Digital Archives, Digital Forensics, and Open Source Search: Developing Together

Mark A. Matienzo, Yale University Library
Open Source Search Conference
Chantilly, VA
October 2, 2012
About Me

• I am an archivist
• Occasionally I develop software
• I am not a digital forensics “expert”
Digital Archives at Yale
Digital Forensics in the Archival Domain

• Increasing use of digital forensics tools/methodologies within the context of digital archives programs (Kirschenbaum et al. 2010)

• Technology-focused work (John 2008; Woods & Brown 2009; AIMS Work Group 2012; BitCurator 2012)

• Methodology-focused work (Duranti 2009; Xie 2011)
Significant Barriers to use of Digital Forensics in Archives

- Cost (Kirschenbaum et al. 2010; Daigle 2012)
- Complexity (Kirschenbaum et al. 2010; Daigle 2012)
- Digital archives as an emerging market for forensics
Potential of Open Source Digital Forensics Software

• Requires additional tool development work to be useful for archivists (Kirschenbaum et al. 2010)

• Requires additional integration work (Lee et al. 2012)
Institutional Context

• Focus on implementation of and development with open source digital forensics software at Yale University Library

• Work must support accessioning, arrangement, description, and management of born-digital archival material

• Material received on physical media as primary focus
Design Principles

- Use and develop with open source digital forensics software to support accessioning, arrangement, and description of born-digital archival records.

- Focus on first two phases (preservation and searching) of Carrier’s (2005) model of digital investigation process.

- Curation micro-services (Abrams, et al. 2010) as philosophical basis to guide development and implementation.

- Digital objects needing management are both disk images themselves (Woods, Lee, and Garfinkel 2011) and bitstreams that they contain.

- Intention of forensic soundness, but assume much of state is lost.
## Micro-services as Design Philosophy*

<table>
<thead>
<tr>
<th>Principles</th>
<th>Preferences</th>
<th>Practices</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Granularity</td>
<td>• Small and simple over large and complex</td>
<td>• Define, decompose, recurse</td>
</tr>
<tr>
<td>• Orthogonality</td>
<td>• Minimally sufficient over feature-laden</td>
<td>• Top down design, bottom up implementation</td>
</tr>
<tr>
<td>• Parsimony</td>
<td>• Configurable over the prescribed</td>
<td>• Code to interfaces</td>
</tr>
<tr>
<td>• Evolution</td>
<td>• The proven over the merely novel</td>
<td>• Sufficiency through a series of incrementally necessary steps</td>
</tr>
</tbody>
</table>

*Outcomes over means

*UC Curation Center/California Digital Library, 2010
Workflow

Start accessioning process

Retrieve media

Assign identifiers to media

Write-protect media

Record identifying characteristics of media as metadata

Create image

Package images and metadata for ingest

Extract filesystem- and file-level metadata

Ingest transfer package

Document accessioning process

End accessioning process
Disk Image Acquisition

- Requires a combination of hardware (drives/media readers, controller cards, write blockers) and software
- In some cases, hardware requires specific software (e.g. floppy disk controller cards that sample magnetic flux transitions)
- Goal: sector image interpretable by multiple tools
Metadata Extraction

- Use open source digital forensics software (Sleuth Kit, fiwalk) and other open source tools to characterize media, volume, file system, and file information

- Attempt to repurpose this information as descriptive, structural, and/or technical metadata to support accessioning, appraisal, and processing

- Extracted metadata expressed in Digital Forensics XML

- Easily extensible and straightforward to process
Extraction Plugins

- Created Fiwalk plugins to perform additional analysis and evaluation of files/bitstreams within disk images
- Virus identification plugin using ClamAV/pyclamd
- File format identification against PRONOM format registry using Open Planets Foundation’s FIDO
- Code (including additional plugins) available online: https://github.com/anarchivist/fiwalk-dgi/
Gumshoe

• Prototype web application to provide search/browse interface to metadata extracted from disk images
• Built as a Ruby on Rails application using Blacklight
• http://github.com/anarchivist/gumshoe
Blacklight

- [http://projectblacklight.org](http://projectblacklight.org)
- Ruby gem for use in Rails applications
- Provides discovery layer over Solr indexes, with support for faceting, bookmarking, etc.
- Use is fairly common in library community
- Implementers include Stanford, Columbia, NC State, UVA, WGBH, National Agricultural Library (AGNIC) ...
Indexing Process

1. Start indexing process
2. Disk image or DFXML
3. Parse DFXML
4. Extract metadata to DFXML
5. Extract strings?
   - Yes: Perform string extraction
   - No: Normalize output for each file
7. Construct Solr document for each file
8. Post documents to Solr
9. End indexing process
Data Normalization

- Depends on DFXML gem
- Translate metadata-layer data to more easily searchable or human-readable version (e.g. file type/file system codes to text labels; certain flags to booleans)
- Data type coercion (integers-as-strings to integers)
- Prepend full path data to filename
- Transform timestamps to ISO8601
Features

- Basic browse view, with sorting by filename, size, modification/access/creation times
- Faceting by disk image, extension, file format, file type
- Searching based on metadata values (e.g. checksums), file content (still under development; somewhat slow)
- Basic bookmarking
Advantages

• Faster (and more forensically sound) to extract metadata once rather than having to keep processing an image

• Possibility of developing better assessments during accessioning process (significance of directory structure, accuracy of timestamps)

• Integrating additional extraction processes and building supplemental tools is simple
Limitations

- Use of tools limited to specific types of file systems
- Requires additional integration and data normalization to work with additional tools
- DFXML is not (currently) a metadata format common within domains of archives/libraries; somewhat in flux
- Extracted metadata harder for archivists to repurpose in some cases based on level of granularity
Work in Progress

- BitCurator project under development; early release available for testing: http://wiki.bitcurator.net

- The Sleuth Kit and related tools under continuing development (Autopsy, fiwalk, etc.): http://sleuthkit.org

- Additional testing, development integration under work at Yale and NYPL
Thanks!

Mark A. Matienzo
mark@matienzo.org
http://matienzo.org
@anarchivist
References