Bringing Bits to the User: BitCurator and BitCurator Access

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

Coalition for Networked Information (CNI) Membership Meeting
December 14-15, 2015
Washington, DC
What are we to do with this stuff?

Goals 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

| Provenance | • Reflect “life history” of records  
• Records from a common origin or source should be managed together as an aggregate unit |
| Original Order | Organize and manage records in ways that reflect their arrangement within the creation/use environment |
| Chain of Custody | • “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”\(^2\) |

Bit digital is different.

See:
<table>
<thead>
<tr>
<th>Level</th>
<th>Label</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Aggregation of objects</td>
<td>Set of objects that form an aggregation that is meaningful encountered as an entity</td>
</tr>
<tr>
<td>7</td>
<td>Object or package</td>
<td>Object composed of multiple files, each of which could also be encountered as individual files</td>
</tr>
<tr>
<td>6</td>
<td>In-application rendering</td>
<td>As rendered and encountered within a specific application</td>
</tr>
<tr>
<td>5</td>
<td>File through filesystem</td>
<td>Files encountered as discrete set of items with associate paths and file names</td>
</tr>
<tr>
<td>4</td>
<td>File as “raw” bitstream</td>
<td>Bitstream encountered as a continuous series of binary values</td>
</tr>
<tr>
<td>3</td>
<td>Sub-file data structure</td>
<td>Discrete “chunk” of data that is part of a larger file</td>
</tr>
<tr>
<td>2</td>
<td>Bitstream through I/O equipment</td>
<td>Series of 1s and 0s as accessed from the storage media using input/output hardware and software (e.g. controllers, drivers, ports, connectors)</td>
</tr>
<tr>
<td>1</td>
<td>Raw signal stream through I/O equipment</td>
<td>Stream of magnetic flux transitions or other analog electronic output read from the drive without yet interpreting the signal stream as a set of discrete values (i.e. not treated as a digital bitstream that can be directly read by the host computer)</td>
</tr>
<tr>
<td>0</td>
<td>Bitstream on physical medium</td>
<td>Physical properties of the storage medium that are interpreted as bitstreams at Level 1</td>
</tr>
</tbody>
</table>
## Interaction Examples

### Level

<table>
<thead>
<tr>
<th>Aggregation of objects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

### Gazette.govt.nz

#### Search Bar
- Search for notices:
  - Enter a keyword
  - Advanced search
  - Find a specific notice:
    - Year
    - Page number
    - Notice number
    - Find

#### Browse Gazette Issues

- **2014**

<table>
<thead>
<tr>
<th>Issue</th>
<th>Title</th>
<th>File Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ No. 124</td>
<td>Special: 2014 General Election - Election of List Candidates.</td>
<td>PDF (68kb)</td>
</tr>
<tr>
<td>+ No. 123</td>
<td>Special: 2014 General Election Amended Official Result for the Te Tai Tokerau Electoral District.</td>
<td>PDF (68kb)</td>
</tr>
<tr>
<td>+ No. 122</td>
<td>Principal Edition, 9 October 2014</td>
<td>PDF (261kb)</td>
</tr>
<tr>
<td>+ No. 121</td>
<td>Customs Edition, 7 October 2014</td>
<td>PDF (55kb)</td>
</tr>
<tr>
<td>+ No. 120</td>
<td>Special: 2014 General Election Results of the Official Count, 4 October 2014.</td>
<td>PDF (206kb)</td>
</tr>
<tr>
<td>+ No. 119</td>
<td>Principal Edition, 2 October 2014</td>
<td>PDF (414kb)</td>
</tr>
<tr>
<td>+ No. 118</td>
<td>Supplement: Financial Markets Authority: Authorised Futures Dealers Notices, 30 September 2014</td>
<td>PDF (89kb)</td>
</tr>
</tbody>
</table>
Interaction Examples

Level
Aggregation of objects

Object or package

In-application rendering
File through filesystem
File as “raw” bitstream
Sub-file data structure
Bitstream through I/O equipment
Raw signal stream through equipment
Bitstream on physical medium

Browse Gazette issues

- 2014
  + No. 123 Special: 2014 General Election Amended Official Result for the Te Tai Tokerau Electoral District.
  + No. 122 Principal Edition, 9 October 2014
  - No. 120 Special: 2014 General Election Results of the Official Count, 4 October 2014.

