Taking Bitstreams Seriously: Digital Forensics and the BitCurator Environment

Cal Lee
School of Information and Library Science
University of North Carolina, Chapel Hill

Digital Preservation 2013
July 23-25, 2013
Alexandria, Virginia
**Two Main Acquisition Paths**

<table>
<thead>
<tr>
<th>Systematic Transfer</th>
<th>Dealing with Whatever you Get</th>
</tr>
</thead>
<tbody>
<tr>
<td>Close pre-coordination between Producer and Archive*</td>
<td>Little pre-coordination between Producer and Archive</td>
</tr>
<tr>
<td>Archive has (at least some) say in how materials are produced, packaged and transferred</td>
<td>Archive has relatively little say in how materials are produced, packaged and transferred</td>
</tr>
<tr>
<td>Relatively little need to engage in guess work or ad hoc description after the transfer</td>
<td>Substantial need to engage in guess work or ad hoc description after the transfer</td>
</tr>
</tbody>
</table>

*Archive in the OAIS sense – can be library, archives, museum, data center, …*
Examples of Systematic Transfer Developments

• Protocols and tools for transfer (e.g. BagIt)¹
• Systematic and predictable ingest workflows²
• Substantial, well-documented Producer-Archive interactions (e.g. PAIMAS)³

Sometimes things are a little messier
Applying Digital Forensics to Digital Collections – Previous Work*

• Ross and Gow (1999) - potential relevance of advances in data recovery and digital forensics to collecting institutions
• More recently - active stream of literature related to use of forensic tools and methods for digital collections, including activities at the British Library, National Library of Australia and Indiana University
• PERPOS (Georgia Tech) – has applied data capture and extraction to US presidential materials
• “Computer Forensics and Born-Digital Content in Cultural Heritage Collections” - symposium and report (2010)
• Born Digital Collections: An Inter-Institutional Model for Stewardship (AIMS) - framework for the stewardship of born-digital materials, including digital forensics methods
• Digital Records Forensics project - has articulated connections between the concepts of digital forensics and archival science
• Two Open Planets Foundation (OPF) Hackathons this year – one in Copenhagen, one in Chapel Hill

• Funded by Andrew W. Mellon Foundation
  – Phase 1: October 1, 2011 – September 30, 2013
  – Phase 2 – October 1, 2013 – September 30, 2014

• Partners: School of Information and Library Science (SILS) at UNC and Maryland Institute for Technology in the Humanities (MITH)
Core BitCurator Team

- Cal Lee, PI
- Matt Kirschenbaum, Co-PI
- Kam Woods, Technical Lead
- Alex Chassonoff, Project Manager (UNC)
- Sunitha Misra, GA (UNC)
- Porter Olsen, GA (MITH)
## Two Groups of Advisors

### Professional Experts Panel
- Bradley Daigle, University of Virginia Library
- Erika Farr, Emory University
- Jennie Levine Knies, University of Maryland
- Jeremy Leighton John, British Library
- Leslie Johnston, Library of Congress
- Naomi Nelson, Duke University
- Erin O’Meara, Gates Archive
- Michael Olson, Stanford University Libraries
- Gabriela Redwine, Harry Ransom Center, University of Texas
- Susan Thomas, Bodleian Library, University of Oxford

### Development Advisory Group
- Barbara Guttman, National Institute of Standards and Technology
- Jerome McDonough, University of Illinois
- Mark Matienzo, Yale University
- Courtney Mumma, Artefactual Systems
- David Pearson, National Library of Australia
- Doug Reside, New York Public Library
- Seth Shaw, University Archives, Duke University
- William Underwood, Georgia Tech
BitCurator Goals

• Develop a system for collecting professionals that incorporates the functionality of open-source digital forensics tools

• Address two fundamental needs not usually addressed by the digital forensics industry:
  – incorporation into the workflow of archives/library ingest and collection management environments
  – provision of public access to the data
BitCurator Environment

• Bundles, integrates and extends functionality (primarily data capture and reporting) of open source software: fiwalk, bulk extractor, Guymager, The Sleuth Kit, sdhash and others

• Can be run as:
  – Self-contained environment (based on Ubuntu Linux) running directly on a computer (download installation ISO)
  – Self-contained Linux environment in a virtual machine using VirtualBox
  – As individual components run directly in your own Linux environment or (whenever possible) Windows environment
Acknowledgement to Simson Garfinkel

• Digital forensics scholar at Naval Postgraduate School
• Responsible for:
  – fiwalk
  – Bulk Extractor
  – Digital Forensics XML (DFXML) metadata conventions
  – forensicswiki.org
  – digitalcorpora.org
BitCurator-Supported Workflow

