Enabling Digital Forensics Practices in Libraries, Archives and Museums: The BitCurator Experience

By

Christopher Lee and Kam Woods

Presented At

The Digital Forensic Research Conference

DFRWS 2014 USA Denver, CO (Aug 3rd - 6th)

DFRWS is dedicated to the sharing of knowledge and ideas about digital forensics research. Ever since it organized the first open workshop devoted to digital forensics in 2001, DFRWS continues to bring academics and practitioners together in an informal environment. As a non-profit, volunteer organization, DFRWS sponsors technical working groups, annual conferences and challenges to help drive the direction of research and development.

http://dfrws.org
Enabling Digital Forensics Practices in Libraries, Archives and Museums: The BitCurator Experience

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

Digital Forensics Research Workshop
August 3-6, 2014
Denver, CO
Goals of Libraries, Archives and Museums (LAMs) When Acquiring Materials

• Ensure integrity of materials
• Allow users to make sense of materials and understand their context
• Prevent inadvertent disclosure of sensitive data
### Fundamental Archival Principles

<table>
<thead>
<tr>
<th><strong>Provenance</strong></th>
<th><strong>Original Order</strong></th>
<th><strong>Chain of Custody</strong></th>
</tr>
</thead>
</table>
| • Reflect “life history” of records  
• Records from a common origin or source should be managed together as an aggregate unit | Organize and manage records in ways that reflect their arrangement within the creation/use environment | • “Succession of offices or persons who have held materials from the moment they were created”¹  
• Ideal recordkeeping system would provide “an unblemished line of responsible custody”² |

---


Digital Forensics Can Help to Fulfill Archival Principles

Provenance  • Identify, extract and save essential information about context of creation

Original Order  • Reflect original folder structures, files associations, related applications and user accounts

Chain of Custody  • Documentation of how records were acquired and any transformations to them
  • Use well-established hardware and software mechanisms to ensure that data haven’t been changed inadvertently

Identifying Sensitive Information  • Identify personally identifying information, regardless of where it appears
  • Flag for removal, redaction, closure or restriction
From Bitstreams to Heritage:

Putting Digital Forensics into Practice in Collecting Institutions

Christopher A. Lee, Kam Woods, Matthew Kirschenbaum, and Alexandra Chassanoff

http://www.bitcurator.net/docs/bitstreams-to-heritage.pdf
Digital Forensics Lab @ UNC School of Information and Library Science
Digital Forensics in LAMs

• In recent years, LAMs have discovered the value of applying various digital forensics methods, for example:
  – use of write blockers
  – generation of disk images
  – applying cryptographic hashes to files
  – capture of Digital Forensics XML (DFXML)
  – scanning bitstreams for personally identifying information
Need for Adaptation of Tools and Tasks for LAM Users

• While existing digital forensics tools provide valuable functionality, they don’t always fit well into primary workflows of LAMs.

• For example, LAMs are particularly concerned with:
  – structure and persistence of metadata
  – provisions for providing public access to data
  – support for older technologies (e.g. floppy disks, HFS)
• 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 Chassanoff, Project Manager
- Sunitha Misra, Software Developer (UNC)
- Kyle Bickoff, 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, Digital Public Library of America</td>
</tr>
<tr>
<td>Jeremy Leighton John, British Library</td>
<td>Courtney Mumma, Artefactual Systems</td>
</tr>
<tr>
<td>Leslie Johnston, US National Archives and Records Administration</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, Beinecke, Yale University</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 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 Elements

- Acquisition
- Reporting
- Redaction
- Metadata Export
Mounted Devices set to Read-Only by Default*

*Not to replace hardware-based write blocking, but useful for various purposes
Creating a Disk Image in Guymager

*Developed by Guy Voncken*
Mounting a Disk Image to Browse the Contents
Mounting a Disk Image to Browse the Contents
Bulk Extractor* – Identifying Potentially Sensitive Information

See: http://www.forensicswiki.org/wiki/Bulk_extractor

*Developed by Simson Garfinkel
Histogram of Email Addresses (Specific Instances in Context on Right)

BitCurator-0.2.0 [Running]

Highlight:
- Match case

Reports
- beoutput
  - domain.txt
  - domain_histogram.txt
  - email.txt
  - email_histogram.txt
  - ether.txt
  - ether_histogram.txt
  - json.txt
  - pcap.txt
  - pcap_histogram.txt
  - url.txt
  - url_histogram.txt
  - url_services.txt
  - winfiles.txt

