BitCurator: An Open-Source Project for Libraries and Archives that Takes Bitstreams Seriously

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

September 17, 2013
National Digital Stewardship Alliance
Infrastructure Working Group
Why should we care about digital forensics?

• **Not** to solve crimes or catch malicious users
• Recognition of how data can be recovered when layers of technology fail or are no longer available
• Capturing information from places that are not immediately visible
• Ensuring that actions taken on files don’t make irreversible changes to essential characteristics (e.g. MAC values)
• Attending to order of volatility – some types of data change more quickly and often than others
• Learning about available tools and techniques to deal with files
• Established practices for documenting acquisition and processing, so others will know what might have changed
• Overlap between technical knowledge required to do digital forensics and ad hoc acquisition of digital materials by libraries/archives
Storage Media Acquisition and Handling Profile for Digital Repositories*

BitCurator

• Funded by Andrew W. Mellon Foundation
  – Phase 1: October 1, 2011 – September 30, 2013
  – Phase 2 – October 1, 2013 – September 30, 2014
• Partners: SILS at UNC and Maryland Institute for Technology in the Humanities (MITH)
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
Core BitCurator Team

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

<table>
<thead>
<tr>
<th>Professional Experts Panel</th>
<th>Development Advisory Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Bradley Daigle, University of Virginia Library</td>
<td>• Barbara Guttman, National Institute of Standards and Technology</td>
</tr>
<tr>
<td>• Erika Farr, Emory University</td>
<td>• Jerome McDonough, University of Illinois</td>
</tr>
<tr>
<td>• Jennie Levine Knies, University of Maryland</td>
<td>• Mark Matienzo, Yale University</td>
</tr>
<tr>
<td>• Jeremy Leighton John, British Library</td>
<td>• Courtney Mumma, Artefactual Systems</td>
</tr>
<tr>
<td>• Leslie Johnston, Library of Congress</td>
<td>• David Pearson, National Library of Australia</td>
</tr>
<tr>
<td>• Naomi Nelson, Duke University</td>
<td>• Doug Reside, New York Public Library</td>
</tr>
<tr>
<td>• Erin O’Meara, Gates Archive</td>
<td>• Seth Shaw, University Archives, Duke University</td>
</tr>
<tr>
<td>• Michael Olson, Stanford University Libraries</td>
<td>• William Underwood, Georgia Tech</td>
</tr>
<tr>
<td>• Gabriela Redwine, Harry Ransom Center, University of Texas</td>
<td></td>
</tr>
<tr>
<td>• Susan Thomas, Bodleian Library, University of Oxford</td>
<td></td>
</tr>
</tbody>
</table>
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 e.g. Virtual Box or VMWare
  – As individual components run directly in your own Linux environment or (whenever possible) Windows environment

*To read about and download the environment, see: http://wiki.bitcurator.net/
BitCurator-Supported Workflow

- Acquisition
- Reporting
- Redaction
- Metadata Export

Source media → Additional supported output formats

- Split raw
- E01

AFF packaged image → Capture and image metadata

Log, device info

Filesystem metadata extraction (fwalk)

Filesystem report, DFXML

File ↔ disk block map

Image analysis (private, sensitive info)

Bulk Extractor

Annotated feature file

Reports: file distribution, sensitive info, hashes, etc. Exported as PDF, .xlsx, plain text as appropriate

Human-readable reporting scripts / plug-in

Analysis (accounts, filesystem activity)

reg2xml (and similar)

XML dump of registry

Reports: user accounts, device usage, environment Exported as PDF, .xlsx, plain text as appropriate

Analysis (file similarity, deduplication)

sdhash

Reports: locations, status of similar files Exported as PDF, .xlsx, plain text as appropriate

