BITCURATOR FOR STATE ARCHIVES

Use the chat box at the right of the screen to tell us who you are, where you’re from, and who is participating with you today.

(To open the chat window, click on the CHAT icon in the upper right corner.)

Connect to the audio portion of the webinar through your phone line or through VoIP.
WELCOME!

Fall 2019 State Electronic Records Initiative webinars

• October: social media
• November: scripting for archivists
• December: developing processing workflows

Spring 2020 SERI webinars will begin in January

SERI ONLINE RECORDINGS

• Managing Digital Content Over Time
  • Identify
  • Select
  • Store
  • Protect
  • Manage
  • Provide access

• [https://archives.utah.gov/community/SERI/SERI-resources.html](https://archives.utah.gov/community/SERI/SERI-resources.html)
NEW SERI VIDEO

- FAQs on Bit Rot
  - https://www.youtube.com/user/StateArchivists
TODAY’S PRESENTER

Cal Lee
School of Information and Library Science
University of North Carolina at Chapel Hill
Some 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
## Some Fundamental 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”\(^2\)

---

But what about these?

<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

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

---

This page lists all the seed queries that are used for monitoring videos related to elections on YouTube. Clicking on a query will show all the results collected over several crawls. Total number of these results are also listed here for each query. The last column in the following table shows how many total results YouTube had for a given query during our latest crawl. Clicking on 'Setup' associated with a query will bring up an interface where the curator can specify what constitutes a "significant" change for a video of that query.

<table>
<thead>
<tr>
<th>#</th>
<th>Query</th>
<th>Setup</th>
<th>Total results so far</th>
<th>Max results on last crawl</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>election 2008</td>
<td>Setup</td>
<td>574</td>
<td>6150</td>
</tr>
<tr>
<td>2</td>
<td>US election 2008</td>
<td>Setup</td>
<td>349</td>
<td>795</td>
</tr>
<tr>
<td>3</td>
<td>United States election 2008</td>
<td>Setup</td>
<td>216</td>
<td>257</td>
</tr>
<tr>
<td>4</td>
<td>presidential election 2008</td>
<td>Setup</td>
<td>206</td>
<td>1820</td>
</tr>
<tr>
<td>5</td>
<td>campaign 2008</td>
<td>Setup</td>
<td>273</td>
<td>2530</td>
</tr>
<tr>
<td>6</td>
<td>decision 2008</td>
<td>Setup</td>
<td>168</td>
<td>142</td>
</tr>
<tr>
<td>7</td>
<td>Joe Biden</td>
<td>Setup</td>
<td>209</td>
<td>1080</td>
</tr>
<tr>
<td>8</td>
<td>Hillary Rodham Clinton</td>
<td>Setup</td>
<td>193</td>
<td>353</td>
</tr>
<tr>
<td>9</td>
<td>Christopher Dodd</td>
<td>Setup</td>
<td>267</td>
<td>815</td>
</tr>
<tr>
<td>10</td>
<td>John Edwards</td>
<td>Setup</td>
<td>902</td>
<td>7540</td>
</tr>
<tr>
<td>11</td>
<td>Mike Gravel</td>
<td>Setup</td>
<td>301</td>
<td>1210</td>
</tr>
<tr>
<td>12</td>
<td>Dennis Kucinich</td>
<td>Setup</td>
<td>229</td>
<td>1600</td>
</tr>
<tr>
<td>13</td>
<td>Barack Obama</td>
<td>Setup</td>
<td>861</td>
<td>9140</td>
</tr>
<tr>
<td>14</td>
<td>Bill Richardson</td>
<td>Setup</td>
<td>287</td>
<td>1100</td>
</tr>
<tr>
<td>15</td>
<td>Wesley Clark</td>
<td>Setup</td>
<td>191</td>
<td>375</td>
</tr>
<tr>
<td>16</td>
<td>Al Gore</td>
<td>Setup</td>
<td>613</td>
<td>4910</td>
</tr>
<tr>
<td>17</td>
<td>Tom Vilsack</td>
<td>Setup</td>
<td>89</td>
<td>68</td>
</tr>
<tr>
<td>18</td>
<td>Sam Brownback</td>
<td>Setup</td>
<td>254</td>
<td>404</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 I/O equipment

## Bitstream on physical medium

---

### ContextMiner Alpha 3.0

This page presents contextual information for a video captured over a number of days. Contextual information is defined as the information about a video that may change with time. Usually this information is contributed by the visitors of the video page. See the metadata information for this video. Description of various attributes displayed is given here.