- Acquisition
- Reporting
- Redaction
- Metadata Export
### Metadata Conventions of the BitCurator Tools: Digital Forensics XML (DFXML)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt;dfxml&gt;</td>
<td>Root element, marks the beginning and end of the DFXML metadata file. The <code>&lt;dfxml&gt;</code> element contains the primary elements reported in the walk XML structure: <code>&lt;metadata&gt;</code>, <code>&lt;creator&gt;</code>, <code>&lt;source&gt;</code>, <code>&lt;volume&gt;</code>, and <code>&lt;runstats&gt;</code>.</td>
<td><code>&lt;metadata&gt;</code>, <code>&lt;creator&gt;</code>, <code>&lt;source&gt;</code>, <code>&lt;volume&gt;</code>, <code>&lt;runstats&gt;</code>, <code>&lt;sectorsize&gt;</code>, <code>&lt;pagesize&gt;</code>, <code>&lt;acquisition_seconds&gt;</code></td>
</tr>
<tr>
<td>2</td>
<td>DFXML</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>&lt;metadata&gt;</td>
<td>The <code>&lt;metadata&gt;</code> tag provides header information that defines the metadata in the DFXML document. Includes namespace declaration, namespace schema location, and other information that is used to define the elements used in the XML file. These declarations provide information on the types of standardization schemes used to convey information in the DFXML document. The <code>&lt;metadata&gt;</code> tag may also contain high level descriptive information about the DFXML document rendered in Dublin Core (dc), in order to increase interoperability.</td>
<td><code>&lt;dc:type&gt;</code>, <code>&lt;dc:creator&gt;</code>, <code>&lt;dc:title&gt;</code>, <code>&lt;dc:description&gt;</code>, for more information on Dublin Core element set, see [21].</td>
</tr>
<tr>
<td>4</td>
<td>&lt;creator&gt;</td>
<td>The Creator element provides documentation about the program and computing environment in which the disk analysis (or capture) take place. &lt;Creator&gt; includes tags documenting the program that initiated the capture creating the DFXML file, and other contextual information about the system on which the capture was performed.</td>
<td><code>&lt;program&gt;</code>, <code>&lt;version&gt;</code>, <code>&lt;build_environment&gt;</code>, <code>&lt;execution_environment&gt;</code></td>
</tr>
</tbody>
</table>
Metadata Generation and Reporting

Acquiring Disk Images with Guymager
Exporting Filesystem Content Using fiwalk

Fiwalk produces a DFXML file showing the volumes, directories, and files contained within a disk image.

Image File
/home/bcadmin/Desktop/SampleData/sampleimage.E01

Output XML File
/home/bcadmin/Desktop/SampleData/sampleimage.xml

Command Line Output:
Viewing the Command Line Output

Command Line Output:

```bash
>>> Command Executed for Fiwalk = ['fiwalk', '-f', '-X', '/home/bcadmin/Desktop/SampleData/sampleimage.E01']

>>> Success!!! Fiwalk created the following file(s):
  o /home/bcadmin/Desktop/SampleData/sampleimage.xml
```
DFXML Output from fiwalk