Feature Filter
- Match case

Navigation
- Image File: sampleimage.E01, 42273785, privacy@Motorola.com
- Feature File: email.txt
- Feature Path: 42273785
- Feature: privacy@Motorola.com

Text
- your credit card number, so this information can only be viewed by Motorola. Motorola uses Secure Sockets Layer (SSL) encryption on technology, the highest level of security on the Internet. The SSL protocol provides server authentication, data integrity, and privacy on the Web. This security measure helps ensure that no impostors, eavesdroppers, or vandals get your personal information transmitted, including credit card information, but also verifies the identity of the server and that the original message arrives safely at its destination. However, no data transmission over the Internet can be guaranteed to be 100% secure. As a result, while we strive to protect your personal information, Motorola cannot ensure or warrant the security of any information you transmit to us or from our Web site, and therefore you use our site at your own risk. Once we receive your transmission, we use our best effort to ensure its security on our systems. If you have established a "user profile" on a Motorola website, you may change the information you provided at an
BitCurator Reporting Tool

Run `fiwalk`, annotate the bulk_extractor output, and generate Office / PDF reports.
If you haven't run bulk_extractor yet, use the button to the right to launch and run it first.

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

**Bulk Extractor Feature Directory**
/home/bcadmin/Desktop/SampleData/bulk-extractor-output

**Output Directory (fiwalk output, annotated features, and reports will appear in here)**
/home/bcadmin/Desktop/SampleData/reporting-output

**Config File (Optional)**
/path/to/file

**Command Line Output**
1) /home/bcadmin/Desktop/SampleData/reporting-output/reports/bc_format_bargraph.pdf
2) /home/bcadmin/Desktop/SampleData/reporting-output/reports/format_table.pdf
3) /home/bcadmin/Desktop/SampleData/reporting-output/reports/FiwalkReport.pdf
4) /home/bcadmin/Desktop/SampleData/reporting-output/reports/FiwalkDeletedFiles.pdf
5) /home/bcadmin/Desktop/SampleData/reporting-output/reports/BeReport.pdf

Generating Excel report /home/bcadmin/Desktop/SampleData/reporting-output/reports/fiwalk-output.xml.xlsx

>> Success!!! BitCurator Reports generated in the directory:
/home/bcadmin/Desktop/SampleData/reporting-output/reports
Various Specialized BitCurator Reports

---

**File Format Table**

<table>
<thead>
<tr>
<th>Format</th>
<th>Short Form</th>
<th>Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>AST</td>
<td>AST</td>
<td>6</td>
</tr>
<tr>
<td>CSV</td>
<td>CSV</td>
<td>1</td>
</tr>
<tr>
<td>DOC</td>
<td>Doc</td>
<td>1</td>
</tr>
<tr>
<td>PDF</td>
<td>PDF</td>
<td>1</td>
</tr>
<tr>
<td>TXT</td>
<td>TXT</td>
<td>1</td>
</tr>
<tr>
<td>ZIP</td>
<td>Zip</td>
<td>1</td>
</tr>
</tbody>
</table>

**Disk Image: sampleimage:E01 File counts (by format)**

- AST: 6
- CSV: 1
- DOC: 1
- PDF: 1
- TXT: 1
- ZIP: 1

---

**Report: File System Statistics and Files**

- Syntax File - Copy-Matter: Syn_m
- empty ZIP archive file: 1
- empty XML file: 1
- x86 boot-sector, code offset: 642, D Searches: 0, file < 64 Boot-sector (649): 1
- x86_dos: 1
- x86_dos: 1
- x86_dos: 1
- x86_dos: 1
Look Familiar? Filesystem Metadata - Output from fiwalk*

```
<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</ctime>
  <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>
```

*Developed by Simson Garfinkel*
# Specialized BitCurator Reports

<table>
<thead>
<tr>
<th>File</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>bc_format_bargraph.pdf</td>
<td>histogram of file formats found on the volume</td>
</tr>
<tr>
<td>bulk_extractor_report.pdf</td>
<td>high-level overview of feature locations on disk</td>
</tr>
<tr>
<td>fiwalk_deleted_files.pdf</td>
<td>shows paths to any deleted materials found in a given partition</td>
</tr>
<tr>
<td>fiwalk-output.xml.xlsx</td>
<td>Excel converted DFXML output (file system metadata)</td>
</tr>
<tr>
<td>fiwalk_report.pdf</td>
<td>high-level overview of file system characteristics</td>
</tr>
<tr>
<td>format_table.pdf</td>
<td>long-form file format names for formats shown in bar graph</td>
</tr>
<tr>
<td>premis.xml</td>
<td>PREMIS preservation metadata</td>
</tr>
</tbody>
</table>
Maybe Less Familiar? PREMIS (Preservation) Metadata Generated from Running BitCurator Tools – Recorded as PREMIS Events
Exporting Files from a Disk Image
Nautilus Scripts

