Digital Forensics for Digital Archives

Mark A. Matienzo
Manuscripts and Archives, Yale University Library
Digital Preservation 2012
Arlington, VA
July 25, 2012
Digital Archives at Yale
Digital Forensics
Digital Forensics + Digital Archives
Design Principles

• Use digital forensics software and methodology to support accessioning, arrangement, and description of born-digital archival records

• Mitigate risk of media deterioration and obsolescence

• Prefer open source solutions whenever possible

• Integrate into a larger, but yet-to-be-defined workflow
Applied Methodology

• Use Carrier’s (2005) model of the digital investigation process: Preservation ↔ Searching ↔ Reconstruction

• Volume and file system as main areas for analysis

• Assume much of the state is already lost

• Methods should approach or intend forensic soundness
Documenting Media

- SharePoint-based list
- Unique identifiers for each piece
- Allows basic documentation of imaging process
<table>
<thead>
<tr>
<th>New</th>
<th>Actions</th>
<th>Settings</th>
<th>View: All Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Electronic Records on Media Accessioning Log**

<table>
<thead>
<tr>
<th>Type</th>
<th>Media number</th>
<th>Media Format</th>
<th>Imaging Date</th>
<th>Imaging Successful?</th>
<th>Bag Created?</th>
<th>Metadata Extracted?</th>
<th>Transfer to Storage Date</th>
<th>Examiner</th>
<th>Image Format</th>
<th>Imaging Software</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2011-M-075.0001</td>
<td>CD-R</td>
<td></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
<td>Glick, Kevin</td>
<td></td>
<td>N/A</td>
<td>FAT32</td>
</tr>
<tr>
<td></td>
<td>2011-M-075.0002</td>
<td>DVD-R</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td>Glick, Kevin</td>
<td>ISO</td>
<td>ImgBurn</td>
<td>ISO</td>
</tr>
<tr>
<td></td>
<td>2011-M-075.0003</td>
<td>DVD-R</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td>Glick, Kevin</td>
<td>ISO</td>
<td>ImgBurn</td>
<td>ISO</td>
</tr>
<tr>
<td></td>
<td>2011-M-075.0004</td>
<td>DVD-R</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td>Glick, Kevin</td>
<td>ISO</td>
<td>ImgBurn</td>
<td>ISO</td>
</tr>
<tr>
<td></td>
<td>2011-M-075.0005</td>
<td>DVD-R</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td>Glick, Kevin</td>
<td>ISO</td>
<td>ImgBurn</td>
<td>ISO</td>
</tr>
<tr>
<td></td>
<td>2011-M-075.0006</td>
<td>DVD-R</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td>Glick, Kevin</td>
<td>ISO</td>
<td>ImgBurn</td>
<td>ISO</td>
</tr>
<tr>
<td></td>
<td>2011-M-075.0007</td>
<td>CD-R</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td></td>
<td>Glick, Kevin</td>
<td>ISO</td>
<td>ImgBurn</td>
<td>ISO</td>
</tr>
<tr>
<td></td>
<td>2011-M-075.0008</td>
<td>CD-R</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td></td>
<td>Glick, Kevin</td>
<td>ISO</td>
<td>ImgBurn</td>
<td>ISO</td>
</tr>
<tr>
<td></td>
<td>2011-M-075.0009</td>
<td>CD-R</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td>Glick, Kevin</td>
<td>ISO</td>
<td>ImgBurn</td>
<td>ISO</td>
</tr>
<tr>
<td></td>
<td>2011-M-075.0010</td>
<td>DVD-R</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td>Glick, Kevin</td>
<td>ISO</td>
<td>ImgBurn</td>
<td>ISO</td>
</tr>
<tr>
<td></td>
<td>2011-M-075.0011</td>
<td>CD-R</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td>Glick, Kevin</td>
<td>ISO</td>
<td>ImgBurn</td>
<td>ISO</td>
</tr>
<tr>
<td></td>
<td>2011-M-075.0012</td>
<td>CD-R</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td>Glick, Kevin</td>
<td>ISO</td>
<td>ImgBurn</td>
<td>ISO</td>
</tr>
<tr>
<td></td>
<td>2011-M-075.0013</td>
<td>Zip disk</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td>Glick, Kevin</td>
<td>dd (Raw)</td>
<td>FTK Imager 3.0.0.1443</td>
<td>FAT32</td>
</tr>
<tr>
<td>Media number</td>
<td>Type</td>
<td>Media Format</td>
<td>Imaging Date</td>
<td>Imaging Successful</td>
<td>Bag Created</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-M-075.0001</td>
<td>CD-R</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-M-075.0002</td>
<td>DVD-R</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-M-075.0003</td>
<td>DVD-R</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-M-075.0004</td>
<td>DVD-R</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-M-075.0005</td>
<td>DVD-R</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-M-075.0006</td>
<td>DVD-R</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-M-075.0007</td>
<td>DVD-R</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-M-075.0008</td>
<td>CD-R</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-M-075.0009</td>
<td>CD-R</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-M-075.0010</td>
<td>DVD-R</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-M-075.0011</td>
<td>CD-R</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-M-075.0012</td>
<td>CD-R</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-M-075.0013</td>
<td>Zip disk</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Imaging Media

- Requires a combination of hardware and software
- In some cases, software depends on particular hardware
- No single universal solution for our workflow
Imaging Hardware