<table>
<thead>
<tr>
<th>Date</th>
<th>Title</th>
<th>Type</th>
<th>Act</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 OCT 2014</td>
<td>2014 General Election</td>
<td>Authorities/Other Agencies of State</td>
<td>Electoral Act</td>
</tr>
</tbody>
</table>

+ No. 119 Principal Edition, 2 October 2014
## Interaction Examples

<table>
<thead>
<tr>
<th>Level</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregation of objects</td>
<td></td>
</tr>
<tr>
<td>Object or package</td>
<td></td>
</tr>
</tbody>
</table>

### In-application rendering

- File through filesystem
- File as “raw” bitstream
- Sub-file data structure
- Bitstream through I/O equipment
- Raw signal stream through I/O equipment
- Bitstream on physical medium
File System

- Access controls
- File names & identifiers
- File size (length)
- Where to find files in storage (sectors and clusters)
- MAC times
  - Modified – when the content was last changed
  - Accessed – time file was last accessed (by person or software)
  - Changed – last time metadata changed
  - Created – (implemented inconsistently, if at all, across different file systems)
Interaction Examples

Level

Aggregation of objects
Object or package
In-application rendering
File through filesystem
File as “raw” bitstream
Sub-file data structure
Bitstream through I/O equipment
Raw signal stream through I/O equipment
Bitstream on physical medium
Hex Dump

- A more compact and more humanly readable way of conveying a stream of bits
- Uses hexadecimal notation
  - Each character represents one of 16 possible values (0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F)
  - Conveniently, a series of two characters represented in hexadecimal can represent exactly one byte ($2^8 = 256$ possible values) of data, because $16^2 = 256$
- Hex dumps from computer’s memory often used for debugging or reverse engineering software and for data recovery
In the BitCurator environment:
Interaction Examples

Level

Aggregation of objects

Object or package

In-application rendering

File through filesystem

File as “raw” bitstream

Sub-file data structure

Bitstream through I/O equipment

Raw signal stream through equipment

Bitstream on physical medium
In BitCurator environment: Right Click on File or Directory and Calculate MD5
Calculate MD5 (Files and Directories)

Please choose the way you want the MD5 hash to be presented:
(1 file(s) selected)

- Handling
  - Display on screen
  - Save to file (the selected filename + .md5 extension)

Cancel  OK

"bitcurator-grub.png" selected (43.3 kB)
The MD5 hash of the selected file:

```
keb2622125be1231b0fc9babee27942d /home/bcadmin/Pictures/bitcurator-grub.png
```
# Interaction Examples

