Developing a file format signature in 15 minutes

The TZX format

David Clipsham

28 January 2013

The National Archives

PRONOM

The technical registry PRONOM

- File format registry
- Over 900 entries (PUIDs)
- Format extensions, mime/media types, links to documentation
- File format signatures
- http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

DROID

- File format identification utility
- Scans internal byte sequences of files
- Uses PRONOM registry signature files at its core
- Both command line and GUI interfaces
- Embedded within Tessella SDB
- http://www.nationalarchives.gov.uk/information-management/projects-and-work/droid.htm

Process overview

- Locate sample files, and scan with latest DROID sig file
- Observe common byte sequences
- Locate technical specification (where available) – use to confirm findings
- Further format research if necessary
- Finalise signature
- Generate test signature, and run against sample files
- Test against further file format corpus

Where to begin?

- Sample files
  - Existing datasets
  - ‘”parent directory” .<fileextension>’
  - ‘filetype:<fileextension>’

- Official format specifications

- Other tools (Filext, LoC’s digitalpreservation.gov/formats/index.shtml, Gary Kessler’s File Signature list,TRID)
- TNA’s signature generation utility

http://test.linkeddatapronom.nationalarchives.gov.uk/sigdev/index.htm

- TNA’s guide to developing file format signatures

http://www.nationalarchives.gov.uk/documents/information-management/pronom-file-signature-research.pdf

The TZX format

- A format for archiving ZX Spectrum programs
- Used with ZX emulation programs
- Large hobbyist community – lots of information available
- A converted .wav stream of the tape data
- World of Spectrum Archive: 10,000’s of examples

Common pitfalls

- Container formats – files that ID as OLE2 or zip (fmt/111, fmt/189, x-fmt/263) are very likely to be container signatures. Container signatures are similar to archive formats, in that they usually hold a number of smaller files. These together make up the file. The container can be opened with a tool like 7zip, or by changing the format extension to .zip

•Formats which are subsets of other formats – XSL is a subset of XML. PDF/A is a subset of PDF. To avoid multiple identification clashes, within PRONOM priority information is set to give the subset format precedence over its parent

•Archive formats – particularly web archives, which typically contain a short header then raw html data, so ID of HTML is common

•The header, or common byte sequences are usually found near the top of the file – BUT NOT ALWAYS!

Workshop goals

- Pick a format that is of interest to you
- Source some sample files
- Use a hex editor to attempt to determine commonalities
- Conduct research to confirm findings
- Use signature generation tool to create and test your signature
- Share your findings with the group
- Have fun!

Tools

- Hex Editor – any can be used. HxD is Freeware: http://mh-nexus.de/en/hxd/
- Signature development utility - http://test.linkeddatapronom.nationalarchives.gov.uk/sigdev/index.htm
- DROID - http://www.nationalarchives.gov.uk/information-management/projects-and-work/droid.htm
- PRONOM - http://www.nationalarchives.gov.uk/PRONOM/Default.aspx