The BitCurator Environment and BitCurator Access Tools

Christopher (Cal) Lee
UNC School of Information and Library Science

Harvard Email Archiving Stewardship Tools (EAST) Workshop
March 2, 2016
Cambridge, MA

BitCurator Access
The Andrew W. Mellon Foundation
BitCurator

- 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
- Porter Olsen, Community Lead
- Alex Chassanoff, Project Manager
- Sunitha Misra, Software Developer (UNC)
- Kyle Bickoff, GA (MITH)
- Amanda Visconti, 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
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
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/
Most of the tasks we cover in this class are explained in the Quick Start Guide. The most recent version always available at: http://wiki.bitcurator.net/
BitCurator Consortium

- Continuing home for hosting, stewardship and support of BitCurator (and BitCurator Access) 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
- The most important member benefit is assurance that the BitCurator software will persist in future years

http://www.bitcurator.net/bitcurator-consortium/
Membership is open to libraries, archives, museums, and other institutions worldwide that seek a collaborative community within which they may explore and apply forensics approaches and solutions to their digital collections.

How to Use BitCurator

- Acquire and process digital collections.
- Maintain the original order of digital materials.
- Survey the extent and composition of digital collections.
- Redact personally identifiable information.
- Extract technical and preservation metadata.
- Package digital materials for archival storage.

Learn more about getting started.

Member Benefits

- Use of the members-only BCC mailing list and help desk
- Access to the members-only videos and documentation
- Prioritized requests for BitCurator feature development
- Opportunities to serve on the BCC committees
- Voting rights for community governance
- Professional development opportunities
- Discounts for events including the BitCurator User Forum®
BitCurator-Supported Workflow

Acquisition
- Imaging (image, gyrimag)
- Source media
- AFF packaged image
- Capture and image metadata
- Log, device info
- Additional supported output formats
  - Split raw
  - E01

Analysis (accounts, filesystem activity)
- Filesystem metadata extraction (findwalk)
- File <-> disk block map
- Image analysis (private, sensitive info)
  - Bulk Extractor
  - Human-readable reporting scripts / plug-in
  - Annotated feature file
  - Reports: file distribution, sensitive info, hashes, etc.
  - Exported as PDF, .xlsx, plain text as appropriate

Analysis (file similarity, deduplication)
- reg2xml (and similar)
- XML dump of registry
- sdhash
- Reports: user accounts, device usage, environment
  - Exported as PDF, .xlsx, plain text as appropriate

Redaction of disk image
- Ruleset (patterns, hashes, etc) describing what to redact
- Python redaction scripts

Redaction
- Metadata export
- Python (xml) export scripts
- METS, MODS, EAD as required by user

Reporting
- Acquisition
- Reporting
- Redaction
- Metadata Export

See: http://bitcurator.net
Creating a Disk Image using Guymager
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: 9c0de6c8532d7a66ddcf01861dfb6535
Two Ways to Interact with Disk Images

• Mount them like regular drives:
  – Disk Utility in Mac OS X (for ISO images)
  – ewfmount
  – MagicDisc (for ISO images)
  – OSFMount
  – BitCurator (mounting scripts built into the environment)

• Inspect them as forensic objects
  – FTK Imager
  – The Sleuth Kit (TSK)
  – BitCurator (Disk Image Access tool)
Mounting a Forensically Packaged Disk Image in the BitCurator Environment
Identifying Potentially Sensitive Data using Bulk Extractor - Scanning Options

See: http://www.forensicswiki.org/wiki/Bulk_extractor
SSNs and DOBs identified in large PST collection using bulk_extractor
Exporting Filesystem Content Using fiwalk

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

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

Command Line Output

>> Generating XML File
/home/bcadmin/Desktop/SampleData/sampleimage.xml

>> Invoking command 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
- Provenance metadata - about the disk capture process
- Technical metadata - about the specific storage partition(s) on the disk
Exporting Filesystem Metadata - Output from fiwalk (XML)

```
<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>
  <crtime>1257729636</crtime>
  <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>
```
This is the schema repository for Digital Forensics XML, version 1.1.1.

If you intend to use the dfxml.xsd file as a DFXML document validator, note that you will also need to download two accompanying .xsd files under the "ref" directory. The easiest way to do this is by downloading the repository as a Git clone, or by downloading the [zip archive](https://github.com/dfxml/working-group/dfxml_schema) from the Github page.