## Level

### Aggregation of objects

Object or package

### In-application rendering

File through filesystem

### File as “raw” bitstream

Sub-file data structure

### Bitstream through I/O equipment

Raw signal stream through equipment

### Bitstream on physical medium

---

**WinZip - Management-curriculum.zip**

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Modified</th>
<th>Size</th>
<th>Ratio</th>
<th>Packed</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>.rels</td>
<td>XML Document</td>
<td>1/1/1980 12:00 AM</td>
<td>590</td>
<td>59%</td>
<td>243</td>
<td>_rels\</td>
</tr>
<tr>
<td>[Content_Types].xml</td>
<td>XML Document</td>
<td>1/1/1980 12:00 AM</td>
<td>1,445</td>
<td>74%</td>
<td>370</td>
<td>\docProps\</td>
</tr>
<tr>
<td>app.xml</td>
<td>XML Document</td>
<td>1/1/1980 12:00 AM</td>
<td>1,041</td>
<td>50%</td>
<td>519</td>
<td>\docProps\</td>
</tr>
<tr>
<td>core.xml</td>
<td>XML Document</td>
<td>1/1/1980 12:00 AM</td>
<td>633</td>
<td>48%</td>
<td>331</td>
<td>\docProps\</td>
</tr>
<tr>
<td>document.xml</td>
<td>XML Document</td>
<td>1/1/1980 12:00 AM</td>
<td>34,242</td>
<td>90%</td>
<td>3,454</td>
<td>word\</td>
</tr>
<tr>
<td>fontTable.xml</td>
<td>XML Document</td>
<td>1/1/1980 12:00 AM</td>
<td>950</td>
<td>72%</td>
<td>265</td>
<td>word_rels\</td>
</tr>
<tr>
<td>fontTable.rls</td>
<td>XML Document</td>
<td>1/1/1980 12:00 AM</td>
<td>1,831</td>
<td>72%</td>
<td>510</td>
<td>word\</td>
</tr>
<tr>
<td>numbering.xml</td>
<td>XML Document</td>
<td>1/1/1980 12:00 AM</td>
<td>6,306</td>
<td>87%</td>
<td>845</td>
<td>word\</td>
</tr>
<tr>
<td>settings.xml</td>
<td>XML Document</td>
<td>1/1/1980 12:00 AM</td>
<td>1,833</td>
<td>57%</td>
<td>791</td>
<td>word\</td>
</tr>
<tr>
<td>styles.xml</td>
<td>XML Document</td>
<td>1/1/1980 12:00 AM</td>
<td>15,692</td>
<td>87%</td>
<td>2,071</td>
<td>word\</td>
</tr>
<tr>
<td>theme1.xml</td>
<td>XML Document</td>
<td>1/1/1980 12:00 AM</td>
<td>6,992</td>
<td>76%</td>
<td>1,686</td>
<td>word_theme\</td>
</tr>
<tr>
<td>webSettings.xml</td>
<td>XML Document</td>
<td>1/1/1980 12:00 AM</td>
<td>260</td>
<td>28%</td>
<td>187</td>
<td>word\</td>
</tr>
</tbody>
</table>

Selected 1 file, 34KB

Total 12 files, 71KB
Example of EXIF Metadata from a JPEG File (Generated Using exiftool*)