Reduction of disk image

Ruleset (patterns, hashes, etc) describing what to redact

Python redaction scripts

Redaction

Metadata export

Python ( lxml) export scripts

METS, MODS, EAD as required by user

See: http://bitcurator.net
High-Level view of Metadata Generation and Reporting

<table>
<thead>
<tr>
<th>Tag name</th>
<th>Element name</th>
<th>Description</th>
<th>May contain</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;dfxml&gt;</td>
<td>DFXML</td>
<td>Root element, marks the beginning and end of the DFXML metadata file. The &lt;dfxml&gt; element contains the primary elements reported in the disk's xml structure: &lt;metadata&gt;, &lt;creator&gt;, &lt;source&gt;, &lt;volume&gt;, and &lt;runstats&gt;.</td>
<td>&lt;metadata&gt;, &lt;creator&gt;, &lt;source&gt;, &lt;volume&gt;, &lt;runstats&gt;, &lt;sector_size&gt;, &lt;pagesize&gt;, &lt;acquisition_seconds&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;metadata&gt;</td>
<td>Metadata</td>
<td>The &lt;metadata&gt; 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 &lt;metadata&gt; tag may also contain high level descriptive information about the DFXML document rendered in Dublin Core (dc), in order to increase interoperability.</td>
<td><a href="">dc:type</a>, <a href="">dc:creator</a>, <a href="">dc:title</a>, <a href="">dc:description</a>; for more information on Dublin Core element set, see (21).</td>
</tr>
<tr>
<td></td>
<td>Creator</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 took place.</td>
<td>&lt;program&gt;, &lt;version&gt;, &lt;build_environment&gt;, &lt;execution_environment&gt;</td>
</tr>
</tbody>
</table>

http://www.bitcurator.net/2013/02/06/dfxml-tag-library/
Acquiring Disk Images with Guymager
Exporting Filesystem Content Using fiwalk
Viewing the Command Line Output

Command Line Output:

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

>>> Success!!! Fiwalk created the following file(s):
  o /home/bcadmin/Desktop/SampleData/sampleimage.xml
```
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="libwft" 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>
</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>
  <crt ime>1257729636</crt ime>
  <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>
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

<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
FIUF: 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.txt</td>
<td>22991360-ZIP-115</td>
<td><a href="mailto:charlie@m57.biz">charlie@m57.biz</a></td>
</tr>
<tr>
<td>Email/Charlie_Email.txt</td>
<td>22991360-ZIP-148</td>
<td><a href="mailto:terry@m57.biz">terry@m57.biz</a></td>
</tr>
<tr>
<td>Email/Charlie_Email.txt</td>
<td>22991360-ZIP-55</td>
<td><a href="mailto:pat@m57.biz">pat@m57.biz</a></td>
</tr>
<tr>
<td>Email/Charlie_Email.txt</td>
<td>22991360-ZIP-648</td>
<td><a href="mailto:pat@m57.biz">pat@m57.biz</a></td>
</tr>
<tr>
<td>Email/Charlie_Email.txt</td>
<td>22991858-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:rubinfriz31@mail.com">rubinfriz31@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:rubinfriz31@mail.com">rubinfriz31@mail.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)
Quick Start Guide
Most recent version always available at:
http://wiki.bitcurator.net/

BitCurator Quick Start Guide
v0.3.4

Last updated: September 8, 2013
Open Source Software Strategy

• Code released under GPL, v3 – available through GitHub
• Existing code that we incorporate is generally GPL or public domain (government products)
• Packaging elements of the code to be integrated into other environments (e.g. Archivematica)
• Regular contact with individuals and organizations responsible for other development efforts
• Phase 2 of BitCurator involves both further development and significant community engagement, including hiring of Community Lead (Porter Olsen)
Engagement Efforts

- Oodles of conference papers, presentations, posters, demos, workshops, tutorials, webinars
- BitCurator users mailing list
- Web site, wiki, social media
- Professional education offerings:
  - Two-day class for Digital Archives Specialist (DAS) curriculum of Society of American Archivists (SAA) – Cal Lee and Kam Woods
  - “Born-Digital Materials: Theory & Practice” Course for Rare Book School – Matt Kirschenbaum and Naomi Nelson
  - Courses for graduate students at UNC SILS
Dedicated Digital Forensics Lab at UNC SILS
Hackathon: Tackling Real-World Collection Challenges with Digital Forensics Tools and Methods

- June 3-5, 2013, Chapel Hill, NC
- Collaboration between OPF, UNC SILS and Library of Congress
- Two categories of participants: collection owners, developers
- Participants from Germany, Netherlands, US

http://sils.unc.edu/events/2013/hackathon-opf

Institutions represented:

- Baker Library, Harvard Business School
- Duke University
- Florida State University
- Massachusetts Institute for Technology
- Library of Congress
- National Snow and Ice Data Center
- New York Public Library
- Open Planets Foundation
- Sloane Art Library, UNC
- Stanford University Libraries
- UNC Libraries
- University of Freiburg
- UNC SILS
- Wake Forest University
Sustainability Plans

• The primary model we’re exploring is a membership-based consortium.

• Looking at existing consortia for potential models (e.g. ArchivesSpace, IIPC)

• Members of Consortium would pay an annual fee, which could entitle them to member benefits such as:
  – Access to dedicated personnel who can assist with integration of BitCurator into their institutions’ particular workflows
  – Ability to submit development requests
  – Priority notification of new code releases
  – BitCurator online training events

• This is all very tentative. Draft sustainability plan under development now.
Sources for BitCurator 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