To report issues, questions, or feature requests, please either:

- File a Github issue, seeing first if it is already filed, [here](https://github.com/dfxml/working-group/dfxml_schema).
- Email the [dfxml@nist.gov](mailto:dfxml@nist.gov) mailing list. If you wish to join the mailing list, send an email to [dfxml-subscribe@nist.gov](mailto:dfxml-subscribe@nist.gov) (no subject or message body is necessary), and a moderator will grant access.

[https://github.com/dfxml-working-group/dfxml_schema](https://github.com/dfxml-working-group/dfxml_schema)
PREMIS (Preservation) Metadata Generated from Running BitCurator Tools – Recorded as PREMIS Events

```xml
<?xml version="1.0" encoding="UTF-8"?>
<premis xmlns="info:lc/xmlns/premis-v2" version="2.0" xsi="http://www.w3c.org/2001/XMLSchema-instance">
  <object>
    <objectIdentifier>
      <objectIdentifierType>6d4e39d6-b8dc-11e3-a80f-080027f8dfea</objectIdentifierType>
      <objectIdentifierValue>/home/bcadmin/Desktop/terry-work-usb-2009-12-11.E01</objectIdentifierValue>
    </objectIdentifier>
  </object>
  <event>
    <eventIdentifier>
      <eventIdentifierType>0d4ea1ce-b8dc-11e3-a80f-080027f8dfea</eventIdentifierType>
      <eventIdentifierValue>E01/home/bcadmin/Desktop/terry-work-usb-2009-12-11.E01</eventIdentifierValue>
    </eventIdentifier>
    <eventType>Capture</eventType>
    <eventDateTime>Wed Jan 19 12</eventDateTime>
    <eventOutcomeInformation>
      <eventOutcome>E01</eventOutcome>
      <eventOutcomeDetailVersion>20100226</eventOutcomeDetailVersion>
      , Image size: 512</eventOutcomeInformation>
  </event>
  <event>
    <eventIdentifier>19882604-b8dc-11e3-93f0-080027f8dfea</eventIdentifier>
    <eventIdentifierValue>bulk_extractor -o /home/bcadmin/Desktop/demo1 /home/bcadmin/Desktop/terry-work-usb-2009-12-11.E01</eventIdentifierValue>
    <eventType>Feature Stream Analysis</eventType>
    <eventDateTime>2014-03-31T13:49:59Z</eventDateTime>
    <eventOutcomeInformation>
      <eventOutcome>Bulk Extractor Output</eventOutcome>
      <eventOutcomeDetailVersion>1.4.4</eventOutcomeDetailVersion>
    </eventOutcomeInformation>
  </event>
</premis>
```
Various Specialized BitCurator Reports
### Other Functionality to Meet Identified User Needs:

<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, FIDO</td>
</tr>
<tr>
<td>Scan for viruses</td>
<td>ClamTK</td>
</tr>
<tr>
<td>Examine, copy and extract information from old Mac disks</td>
<td>HFS Utilities (including HFS Explorer)</td>
</tr>
<tr>
<td>Capture AV file metadata</td>
<td>MediaInfo, FFProbe</td>
</tr>
<tr>
<td>Extract text from older binary (.doc) Word files</td>
<td>antiword</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>Extract and analyze data from Windows Registry files</td>
<td>regripper</td>
</tr>
<tr>
<td>Identify files that are partially similar but not identical</td>
<td>sdhash, ssdeep</td>
</tr>
<tr>
<td>Package files for storage and/or transfer</td>
<td>BagIt (Java) library, Bagger</td>
</tr>
<tr>
<td>File preview (left-click on file then hit space bar)</td>
<td>gnome-sushi</td>
</tr>
<tr>
<td>Play and examine metadata from AV media files</td>
<td>VLC media player</td>
</tr>
<tr>
<td>Damaged/lost partition recovery</td>
<td>TestDisk</td>
</tr>
<tr>
<td>Damaged/lost file recovery</td>
<td>PhotoRec</td>
</tr>
<tr>
<td>Identify the filesystem on a disk</td>
<td>disktype</td>
</tr>
<tr>
<td>Index and search for keywords in documents</td>
<td>recoll</td>
</tr>
<tr>
<td>Find blacklist data by using hashes calculated from hash blocks</td>
<td>hashdb</td>
</tr>
<tr>
<td>Generate hashes of files and blocks</td>
<td>md5deep (more features than md5sum)</td>
</tr>
</tbody>
</table>
Command Line Operations – Open Up Many More Possibilities

- stringing tools together
- performing batch operations
- changing parameters from their default values
- using tools that are only available through the command line (no GUI)
Readpst
readpst albert_meyers_000_1_1.pst
Opening PST file and indexes...

