive valid files from data producers and service partners. A step to reach this goal would be an easy way to allow external providers to validate their files before they hand it over to the library. A **web service**, where anybody could upload files and check them against the standard (and a custom policy) would serve such a purpose. If the file is not valid, a **repair-possibility** for the error encountered in the file would help the data producer/provider, to hand over only valid files.

So, in no particular order a short overview of the requirements have regarding open source software for validation (besides from being open source):

- easy implementation into existing software and workflows
- simple Graphical User Interface
- Command Line Interface
- checking against file format standards
- checking against custom policy
- use standards for reports
- perform (fast) on a large amount of files
- webservice (standards and policies)
- repair-possibility

We?ve tested two open source tools for different file formats: veraPDF (PDF/A), and mediaConch (for Matroska/FFV1 ? for film). I am very happy to state that all of our requirements are fulfilled (or are possible to fulfill) by these tools.

VeraPDF

We have tested most of the requirements and the tool performs well. As we receive a lot of PDFs that do not need to fulfill the PDF/A standard but nevertheless must be without password, we created our own policy that checks whether a pdf is password encrypted or not. What we haven't tested yet is the integration into our existing software, but the implementation is discussed in the Rosetta Format Library Working Group (FLWG), a user group responsible for ? amongst other things ? deciding which tools should be rolled out within the Rosetta archive framework. Until now I haven't seen a web service based on veraPDF, but as the software is open source and well documented, it should be feasible to build one. Due to the fact that we could not find a file that could be repaired by veraPDF, we haven't tested the given repair functionality. But it is possible to fix the PDF document metadata, e.g. if a file does not conform to the standard, the PDF/A flag can be removed automatically.

MediaConch

Most of our requirements were tested and are fulfilled by MediaConch. There are only a few requirements that we have not tested yet. One of these is the implementation into existing software ? as with veraPDF, the FLWG is also discussing the integration of MediaConch. On the other hand we do not have special requirements for mkv / ffv1 files yet, so we couldn't test a custom policy ? simply because we currently do not have a custom policy. Due to the implemented checksums in ffv1 and Matroska there are several ?repair-possibilities? for a file with a bitflip. But as we haven't encountered a corrupted file (yet), we haven't tested this possibility.

Conclusion

Both veraPDF and MediaConch are suitable for our needs regarding digital object validation. I have to admit that I had a lot of fun testing these new tools, looking into the reports and figuring out how to write my own policy. It is worthwhile to work with open source tools ? especially when they are designed to fit the needs of a (digital) archive. And it is worth investing (money and / or time) to foster or to enhance these tools!

If you are interested in the conference you will find the recordings on the youtube channel of the conference.

Thanks to the organizers of this conference ? it was inspiring and encouraging!