Automated Characterization in Preservation Workflows

Stephen Abrams
Harvard University Library
stephen_abrams@harvard.edu
Characterization

- Knowing what you have…
- A stable starting point for (iterative) preservation analysis, planning, and action

Adopted from A. Brown, “Developing Practical Approaches to Active Preservation,” *IJDC* 1:2 (June 2007).
Significant properties

• What is important to know?

• Format is the fundamental characterization property, as it enables the preservation of usable content, not merely bits

• Metadata initiatives regarding format-agnostic and format-specific properties
  – PREMIS, NISO Z39.87, AES-X098B, …

• Community best practice recommendations
  – AHDS Preservation Handbooks
  – InSPECT project
Scale drives the need for automation
Ingest workflow
Ingest workflow

• Push characterization operations as far up-stream as possible

• Early detection of anomalous or problematic data facilitates efficient remediation
Migration workflow
Migration workflow

- The post-migration quality assurance check is one of equivalence rather than equality
- Both sets of characterization information are evaluated relative to a canonical expression of the underlying information content
DROID

• A tool for signature-based format identification
  – Confidence-weighted matching of required and optional internal and external signatures
  – Regular expression-based signature definitions from PRONOM registry
  – GUI, command-line, and Java API interfaces

• National Archives (UK)
  – BSD license
Metadata Extractor

• An extensible system for extracting technical preservation metadata
  – Pluggable format-specific adapters for:
    • BMP, GIF, JPEG, TIFF
    • MP3, WAVE
    • HTML, XML
    • MS Excel, PowerPoint, Word, Works
    • Open Office, PDF
  – GUI and command-line interfaces

• National Library of New Zealand
  – Apache Public License
XCDL / XCEL

- Formal languages for expressing format specifications and the extracted properties of formatted objects
  - Extensible Characterisation Extraction Language (XCEL)
  - Extensible Characterisation Description Language (XCDL)

- Funded by the European Commission as part of the PLANETS project
JHŒVE

• Extensible framework for format identification, validation, and characterization
  – Pluggable format-specific modules for:
    • GIF, JPEG, JPEG 2000, TIFF
    • AIFF, WAVE
    • ASCII, HTML, UTF-8, XML
    • PDF
  – GUI, command-line, and Java API interfaces

• Collaborative project of Harvard University and the JSTOR Electronic-Archive Initiative
  – Funded by Andrew W. Mellon Foundation
  – GNU LGPL license
Format profiles (or subtypes) can be significant
  - For example, TIFF has many variants…
    - TIFF 4.0 – 6.0
    - Baseline Class B, G, P, R; extension Class Y
    - TIFF/IT (ISO 12639)
      - File types CT, LW, HC, MP, BP, BL, FP; conformance levels P1, P2
    - TIFF/EP (ISO 12234-2)
    - GeoTIFF
    - EXIF 2.0, 2.1 (JEIDA-49-1998), 2.2 (JEITA CP-3451)
    - RFC 1314, Class F (RFC 2306)
    - TIFF-FX (RFC 2301)
      - Profiles C, F, L, M, S
    - DNG
JHVE2

• A next generation architecture for format-aware preservation processing
  – Three-fold goals:
    • Re-factor the existing architecture to achieve higher performance, simplify system integration, and encourage third-party enhancement
    • Provide significant new function
    • (Re-) Implement modules

• Collaborative project of Harvard University, Portico, and Stanford University
  – Funded by Library of Congress/NDIIPP
  – Educational Community License
JH\textsuperscript{2}VE\textsuperscript{2} enhancements

• Separate identification from validation
  – In JHOVE the identified format is determined by the first (or last) module that validates
  – JHOVE\textsuperscript{2} will use DROID for signature-based identification

• Standardize the handling of format profiles and error reporting

• Support configurable criteria for validity

• Provide more comprehensive documentation
JHVE2 enhancements

- There is a useful distinction between well-formedness, validity, renderability, and usability
  - Well-formedness and validity are “binary” determinations relative to a specification
  - Renderability is a “binary” determination relative to a specific rendering tool
  - Usability is a “fuzzy” determination relative to local policies and heuristics
JHOVE2 data model

• Implicit assumption in JHOVE: 1 object = 1 file = 1 format

• But what about…
  – TIFF with embedded ICC profile and XMP metadata
    1 object = 1 file = 3 formats
  – JPEG 2000 JPX fragmentation
    1 object = $n$ files = 1 format
  – ESRI Shapefile
    1 object = 3 files = 3 formats

• In JHOVE2: 1 object = $n$ files = $m$ formats
JHVE2 “generic” module API

- Outer iteration over digital objects; inner iteration over processes

```java
while (has-another-object) {
    while (has-another-process) {
        process (object, state);
    }
}
```
JHVE2 modules

• Validation and characterization for:
  – GIF, JPEG, JPEG 2000, TIFF
  – AIFF, WAVE
  – ASCII, HTML, SGML, UTF-8, XML
  – PDF
  – Shapefile
  – ICC

• Symbolic display of selected binary formats

• Assessment based on prior characterization and locally-defined policy rules and heuristics
Why are there no good commercial tools?
Why are there no good commercial tools?

• Identification tools from other domains
  – Data recovery
  – Forensic investigation

• Validation
  – PDF pre-flight
  – Rendering (with error reporting), rather than validation

• Characterization
  – Descriptive, not technical
  – Narrow format support
Questions?

www.significantproperties.org.uk
droid.sourceforge.net
meta-extractor.sourceforge.net
www.planets-project.eu
hul.harvard.edu/jhove

stephen_abrams@harvard.edu