Processing Folder "Deleted Items"
Processing Folder "meyers-a"
  *Personal folders* - 2 items done, 0 items skipped.
Processing Folder "ExMerge - Meyers, Albert"
  "meyers-a" - 1 items done, 0 items skipped.
Processing Folder "Sent Items"
Processing Folder "Inbox"
Processing Folder "Deleted Items"
Processing Folder "Contacts"
  "ExMerge - Meyers, Albert" - 4 items done, 0 items skipped.
  "Contacts" - 4 items done, 0 items skipped.
  "Sent Items" - 12 items done, 0 items skipped.
  "Inbox" - 20 items done, 0 items skipped.

bcadmin@ubuntu:~Desktop/sample-data/pst-test$ readpst albert_meyers_000_1_1.pst
From "kourtneynelson@hotmail.com" Sat Jan 26 00:53:56 2002
Date: Fri, 25 Jan 2002 21:53:56 -0800 (PST)
Message-ID: <HRFHDCGM943CFFVUOYADW03KROBOA@zlsvr22>

Subject: Suz, Katie and Kourtney's Housewarming Party
Return-Path: kourtneynelson@hotmail.com
X-Originating-IP: 192.168.110.237
X-OriginalArrivalTime: 25 Jan 2002 21:54:38.000000 (UTC) FILETIME=02BC8320:01C156EA
X-ZL-from: Suz Warjone, Kourtney Nelson and Katie Sullivan <<<Kourtneynelson@hotmail.com>>
X-ZL-to: Meyers, Albert <<<ENRON/OU=NA/CN=RECIPIENTS/CN=BMeyers>>
X-ZL-subject: Suz, Katie and Kourtney’s Housewarming Party
filename: bert meyers 6-25-02.PST
folder: \ExMerge - Meyers, Albert\Inbox
Status: RO
Cc: X-Libpst-forensic-bcc:
MIME-Version: 1.0
Content-Type: multipart/mixed;
boundary="--boundary-LibPST-Iamunique-1068502869_-
Content-Type: text/plain;
charset="us-ascii"

---boundary-LibPST-Iamunique-1068502869_--
End User Access Scenarios

• Virtualization and emulation
• Mounting the original filesystem
• Accessing (but not mounting) disk images using forensics software
• Remote, dynamic access to disk image contents
• Cross-drive analysis
Two-year project (October 1, 2014 – September 30, 2016) at School of Information and Library Science, University of North Carolina at Chapel Hill

Funded by Andrew W. Mellon Foundation

Developing open-source software to support access to disk images. Core areas of focus:

- Tools and reusable libraries to support web access services for disk images
- Analyzing contents of file systems and associated metadata
- Redacting complex born-digital objects (disk images)
- Emulated access to data from disk images
BitCurator Access Team

Cal Lee – Principal investigator

Kam Woods - Technical Lead and Co-PI

Alex Chassanoff - Project Manager

Sunitha Misra - Software Developer
BitCurator Access Advisory Board

- Geoffrey Brown, Indiana University
- Mark Evans, History Associates
- Erika Farr, Emory University
- Matthew Farrell, Duke University
- Brad Glisson, University of South Alabama
- Matthew Kirschenbaum, Maryland Institute for Technology in the Humanities
- Susan Malsbury, New York Public Library
- Don Mennerich, New York University
- Klaus Rechert, University of Freiburg
- Kari Smith, Massachusetts Institute of Technology
- Bradley Westbrook, ArchivesSpace
- Doug White, National Institute of Standards and Technology
- Carl Wilson, Open Planets Foundation
Automated Redaction and Access Options

Option A: Redact from live image in EaaS via copy-on-write overlay

Option B: EaaS access to previously redacted image

Option C: Browse non-live file system with redaction mask

EaaS = Emulation-as-a-Service. [http://bw-fla.uni-freiburg.de/](http://bw-fla.uni-freiburg.de/)
Automated Redaction and Access Options