--- ExifTool ----
ExifTool Version Number : 9.38
ExifTool Version Number : 9.38
System ----
File Name : IMG_20130823_151811.jpg
File Size : 1785 kB
File Modification Date/Time : 2013:08:23 16:36:44-04:00
File Access Date/Time : 2013:10:14 17:13:02-04:00
File Creation Date/Time : 2013:08:23 16:36:44-04:00
File Permissions : rw-rw-rw-
---- File ----
File Type : JPEG
MIME Type : image/jpeg
Exif Byte Order : Big-endian (Motorola, MM)
Image Width : 2592
Image Height : 1944
Encoding Process : Baseline DCT, Huffman coding
Bits Per Sample : 8
Color Components : 3
YCbCr Sub Sampling : YCbCr4:2:0 (2:2)
---- GPS ----
GPS Img Direction : 83
GPS Img Direction Ref : Magnetic North
GPS Latitude Ref : North
GPS Longitude : 35 deg 55' 2.24"
GPS Longitude Ref : West
GPS Altitude Ref : Above Sea Level
GPS Altitude : 0 m
GPS Time Stamp : 19:18:06
GPS Processing Method : NETWORK
GPS Date Stamp : 2013:08:23
---- IFD0 ----
Orientation : Unknown (0)
Camera Model Name : Galaxy Nexus
Modify Date : 2013:08:23 15:18:11
YCbCr Positioning : Centered
Y Resolution : 72
Resolution Unit : inches
X Resolution : 72
Make : Samsung
---- ExifIFD ----
Create Date : 2013:08:23 15:18:11
Date/Time Original : 2013:08:23 15:18:11
Exif Version : 0220
Flash Energy : 0
Image Unique ID : OAEL01
Exposure Time : 1/17
ISO : 125, 0, 0
Scene Type : Directly photographed
Exposure Index : undef
Components Configuration : Y, Cb, Cr, -
F Number : 2.8
Compressed Bits Per Pixel : 0
Sensing Bits Per Sample : One-chip color area
Exposure Program : Aperture-priority AE
Aperture Value : 2.6
Brightness Value : 0
Subject Distance Range : Unknown
Shutter Speed Value : 1/15
Subject Distance : 0 m
Saturation : Normal
Color Space : sRGB
Contrast : Normal
Metering Mode : Multi-spot
Flashpix Version : 
Exposure Compensation : 0
Exif Image Width : 2592
Max Aperture Value : 2.6
Sharpness : Normal
Exif Image Width : 2592
Focal Length : 3.4 mm
Digital Zoom Ratio : 1
Light Source : Fluorescent
Scene Capture Type : Standard
Flash : Off, Did not fire
Custom Rendered : Custom
White Balance : Auto
Exposure Mode : Auto
---- IFD1 ----
Compression : JPEG (old-style)
Image Width : 160
Image Height : 120
Thumbnail Offset : 1239
Thumbnail Length : 7164
---- Composite ----
Aperture : 2.8
GPS Altitude : 0 m Above Sea Level
GPS Date/Time : 2013:08:23 19:18:06Z
GPS Latitude : 35 deg 55' 2.24" N
GPS Longitude : 79 deg 2' 57.55" W
GPS Position : 35 deg 55' 2.24" N, 79 deg 2' 57.55" W
Image Size : 2592x1944
Shutter Speed : 1/17
Focal Length : 3.4 mm
Light Value : 6.7

*http://www.sno.phy.queensu.ca/~phil/exiftool/ (Also available through the BitCurator environment)
FIND
THE ONE SONG
BEFORE YOU ENTER THE LIGHT
THE GLORY
LIKE A SUNSET
ONE SONG
TO REDEEM THIS EMPTY LIFE

TIME FLIES
AND THEN- NO NEED TO ENDURE ANYMORE
TIME DIES
(A knock on the door)

THE DOOR

08. LIGHT MY CANDLE

ROGER
WHAT'D YOU FORGET?

(HE opens the door. MIMI stands, with a candle.)
Interaction Examples

Level

Aggregation of objects

Object or package

In-application rendering

File through filesystem

File as “raw” bitstream

Sub-file data structure

Bitstream through I/O equipment

Raw signal stream through I/O equipment

Bitstream on physical medium
Level
Aggregation of objects
Object or package
In-application
File through filesystem
File as "raw" bitstream
Sub-file data structure
Bitstream through I/O equipment
Raw signal stream through I/O equipment
Bitstream on physical medium

Interaction Examples
## Interaction Examples

### Level

<table>
<thead>
<tr>
<th>Aggregation of objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object or package</td>
</tr>
<tr>
<td>In-application rendering</td>
</tr>
<tr>
<td>File through filesystem</td>
</tr>
<tr>
<td>File as “raw” bitstream</td>
</tr>
<tr>
<td>Sub-file data structure</td>
</tr>
<tr>
<td>Bitstream through I/O equipment</td>
</tr>
<tr>
<td>Raw signal stream through I/O equipment</td>
</tr>
<tr>
<td><strong>Bitstream on physical medium</strong></td>
</tr>
</tbody>
</table>

---

Three Complicating Factors:

1. Medium Failure / Bit Rot

2. Obsolescence

3. Volatility
Bit Rot

- Preventing measures can help (proper storage and handling), but bits on a given medium will eventual flip or become unreadable

- Bit rot and advantages of newer media both call for periodic refreshing and reformatting

- Ensuring the **integrity of the bit stream** when transferring from one medium to another is extremely important


Obsolescence

“Obsolete power corrupts obsoletely.”