- Scripts that can be run using Nautilus (GNOME file manager)
- Most provide more convenient access (right click and menu selection) to functions performed by applications that could also be run directly
Right Click on File or Directory and Calculate MD5
The MD5 hash of the selected file:

```
keb2622125be1231b0fc9babee27942d /home/bcadmin/Pictures/bitcurator-grub.png
```

"bitcurator-grub.png" selected (43.3 kB)
Quick Access to a Hex View:
ewfinfo 20130416

Acquire information
  Acquisition date: Wed Jan 19 12:09:18 2011
  System date: Wed Jan 19 12:09:18 2011
  Operating system used: Linux
  Software version used: 20100226
  Password: N/A

EWF information
  File format: EnCase 6
  Sectors per chunk: 64
  Error granularity: 64
  Compression method: deflate
  Compression level: best compression
  Set identifier: 4eb6701d-6cf0-2f4a-a0c6-0cb5d5e20959

Media information
  Media type: fixed disk
  Is physical: yes
  Bytes per sector: 512
  Number of sectors: 2068480
  Media size: 1010 MiB (1059061760 bytes)

Digest hash information
  MD5: 9c0de6c8532d7a66ddcfc01861dfb6535
Quick Start Guide
Most recent version always available at:
http://wiki.bitcurator.net/

BitCurator
Quick Start Guide v0.9.16

Last updated: August 3, 2014
<table>
<thead>
<tr>
<th>Function</th>
<th>Tool(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify duplicate files</td>
<td>FSLint</td>
</tr>
<tr>
<td>Characterize files</td>
<td>FITS</td>
</tr>
<tr>
<td>Examine, copy and extract information from old Mac disks</td>
<td>HFSEExplorer</td>
</tr>
<tr>
<td>Package files for storage and/or transfer</td>
<td>BagIt (Java) library</td>
</tr>
<tr>
<td>Scan for viruses</td>
<td>ClamTK</td>
</tr>
<tr>
<td>Read contents of Microsoft Outlook PST files</td>
<td>readpst</td>
</tr>
<tr>
<td>Examine embedded header information in images</td>
<td>pyExifToolGUI</td>
</tr>
<tr>
<td>Generate images of problematic disks or particular disk types</td>
<td>dd, dcfldd, ddrescue, cdrdao (in addition to Guymager)</td>
</tr>
<tr>
<td>Identify files that are partially similar but not identical</td>
<td>sdhash, ssdeep</td>
</tr>
</tbody>
</table>
Considerable Uptake

- 118 members of the BitCurator Users mailing list
- 640 individuals who have participated in BitCurator-related events (workshops, tutorials, hackathons)
- Numerous publications and reports by library/archives students and professionals about their testing and use of the software
- 934 Twitter followers (whatever that means)
Open Source Development Strategy

• Rapid development with numerous iterations based on several channels of user feedback
• Code released under GPL, v3 (perhaps moving to Apache) – available through GitHub
• Existing code incorporated is generally GPL or public domain (government products)
• Packaging elements of the code to be integrated into other environments (e.g. Archivematica)
• Significant community engagement, including work of Community Lead (Porter Olsen)
BitCurator Consortium

• Continuing home for hosting, stewardship and support of BitCurator tools and associated user engagement
• Administrative home: Educopia Institute
• Funding based on membership dues
• Institutions as members, with two categories of membership: Charter and General
• Software and documentation will continue to be free and open source, but membership provides further benefits (e.g. support, training, development priority)
DIMAC (Disk Image Access for the Web)

- Developed by Sunitha Misra and Kam Woods
- To dynamically navigate and download contents of a disk image, without having to download or mount the full image
- See: https://github.com/kamwoods/dimac
- Demo at: http://www.youtube.com/watch?v=BwiWFqxYzQ8

APPLYING FORENSICS TO PRESERVING THE PAST: CURRENT ACTIVITIES AND FUTURE POSSIBILITIES

• Organizers: Cal Lee, Jeremy Leighton John, Susan Thomas

• To be held at Digital Libraries 2014, London, September, 8-12, 2014

• One-day event, split across an afternoon and following morning (Sept 11-12)

• Short papers and talks, group discussion and formation of steps for further action
Thank You!

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