ICA Africa Programme
Programme pour l'Afrique

ICA Training Programme
Programme de Formation

Digital Records Curation Programme

Week 3:
File Profiling Workshop

Learning Outcomes

At the end of this class, students should be able to:

- explain what file profiling is and why we do it
- understand the concepts of representation information
- understand checksums and technical registries
- use file profiling tools at a basic level

Digital objects are composed of binary code and expressed in file formats.

Selecting File Formats

TNA advises consideration of:

- Ubiquity (subjective but widely known)
- Support (number of compatible programmes and their ubiquity)
- Disclosure (openness of technical specs)
- Documentation quality (detailed enough to recreate?)
- Stability (rarely changing, with new versions backwards compatible)
- Ease of identification and validation (availability of validation tools and preference for formats with file signatures and version information within the file structure)

http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf

Selecting File Formats

TNA advises consideration of:

- Intellectual Property Rights (over technologies used by the format, such as image compression algorithms)
- Metadata Support (does the format allow inclusion of metadata)
- Complexity (the more complex the format, the more difficult and expensive to preserve)
- Interoperability (platform independent and used across programmes)
- Viability (formats with error-detection facilities are preferred)
- Reusability (can the original functionality be maintained?)

http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf

Identifying the files you have

- Tools have been developed to help us identify the tools we have
- DROID and JHOVE are two examples that we will use today
- The process of identifying files is known as ‘file profiling’
- File profiling produces a lot of useful information about files, including their formats, size and some information that is important in their ongoing preservation, such as their representation information

Representation Information

Representation information is any information required to understand and render both the digital material and the associated metadata.

Digital objects are not easily understandable to us without further data to interpret them. Representation information is information that allows raw data to be understood.

Standard representation information includes: file pathname or URI, last modification date, byte size, format, format version, media (MIME) type, format profiles, and optionally, checksums.

Checksums

- A checksum is a long string of alphanumeric characters that act as ‘digital fingerprints’ for digital objects, e.g. 96b13dbbc9f3bc569ddad9745f64b9cdb43ea9ae
- Created using checksum algorithms (cryptographic hash functions) such as MD5, SHA-1 (Secure Hash Algorithm 1), SHA-256 and SHA-512

Technical Registries

- File profiling depends on technical registries
- Technical registries are essentially databases
- They are used in digital preservation to enable organisations to maintain definitions of the formats, format properties, software, migration pathways etc. needed to preserve content over the long term.
- PRONOM is TNA's technical registry. DROID and other file profiling tools use PRONOM as a source of information.

File Profiling Tools

- DROID
  - Digital Record Object IDentification tool
  - Developed by TNA to perform automated batch identification of file formats
- JHOVE
  - Java tool developed by Harvard University to allow the automatic identification, validation and characterisation of a range of digital object types.

File Profiling Exercises

Any questions?