- Drives (eg floppy drives, flash cardreaders)
- Interface cards (Catweasel, Kryoflux, FC5025)
- Writeblockers
Imaging Software

• FTK Imager (proprietary; gratis)

• Hardware-specific imaging software for floppy interface cards

• Other software tested: dd, Guymager, etc.
Metadata Extraction

- Desire to repurpose existing information as archival description and reports to other staff

- Ideal output is XML; can be packaged with disk images going into medium- or long-term storage

- Tools: Fiwalk/The Sleuth Kit; FTK Imager; testing others

- Have integrated file-format identification (using OPF’s FIDO) and virus/malware recognition (using ClamAV) using Fiwalk’s plugin architecture
Sample DFXML Output

```xml
<?xml version='1.0' encoding='UTF-8'?>
<dfxml version='1.0'>
  <metadata
    xmlns='http://www.forensicswiki.org/wiki/Category:Digital_Forensics_XML'
    xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'
    xmlns:dc='http://purl.org/dc/elements/1.1'/>
    <dc:type>Disk Image</dc:type>
  </metadata>
  <creator version='1.0'>
    <!-- provenance information re: extraction - software used; operating system -->
  </creator>
  <source>
    <image_filename>2004-M-088.0018.dd</image_filename>
  </source>
  <volume offset='0'>
    <!-- partitions within each disk image -->
    <fileobject>
      <!-- files within each partition -->
    </fileobject>
  </volume>
  <runstats>
    <!-- performance and other statistics -->
  </runstats>
</dfxml>
```
Sample DFXML Output

<fileobject>
  <filename>_ublist1.wpd</filename>
  <partition>1</partition>
  <id>1</id>
  <name_type>r</name_type>
  <filesize>202152</filesize>
  <unalloc>1</unalloc>
  <used>1</used>
  <inode>3</inode>
  <meta_type>1</meta_type>
  <mode>511</mode>
  <nlink>0</nlink>
  <uid>0</uid>
  <gid>0</gid>
  <mtime>2001-02-22T22:30:52Z</mtime>
  <atime>2001-02-22T05:00:00Z</atime>
  <ctime>2001-02-22T22:31:54Z</ctime>
  <libmagic>(Corel/WP)</libmagic>
  <byte_runs>
    <byte_run file_offset='0' fs_offset='16896' img_offset='16896' len='512'/>
  </byte_runs>
  <hashdigest type='md5'>d7bc22242c0a88fd8b68712980d5ab28</hashdigest>
  <hashdigest type='sha1'>64bf2bdf82e33fcda50158804483ac611e753db5</hashdigest>
</fileobject>
Analysis/Processing

• Once acquired, we can perform additional analysis or reporting to captured assets or records

• Few tools are easily useable by archivists (BitCurator toolset under development will help)

• Additional forensic tools can be used for archival arrangement and description of this information
Forensic Toolkit

• Proprietary application to analyze files, filesystems, etc.

• Provides full-text indexing, tagging, bookmarking, file presentation/viewing, and reporting

• Used at Yale, Stanford, and other institutions for archival processing of born-digital records

• Still a challenge to use given the complexity of the application
Guidelines for Media Sanitization

Recommendations of the National Institute of Standards and Technology

Richard Kissel
Matthew Scholl
Steven Skolnicko
Xing Li
Gumshoe

• Prototype based on Blacklight (Ruby on Rails + Solr)
• Indexing code works with fiwalk output or directly from a disk image
• Populates Solr index with all file-level metadata from fiwalk and, optionally, text strings extracted from files
• Provides searching, sorting and faceting based on metadata extracted from filesystems and files
• Code at http://github.com/anarchivist/gumshoe
<table>
<thead>
<tr>
<th>File Path</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>/home/ubuntu/Desktop/MyStuff/SEC Documents/spch121708cc-idata.wmv</td>
<td></td>
</tr>
</tbody>
</table>
| **Filename**: spch121708cc-idata.wmv  
| **Full Path**: /home/ubuntu/Desktop/MyStuff/SEC Documents  
| **Image file**: ubnist1_casper_rw_gen2  
| **Type**: Regular file  
| **Size (bytes)**: 37887210  
| **Inode number**: 15697  
| **MD5**: 8e7d1611c0b870f658529d94556f9a21  
| **Format (libmagic)**: Microsoft ASF  
| **Modification Time**: 2008-12-17T17:10:00Z  
| **Access Time**: 2008-12-29T05:35:21Z  
| **Change Time**: 2008-12-29T05:35:21Z |
| /Compressed/logfile1.txt | 
| **Filename**: logfile1.txt  
| **Full Path**: /Compressed  
| **Image file**: nts1_gen2  
| **Type**: Regular file  
| **Size (bytes)**: 21888890  
| **Inode number**: 48 |
Advantages

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

- Develop better assessments during accessioning process (directory structure significant? timestamps accurate?)

- Integrating additional extraction processes and building supplemental tools takes less time.
Limitations

• Use of tools limited to specific types of filesystems
• Additional software requires additional integration and data normalization
• DFXML is not (currently) a metadata format common within domains of archives/libraries
• Extracted metadata maybe harder to repurpose for descriptive purposes 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 development (Autopsy, fiwalk, etc.): http://sleuthkit.org

- Additional testing and integration under work at Yale, using DFXML as common schema whenever possible

- Possible development of a new media log to record media/imaging metadata and workflow status
Thanks!

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