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```xml
<dfxml version="1.0">
  <metadata>
    <dc:type>Disk Image</dc:type>
  </metadata>
  <creator version="1.0">
    <program>fiwalk</program>
    <version>4.1.0</version>
    <build_environment>
      <compiler>GCC 4.6</compiler>
      <library name="afflib" version="3.7.1"/>
      <library name="libwfd" version="20130416"/>
    </build_environment>
    <execution_environment>
      <command_line>
        fiwalk -X/home/bcadmin/Desktop/SampleData/sampleimage.xml /home/bcadmin/Desktop/SampleData/sampleimage.E01
      </command_line>
      <start_time>2013-07-20T05:34:37Z</start_time>
    </execution_environment>
  </creator>
  <source>
    <image_filename>/home/bcadmin/Desktop/SampleData/sampleimage.E01</image_filename>
  </source>
  <volume_offset="0">
    <partition_offset>0</partition_offset>
    <sector_size>512</sector_size>
    <block_size>512</block_size>
    <ftype>2</ftype>
  </volume_offset>
</dfxml>
```
<fileobject>
   <filename>Documents and Settings/All Users/Documents/
            My Pictures/Sample Pictures/Blue hills.jpg</filename>
   ...
   <filesize>28521</filesize>
   <alloc>1</alloc>
   <used>1</used>
   <inode>6245</inode>
   ...
   <uid>0</uid>
   <gid>0</gid>
   <mtime>1208174400</mtime>
   <ctime>1257729636</ctime>
   <atime>1257729636</atime>
   <crttime>1257729636</crttime>
   <seq>2</seq>
   <libmagic>JPEG image data, JFIF standard 1.02</libmagic>
   <byte_runs>
      <run file_offset='0' fs_offset='0' img_offset='363200512' 
            len='0'/>
   </byte_runs>
   <hashdigest type='MD5'>
      6fb2a38dc107eacb41cf1656e899cf70
   </hashdigest>
   <hashdigest type='SHA1'>
      4eee44b18576e84de7b163142b537d2fe6231845
   </hashdigest>
</fileobject>
Identifying “Features” of Interest in Disk Images or Directories

Bulk Extractor
<table>
<thead>
<tr>
<th>Scanner</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>scan-accts</td>
<td>Looks for phone numbers, credit card numbers, etc</td>
</tr>
<tr>
<td>scan_base64</td>
<td>Decodes BASE64 text</td>
</tr>
<tr>
<td>scan_kml</td>
<td>Detects KML (Keyhole Markup Language) files – used to identify geographic locations</td>
</tr>
<tr>
<td>scan_gps</td>
<td>Detects XML from Garmin GPS devices</td>
</tr>
<tr>
<td>scan_aes</td>
<td>Detects in-memory AES (Advanced Encryption Standard) keys from the key schedules</td>
</tr>
<tr>
<td>scan_json</td>
<td>Detects JavaScript Object Notation files</td>
</tr>
<tr>
<td>scan_exif</td>
<td>Detects EXIF structures from JPEG files</td>
</tr>
<tr>
<td>scan_zip</td>
<td>Detects and decompresses ZIP files and zlib streams</td>
</tr>
<tr>
<td>scan_gzip</td>
<td>Detects and decompresses GZIP files and gzip streams</td>
</tr>
<tr>
<td>scan_pdf</td>
<td>Extracts text from some kinds of PDF files</td>
</tr>
<tr>
<td>scan_hiber</td>
<td>Detects and decompresses Windows hibernation file fragments</td>
</tr>
<tr>
<td>scan_winprefetch</td>
<td>Detects and extracts fields from windows prefetch files and file fragments</td>
</tr>
</tbody>
</table>
bulk_extractor scan completed. See Status, below, for details.

Options
bulk_extractor -o
/home/bcadmin/Desktop
/home/bcadmin/Desktop

Status
Elapsed time: 0.4985 sec.
Overall performance: 2.958 MBytes/sec.
Total email features found: 0

Done.

Report is Ready
bulk_extractor has completed.
Report bulk-extractor-output has been opened and is ready for viewing.

Close
Histogram of Email Addresses (Specific Instances in Context on Right)
Matching Bulk Extractor Output (Based on Byte Offsets) to fiwalk Output (Based on Filesystem Location)
Generating BitCurator Reports
# Report: File System Statistics and Files

## Technical Metadata

Disk Image: image_filename: charlie-work-usb-2009-12-11.aff

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SECTORSIZE</td>
<td>1024</td>
</tr>
<tr>
<td>FTYPE STR</td>
<td>ntfs</td>
</tr>
<tr>
<td>PARTITION OFFSET</td>
<td>512</td>
</tr>
<tr>
<td>BLOCK SIZE</td>
<td>4096</td>
</tr>
<tr>
<td>ACQUISITION SECONDS</td>
<td>73</td>
</tr>
<tr>
<td>FIRST BLOCK</td>
<td>0</td>
</tr>
<tr>
<td>BLOCK COUNT</td>
<td>258559</td>
</tr>
<tr>
<td>LAST BLOCK</td>
<td>258558</td>
</tr>
<tr>
<td>PAGESIZE</td>
<td>16777216</td>
</tr>
<tr>
<td>FTYPE</td>
<td>1</td>
</tr>
<tr>
<td>IMAGE_FILENAME</td>
<td>charlie-work-usb-2009-12-11.aff</td>
</tr>
<tr>
<td>Number of Files</td>
<td>128</td>
</tr>
<tr>
<td>Total Directories</td>
<td>23</td>
</tr>
<tr>
<td>Total Deleted Files</td>
<td>0</td>
</tr>
<tr>