**Query:** Rudy Giuliani

*Lisa & Crush feat. Giulani*

Collaboration with the very talented JackDanyells, who came up with the concept for this video. Check out his channel at: http://www.youtube.com/jackdanyells -Lyrics by JackDanyells -Vocal melody composed and sung by me -Royalty free background music from sounddogs.com

Comedy

Crawling since 2007-07-19

---

<table>
<thead>
<tr>
<th>Color coding for % changes</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
<th>5.0</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Crawl date</th>
<th>Rank</th>
<th>Views</th>
<th>Ratings</th>
<th>Avg Rating</th>
<th>Comments</th>
<th>Links</th>
<th>Favorite</th>
<th>Honors</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2007-07-31</td>
<td>5</td>
<td>27237</td>
<td>301</td>
<td>3.74</td>
<td>288</td>
<td>5</td>
<td>44</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>2</td>
<td>2007-08-01</td>
<td>5</td>
<td>27452</td>
<td>303</td>
<td>3.73</td>
<td>290</td>
<td>5</td>
<td>44</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>3</td>
<td>2007-08-02</td>
<td>5</td>
<td>27760</td>
<td>307</td>
<td>3.72</td>
<td>291</td>
<td>5</td>
<td>45</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>4</td>
<td>2007-08-03</td>
<td>5</td>
<td>28086</td>
<td>309</td>
<td>3.71</td>
<td>291</td>
<td>5</td>
<td>45</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>5</td>
<td>2007-08-04</td>
<td>5</td>
<td>28308</td>
<td>310</td>
<td>3.71</td>
<td>291</td>
<td>5</td>
<td>45</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>6</td>
<td>2007-08-05</td>
<td>5</td>
<td>28642</td>
<td>314</td>
<td>3.59</td>
<td>294</td>
<td>5</td>
<td>45</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>7</td>
<td>2007-08-06</td>
<td>5</td>
<td>29080</td>
<td>314</td>
<td>3.69</td>
<td>296</td>
<td>5</td>
<td>45</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>8</td>
<td>2007-08-07</td>
<td>5</td>
<td>29265</td>
<td>318</td>
<td>3.66</td>
<td>298</td>
<td>5</td>
<td>45</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>9</td>
<td>2007-08-08</td>
<td>5</td>
<td>29431</td>
<td>319</td>
<td>3.65</td>
<td>299</td>
<td>5</td>
<td>46</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>10</td>
<td>2007-08-09</td>
<td>5</td>
<td>30045</td>
<td>320</td>
<td>3.54</td>
<td>300</td>
<td>5</td>
<td>47</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>11</td>
<td>2007-08-10</td>
<td>5</td>
<td>30364</td>
<td>323</td>
<td>3.62</td>
<td>302</td>
<td>5</td>
<td>47</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>12</td>
<td>2007-08-11</td>
<td>5</td>
<td>30461</td>
<td>324</td>
<td>3.62</td>
<td>302</td>
<td>5</td>
<td>48</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>13</td>
<td>2007-08-12</td>
<td>5</td>
<td>30540</td>
<td>324</td>
<td>3.62</td>
<td>302</td>
<td>5</td>
<td>48</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>14</td>
<td>2007-08-13</td>
<td>5</td>
<td>30697</td>
<td>326</td>
<td>3.61</td>
<td>306</td>
<td>5</td>
<td>49</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>15</td>
<td>2007-08-14</td>
<td>5</td>
<td>30848</td>
<td>325</td>
<td>3.61</td>
<td>306</td>
<td>5</td>
<td>49</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>16</td>
<td>2007-08-15</td>
<td>5</td>
<td>31026</td>
<td>326</td>
<td>3.61</td>
<td>306</td>
<td>5</td>
<td>49</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>17</td>
<td>2007-08-16</td>
<td>5</td>
<td>31181</td>
<td>326</td>
<td>3.61</td>
<td>306</td>
<td>5</td>
<td>49</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>18</td>
<td>2007-08-17</td>
<td>5</td>
<td>31439</td>
<td>327</td>
<td>3.61</td>
<td>307</td>
<td>5</td>
<td>51</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>19</td>
<td>2007-08-18</td>
<td>5</td>
<td>31662</td>
<td>331</td>
<td>3.59</td>
<td>309</td>
<td>5</td>
<td>51</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>20</td>
<td>2007-08-19</td>
<td>5</td>
<td>31792</td>
<td>332</td>
<td>3.59</td>
<td>309</td>
<td>5</td>
<td>51</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>21</td>
<td>2007-08-20</td>
<td>5</td>
<td>31937</td>
<td>335</td>
<td>3.57</td>
<td>310</td>
<td>5</td>
<td>51</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>22</td>
<td>2007-08-21</td>
<td>5</td>
<td>32133</td>
<td>335</td>
<td>3.57</td>
<td>311</td>
<td>5</td>
<td>52</td>
<td>0</td>
<td>--</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 I/O equipment