- Ted Nelson

The technology associated with interpreting the representation at each of the layers can change or become less available
Order of Volatility

- Some types of data change much more quickly and often than others
- Important to recognize in order to recover data from a computer system or media, while ensuring that actions don’t make irreversible changes to their record characteristics
- Example: If the contents of the browser cache are important to you, capture the cache before using the browser
Digital Forensics in Libraries, Archives and Museums (LAMs)

In recent years, LAMs have been 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 Digital Forensics Tools and Tasks for LAMs

- Existing digital forensics tools provide valuable functionality, but 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)
They’re using a lot of this stuff:
Here’s what it looks like in libraries and archives:
Stanford University Libraries and Academic Information Resources (SULAIR)
UNC School of Information and Library Science (Manning 213)
USB 3.5” Floppy Disk Drive
Still available new from online retailers, look for a drive that can read both 1.44 MB (HD) and 800 KB (DD) 3.5” diskettes. Most drives support HD diskettes in both PC and Mac format, but only support PC formatted DD diskettes. New units are still available for around $20.

5.25” Floppy Disk Drive
These units are no longer available new, but can still be purchased off of eBay for about $50. We recommend purchasing a number of drives as well as a floppy disk drive cleaning kit.

External USB 250MB Zip Drive
These units are available both new and used. We recommend the 250MB model as it is backwards compatible with the 100MB Zip disks. New units retail for around $200 and used units for around $50.

Device Side Data’s FC5025
The FC5025 is a controller card for 5.25” floppy disk drives that can be used as an internal or external—as seen here—interface. Device Side Data charges $55.25 per controller.

Wiebotech UltraDock Hardware Write Protector
This unit serves as both an interface with IDE and Serial ATA type hard disk drives and as a write protector. Because it is common for the OS to overwrite metadata on a hard drive, write protection ensures that no interactions of the archivist or researcher affects the integrity of the original media. Wiebotech charges $250 for the UltraDock Hardware Write Protector.

Porter Olsen, Building a Digital Curation Workstation with BitCurator (update)
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
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, US National Archives and Records Administration
- Naomi Nelson, Duke University
- Erin O’Meara, Gates Archive
- Michael Olson, Stanford University Libraries
- Gabriela Redwine, Beinecke, Yale University
- 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, Digital Public Library of America
- 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 LAM ingest and collection management environments
  - Provision of public access to the data
BitCurator Environment*

- Bundles, integrates and extends functionality of open source software
- 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
- We’ll be looking at specific BitCurator tools and functions later

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

- Acquisition
- Reporting
- Redaction
- Metadata Export

See: http://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.

Become a member now >

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.

How our members are using BitCurator

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®

Members

McMaster University
Penn State University
Massachusetts Institute of Technology
Duke University
The University of Maryland, MITH
Stanford University
Yale University
The University of Manchester Library
University of Colorado Boulder

A Growing Community

The BitCurator Consortium provides spaces for members to share documentation, develop their skills, and improve the BitCurator environment.

Membership is open >
So let’s now look at specific curation actions and processes…
Write Blocking – One-Way Streets for Data

• Ensures that data can be read from the device, but no bits can be changed
• Doesn’t just prevent changes conscious made by user but also changes made by the system
• Options for write blocking (in order of most to least certain to prevent writes to the drive):
  – Dedicated write blockers
  – Writing blocking tabs or settings on the device itself
  – Software-based write blocking

Image source: http://thinng.com/1555-one-way-sign-seat
Dedicated Hardware Write Blockers
5.25 Inch Floppy – If light can get through, it’s **not** write protected

