Format Identification Benchmarking

Carl Wilson
Open Preservation Foundation
OPF Webinar 2016

The OPF Team

Community Manager : Rebecca McGuinness

- Events (face-to-face/virtual)
- Training (staff development)
- Comms (web/email/social)

Executive Director : Joachim Jung

- Membership (engagement/value)
- Open preservation advocacy
- Operational management

Technical Lead : Carl Wilson

- Infrastructure (host/test)
- Software stewardship (roadmap/maturity/packaging)
- Data corpora

OPF's Role

- Committed to improving the quality of open source DP tools.
- We believe this is best addressed through testing
- We can provide expertise in development and automation of software testing.
- Help people to automate the testing and reporting of data for the Benchmark DP project

What to benchmark?

OPF has had ambitions to test format identification tools because:

- there are a variety of tools available;
- they give differing results, even tools using the same signature data; and
- format identification is a vital “first contact” step in digital preservation workflows.

Where to start?

Our initial aim is to start as simply as possible:

- format identification data is straightforward as the outputs are fairly well defined;
- begin by evaluating three tools based on the same PRONOM signature data; so
- this gives us a fairly straightforward dataset to compare and evaluate.

Initial Aims

Benchmarking can be used to evaluate many properties of software, e.g. performance, we'll start by examining two aspects:

- accuracy, how accurately the tool identifies various formats; and
- coverage, how many formats a tool is capable of identifying.

Ground Truth

In order to answer these questions some concept of ground truth is required for testing. In this context ground truth refers to the accuracy of the data sets used for testing.

Accuracy

- Requires that we have test files where the format is known or is established by a tool known to be correct.
- An initial, naive approach is to simply compare the results of the tools. Where the tools agree then assume that the answer is accurate and investigate the cause of any discrepancies.
- This will be adopted initially as it makes life easy, better yardsticks of accuracy can be established once the comparison framework is in place.

Coverage

Ideally a master list of all possible formats is required, this doesn't exist but we can start by:

- compiling a master list of MIME types recognised by different tools;
- consulting available resources, e.g. IAANA lists, and UK Web archiving data; and
- examining existing corpora to see which formats they hold.

The PRONOM based tools

The initial scope will be simply to compare the 3 format identification tools that use the PRONOM signature files:

- **DROID**: the original PRONOM based ID tool maintained by the UK National Archive and developed in Java;
- **FIDO**: the OPF's python tool that; and
- **Siegfried**: developed in Go by Richard Lehane.

All of these tools use the same signature files and should, in theory, deliver identical results. In practise, they all use different algorithms under the hood and won't agree.

Corpora Used

Initially we're testing against 2 corpora:

- GovDocs selected, a set of 25 thousand files selected from http://digitalcorpora.org/corpora/govdocs, the starting set of 1 million files has been reduced by removing duplicate format examples, although the set still has 4 thousand different PDF examples.
- The OPF's format corpus https://github.com/openpreserve/format-corpus

Simple design

- The test data and results are made publicly available through a web interface.
- The interface builds upon a set of RESTful web services that make the data available as XML and JSON (plus CSV for the raw result data)

Additional Tools to Benchmark

Once we've covered the PRONOM based tools we'll test two more open source tools:

- ***nix File Utility**: the Fine Free File utility that's installed on Unix and Linux systems.
- **Apache Tika**: a content extraction tool that also supports format identification.

Additional Tools

These tools add another level of complexity as:

- the format identifiers are MIME based rather than PRONOM IDs
- the File magic (signatures) data covers many more formats than the other tools, i.e. the coverage is extended.

How To Get Involved

The initial web site will be made available in the first week of March, allowing people to browse and