Option A: Redact from live image in EaaS via copy-on-write overlay

Option B: EaaS access to previously redacted image

Option C: Browse non-live file system with redaction mask

EaaS = Emulation-as-a-Service. [http://bw-fla.uni-freiburg.de/](http://bw-fla.uni-freiburg.de/)
BCA (BitCurator Access) Web Tools

- Integrates digital forensics software libraries and lightweight web-services tools
- Drop disk images in a local or network-accessible location, start up the service, and start browsing
- Most analysis runs server-side (via Sleuthkit and DFXML Python bindings, among others)
- Service is database-agnostic (we’re using postgres)
- Automatic metadata production – Digital Forensics XML (DFXML), PREMIS, others)

https://github.com/kamwoods/bca-webtools

The bca-webtools application provides access to forensically-packaged and raw disk images. Supported file systems include FAT16, FAT32, NTFS, HFS+, and EXT2/3/4. Click on 'Browse' to navigate through the file system(s) within the disk image, or 'Download' to download the complete disk image.

<table>
<thead>
<tr>
<th>Image Name</th>
<th>Info</th>
<th>Browse</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>charlie-work-usb-2009-12-11.E01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>terry-work-usb-2009-12-11.E01</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# bca-webtools - Admin Tools

- Build Image Table
- Build DFXML Table
- Build All Tables
- Drop Image Table
- Drop DFXML Table
- Drop All Tables
- Generate Index
- Clear Index
- Show Image Matrix

## Image Matrix

<table>
<thead>
<tr>
<th>Index</th>
<th>Image name</th>
<th>Image DB?</th>
<th>DFXML DB?</th>
<th>Indexed?</th>
<th>Add Table</th>
<th>Delete Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>charlie-work-usb-2009-12-11.E01</td>
<td>True</td>
<td>False</td>
<td>True</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>terry-work-usb-2009-12-11.E01</td>
<td>True</td>
<td>False</td>
<td>True</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Admin
<table>
<thead>
<tr>
<th>d/r</th>
<th>Filename</th>
<th>Size</th>
<th>Last Modified</th>
<th>Deleted?</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>AttrDef</td>
<td>2560</td>
<td>2009-11-20T17:38:09Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>BadClus</td>
<td>0</td>
<td>2009-11-20T17:38:09Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>Bitmap</td>
<td>32320</td>
<td>2009-11-20T17:38:09Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>Boot</td>
<td>8192</td>
<td>2009-11-20T17:38:09Z</td>
<td>No</td>
</tr>
<tr>
<td>d</td>
<td>Extend</td>
<td>552</td>
<td>2009-11-20T17:38:09Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>LogFile</td>
<td>7405568</td>
<td>2009-11-20T17:38:09Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>MFT</td>
<td>262144</td>
<td>2009-11-20T17:38:09Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>MFTMirr</td>
<td>4096</td>
<td>2009-11-20T17:38:09Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>Secure</td>
<td>0</td>
<td>2009-11-20T17:38:09Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>UpCase</td>
<td>131072</td>
<td>2009-11-20T17:38:09Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>Volume</td>
<td>0</td>
<td>2009-11-20T17:38:09Z</td>
<td>No</td>
</tr>
<tr>
<td>d</td>
<td>.</td>
<td>56</td>
<td>2009-12-03T21:17:01Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>01.zip</td>
<td>108438</td>
<td>2009-11-24T21:21:16Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>astronaut.jpg</td>
<td>713418</td>
<td>2009-11-24T21:33:33Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>astronaut1.jpg</td>
<td>722717</td>
<td>2009-11-24T21:43:42Z</td>
<td>No</td>
</tr>
<tr>
<td>d</td>
<td>Email</td>
<td>56</td>
<td>2009-12-10T22:27:55Z</td>
<td>No</td>
</tr>
<tr>
<td>d</td>
<td>Immortality</td>
<td>56</td>
<td>2009-11-24T21:55:45Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>invspace2.exe</td>
<td>1291720</td>
<td>2009-11-19T18:42:25Z</td>
<td>No</td>
</tr>
<tr>
<td>r</td>
<td>microscope.jpg</td>
<td>136274</td>
<td>2009-11-24T21:27:51Z</td>
<td>No</td>
</tr>
</tbody>
</table>
BitCurator, BitCurator Consortium and BitCurator Access Resources

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

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

BitCurator Access Project and Products
http://access.bitcurator.net/

Twitter: @bitcurator