<td>Total Unused Files</td>
<td>0</td>
</tr>
<tr>
<td>Files with Nlinks &gt; 1</td>
<td>0</td>
</tr>
</tbody>
</table>
# Report: Bulk Extractor Features

**Note:**
- FIUF: Total features unallocated to files
- FLTF: Total features unallocated to files
- FICR: Total features in compressed regions

<table>
<thead>
<tr>
<th>Bulk Extractor Report Files</th>
<th>Feature Instances</th>
<th>FLTF</th>
<th>FIUF</th>
<th>FICR</th>
</tr>
</thead>
<tbody>
<tr>
<td>annotated_telephone.txt</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>annotated_rfc822.txt</td>
<td>258</td>
<td>39</td>
<td>219</td>
<td>110</td>
</tr>
<tr>
<td>annotated_zip.txt</td>
<td>127</td>
<td>8</td>
<td>119</td>
<td>3</td>
</tr>
<tr>
<td>annotated_windirs.txt</td>
<td>466</td>
<td>13</td>
<td>453</td>
<td>180</td>
</tr>
<tr>
<td>annotated_domain.txt</td>
<td>653</td>
<td>48</td>
<td>605</td>
<td>317</td>
</tr>
<tr>
<td>annotated_exif.txt</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>annotated_winpc.txt</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>annotated_email.txt</td>
<td>500</td>
<td>42</td>
<td>458</td>
<td>224</td>
</tr>
</tbody>
</table>
### Report: Bulk Extractor Features

**Feature File: annotated_email.txt**

<table>
<thead>
<tr>
<th>Filename</th>
<th>Position</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email/Charlie_Email</td>
<td>22991360-ZIP-115</td>
<td><a href="mailto:charlie@m57.biz">charlie@m57.biz</a></td>
</tr>
<tr>
<td>Email/Charlie_Email</td>
<td>22991360-ZIP-148</td>
<td><a href="mailto:terry@m57.biz">terry@m57.biz</a></td>
</tr>
<tr>
<td>Email/Charlie_Email</td>
<td>22991360-ZIP-55</td>
<td><a href="mailto:pat@m57.biz">pat@m57.biz</a></td>
</tr>
<tr>
<td>Email/Charlie_Email</td>
<td>22991360-ZIP-648</td>
<td><a href="mailto:pat@m57.biz">pat@m57.biz</a></td>
</tr>
<tr>
<td>Email/Charlie_Email</td>
<td>22991360-ZIP-130</td>
<td><a href="mailto:terry@m57.biz">terry@m57.biz</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22991858-ZIP-37</td>
<td><a href="mailto:pat@m57.biz">pat@m57.biz</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22991858-ZIP-97</td>
<td><a href="mailto:charlie@m57.biz">charlie@m57.biz</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22992313-ZIP-129</td>
<td><a href="mailto:pat@m57.biz">pat@m57.biz</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22992313-ZIP-55</td>
<td><a href="mailto:charlie@m57.biz">charlie@m57.biz</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22992313-ZIP-712</td>
<td><a href="mailto:pat@m57.biz">pat@m57.biz</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22992313-ZIP-732</td>
<td><a href="mailto:pat@m57.biz">pat@m57.biz</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22992834-ZIP-108</td>
<td><a href="mailto:aix.pery@yahoo.com">aix.pery@yahoo.com</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22992834-ZIP-129</td>
<td><a href="mailto:rebinfriz31@mail.com">rebinfriz31@mail.com</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22992834-ZIP-45</td>
<td><a href="mailto:charlie@m57.biz">charlie@m57.biz</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22993168-ZIP-133</td>
<td><a href="mailto:charlie@m57.biz">charlie@m57.biz</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22993168-ZIP-204</td>
<td><a href="mailto:charlie@m57.biz">charlie@m57.biz</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22993168-ZIP-226</td>
<td><a href="mailto:aix.pery@yahoo.com">aix.pery@yahoo.com</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22993168-ZIP-247</td>
<td><a href="mailto:rebinfriz31@mail.com">rebinfriz31@mail.com</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22993168-ZIP-51</td>
<td><a href="mailto:aix.pery@yahoo.com">aix.pery@yahoo.com</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22993587-ZIP-37</td>
<td><a href="mailto:pat@m57.biz">pat@m57.biz</a></td>
</tr>
<tr>
<td>Unknown</td>
<td>22993587-ZIP-97</td>
<td><a href="mailto:charlie@m57.biz">charlie@m57.biz</a></td>
</tr>
</tbody>
</table>
Nautilus Scripts

• Scripts that can be run using the GNOME file manager called Nautilus (Linux analog to Windows Explorer or Mac OS X Finder)

• Can be used in the BitCurator environment or your own Linux environment
MD5 Hashes of Files (Nautilus Script)
BitCurator Tools and Further Information

Get the software
Documentation and technical specifications
Screencasts
Google Group
http://wiki.bitcurator.net/

People
Project overview
Publications
News
http://www.bitcurator.net/

Twitter: @bitcurator