3.5 Inch Floppy – If light can get through, it **is** write protected

Example of Software Write Blocking – Mounted Devices set to Read-Only by Default
Getting below the File System – Low-Level Copying

- Getting an “image” of a storage medium involves working at a level below the file system
  - Can get at file attributes and deleted files not visible through higher-level copy operations
- Most commonly used tool is dd (or variant) - UNIX program for low-level copying and conversion of data from a storage device
- More specialized tools for creating forensic images include:
  - FTK Imager
  - Guymager
  - Imaging utilities in commercial applications (e.g. EnCase)
Main Acquisition Interface for 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
Exporting Files from a Disk Image
Identifying Potentially Sensitive Data using Bulk Extractor - Scanning Options

See: http://www.forensicswiki.org/wiki/Bulk_extractor
Histogram of Email Addresses (Specific Instances in Context on Right)
<table>
<thead>
<tr>
<th>OUTLOOK-36</th>
<th>SSN:</th>
<th>DOB:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>JUNE 16, 1978</td>
</tr>
</tbody>
</table>

SSNs and DOBs identified in large PST collection using bulk_extractor
Generating BitCurator Reports

![BitCurator Reporting Tool interface](image)

- **Computer**
  - Home
  - Imaging Tools
  - Forensics Tools
  - Additional Tools
  - Trash

**BitCurator Reporting Tool**

- **Report Types**
  - Run All
  - Fiwalk XML
  - Annotated Features
  - Reports
  - File Access

- **Fiwalk XML File**
  - `/home/bcadmin/Desktop/SampleData/sampleimage.xml`

- **Annotated Feature Files Directory**
  - `/home/bcadmin/Desktop/SampleData/annotated-features`

- **Output Directory For Reports**
  - `/home/bcadmin/Desktop/SampleData/bc-reports`

- **Config File (optional)**
  - `/Path/To/file`

- **Command Line Output**

  1. `/home/bcadmin/Desktop/SampleData/reporting-output/reports/FiwalkReport.pdf`
  2. `/home/bcadmin/Desktop/SampleData/reporting-output/reports/FiwalkDeletedFiles.pdf`
  3. `/home/bcadmin/Desktop/SampleData/reporting-output/reports/BeReport.pdf`
  4. `/home/bcadmin/Desktop/SampleData/reporting-output/reports/be_format_bargraph.pdf`
  5. `/home/bcadmin/Desktop/SampleData/bc-reports/format_table.pdf`
  6. `/home/bcadmin/Desktop/SampleData/bc-reports/FiwalkReport.pdf`
  7. `/home/bcadmin/Desktop/SampleData/bc-reports/FiwalkDeletedFiles.pdf`
  8. `/home/bcadmin/Desktop/SampleData/bc-reports/beReport.pdf`

- **Generating Excel report**
  - `/home/bcadmin/Desktop/SampleData/bc-reports/sampleimage.xml.xlsx`

>> Success!!! Reports generated in the directory:
- `/home/bcadmin/Desktop/SampleData/bc-reports`
- Provenance metadata - about the disk capture process
- Technical metadata - about the specific storage partition(s) on the disk

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>4.0.2</version>
    <build_environment>
      <compiler>GCC 4.6</compiler>
      <library name="afflib" version="3.7.1"/>
      <library name="libewf" version="20130303"/>
    </build_environment>
    <execution_environment>
      <command_line>
        fiwalk -fX/home/bcadmin/Desktop/SampleData/sampleimage.xml /home/bcadmin/Desktop/SampleData/sampleimage.E01
      </command_line>
      <start_time>2013-03-12T00:08:28Z</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>
    <block_size>2048</block_size>
    <ftype>2048</ftype>
    <ftype_str>iso9660</ftype_str>
    <block_count>36839</block_count>
  </volume_offset>
</dfxml>
```
Operationalizing Original Order - Filesystem Metadata Output from fiwalk

*Developed by Simson Garfinkel*
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 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.
- Email the dfxml@nist.gov mailing list. If you wish to join the mailing list, send an email to dfxml-subscribe@nist.gov (no subject or message body is necessary), and a moderator will grant access.
High-Level view of Metadata Generation and Reporting

PREMIS (Preservation) Metadata Generated from Running BitCurator Tools – Recorded as PREMIS Events

