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)

- Barriers to adoption: cost, complexity, need for additional tool development (Kirschenbaum et al. 2010; Daigle 2012; Lee et al. 2012)

- BitCurator project: http://bitcurator.net
Initial Goals

• 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

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

• Intention of forensic soundness, but assumption that much of the state is lost

• Curation micro-services (Abrams, et al. 2010) as philosophical basis to guide our thinking
## 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>
<tr>
<td></td>
<td>Outcomes over means</td>
<td></td>
</tr>
</tbody>
</table>

*UC Curation Center/California Digital Library, 2010
Workflow
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
Acquired 1,039 disk images from across 69 accessions at Manuscripts and Archives.
Initial Work with Disk Images

- Experimentation with various tools: The Sleuth Kit (3.1+), Autopsy, Pyflag, bulk_extractor ...
- Basic integration/processing with shell scripts or Python
- Discovering fiwalk was my "eureka" moment
Metadata Extraction

- Used 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
File Systems

- Ran metadata extraction on 812 images

![Bar Chart]

- ISO9660: 386
- FAT12: 246
- Unidentified: 155
- HFS+: 14
- FAT16: 11
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/
File Analysis

• Ran enhanced metadata extraction on 619 images (using our plugins)

• Performed analysis on 49,724 files within images

• Successfully identified 43,729 files (147 unique file types) against PRONOM format registry

• Identified 9 files as containing virus signatures (2 unique virus signatures)
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](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

Start indexing process → Disk Image

<table>
<thead>
<tr>
<th>Disk image or DFXML</th>
<th>Extract metadata to DFXML</th>
</tr>
</thead>
</table>

Parse DFXML

<table>
<thead>
<tr>
<th>Disk image or DFXML</th>
<th>DFXML</th>
</tr>
</thead>
</table>

**Extract strings?**

- **Yes** → Perform string extraction
- **No** → Normalize output for each file

Normalize output for each file → Construct Solr document for each file → Post documents to Solr → 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 conversion (e.g. 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/content/creation times

• Faceting by disk image, extension, file format, file type

• Basic bookmarking

• Searching based on metadata values (e.g. checksums), file content (still under development; somewhat slow)
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
- Still struggling with how to best present data to archivists
Work in Progress

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

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

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