```
<?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>6d4e30d6-b8dc-11e3-a80f-080027f8dfe4</objectIdentifierType>
      <objectIdentifierValue>/home/bcadmin/Desktop/terry-work-usb-2009-12-11.E01</objectIdentifierValue>
    </objectIdentifier>
  </object>
  <event>
    <eventIdentifier>
      <eventIdentifierType>0d4ea1ce-b8dc-11e3-a80f-080027f8dfe4</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>
      <eventOutcomeDetailDescription>Image size: 512</eventOutcomeDetailDescription>
    </eventOutcomeInformation>
  </event>
  <event>
    <eventIdentifier>19882604-b8dc-11e3-93f0-080027f8dfe4</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, dd rescue, 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>recall</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>
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
- Source media → Forensic disk image
- Create copy-on-write overlay
- Redact items for session via overlay
- Redacted image copy (raw or forensic repackage)
- Redaction script

Option B: EaaS access to previously redacted image
- Create copy-on-write overlay
- Redacted image copy (raw or forensic repackage)
- Redaction script

Option C: Browse non-live file system with redaction mask
- bca_webtools
- List of items to be redacted (annotated DFXML)
- Analyze with bulk_extractor and fiwalk

Automated Redaction and Access Options

Option A: Redact from live image in EaaS via copy-on-write overlay
- Redact items for session via overlay
- Create copy-on-write overlay
- Redacted image copy (raw or forensic repackage)
- Redaction script
- Forensic disk image
- Source media

Option B: EaaS access to previously redacted image
- Create copy-on-write overlay
- Redacted image copy (raw or forensic repackage)
- Redaction script
- Forensic disk image
- Source media

Option C: Browse non-live file system with redaction mask
- List of items to be redacted (annotated DFXML)
- Analyze with bulk_extractor and fiwalk
- bca_webtools

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

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

Submit

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>Add</td>
<td>Delete</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>Add</td>
<td>Delete</td>
</tr>
</tbody>
</table>

Admin
Browse directories and download files. Items marked "r" in the first column are regular files. Items marked "d" are directories.

<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>invsecr2.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>
Upcoming Events
CurateGear: Enabling the Curation of Digital Collections

January 14, 2016 – Friday Center, Chapel Hill, North Carolina

An interactive day-long event focused on digital curation tools and methods. See demonstrations, hear about the latest developments, and discuss application in professional contexts.

http://ils.unc.edu/digccurr/curategear2016.html
BitCurator User Forum 2016

Join BitCurator users from around the globe as we discuss how we are using the BitCurator software environment. Hosted by the BitCurator Consortium (BCC), this event will be grounded in the practical, real-world experiences of digital archivists and digital curation experts. Come prepared to discuss your current challenges, share emerging BitCurator integrations and workflows, and address the "now what" of handling your digital forensics outputs.

Date: 15 January 2016
Location: Pleasant Family Assembly Room, Wilson Library, University of North Carolina, Chapel Hill, North Carolina

Registration
General Registration - $30
Student Registration - $15
BitCurator Consortium Member Registration - Free
[Register Now]

Program
(details about speakers coming soon)

8:00 - 8:30am  Registration and Coffee
8:30 - 9:00am  Welcome and Introductions
9:00 - 10:15am Panel: Beyond Floppy Disks
Cultural heritage institutions are collecting a larger variety of media than ever before including external hard drives, computers, flash drives, cell phones, and a host of other devices. This panel will discuss the implications for archivists and curators and some strategies to handle this variety of media.

https://bitcuratorconsortium.org/events/buf2016
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