Bitstream on physical medium
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
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
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
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
## Interaction Examples

**Examples**

<table>
<thead>
<tr>
<th>Interaction Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Browsing the contents of an archival collection using a finding aid</td>
<td>Viewing a web page that contains several files, including HTML, a style sheet and several images</td>
</tr>
<tr>
<td>Object or package</td>
<td>Using Microsoft Excel to view an .xls file, watching an online video using a Flash viewer</td>
</tr>
<tr>
<td>In-application rendering</td>
<td>Windows Explorer, typing &quot;ls&quot; at the Unix command prompt to show the contents of a directory</td>
</tr>
<tr>
<td>File through filesystem</td>
<td>Connecting a hard drive to a host computer and then generating a sector-by-sector image of the disk</td>
</tr>
<tr>
<td>File as &quot;raw&quot; bitstream</td>
<td>Connecting a floppy drive to a host computer and then generating a magnetic flux transition image of the disk</td>
</tr>
<tr>
<td>Sub-file data structure</td>
<td>Using a high-power microscope and camera to take a picture of the patterns of magnetic charges on the surface of a hard drive</td>
</tr>
<tr>
<td>Bitstream through I/O equipment</td>
<td>Using a hex editor to view an individual file</td>
</tr>
<tr>
<td>Extracting a tagged data element in an XML document or value of a field in a relational database</td>
<td>Connecting a floppy drive to a host computer and then generating a magnetic flux transition image of the disk</td>
</tr>
<tr>
<td>Bitstream on physical medium</td>
<td>Connecting a floppy drive to a host computer and then generating a magnetic flux transition image of the disk</td>
</tr>
</tbody>
</table>

Interaction Examples

Level
Aggregation of objects

Bitstream on physical medium

Veeco Instruments.
http://www.veeco.com/library/nanotheater_detail.php?type=application&id=78&app_id=34

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 LAM workflows.
- 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)
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
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

https://bitcuratorconsortium.org/
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.

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®
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

UltraBay II (legacy ports)

UltraBay 4D (contemporary ports)
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 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
Four Ways to Interact with Disk Images

Emulation

Mount them like regular drives:
- For ISO images - disk utilities in Mac OS or Windows 8/10
- For forensically packaged disk images: ewfmount, OSFMount, BitCurator (mounting scripts built into the environment)

Inspect them as forensic objects
- FTK Imager
- Autopsy
- BitCurator (Disk Image Access tool)

Dynamically navigate them from within a web browser (BitCurator Access Webtools)
Emulation as a Service

http://bw-fla.uni-freiburg.de/demos.html
Mounting a Forensically Packaged Disk Image in the BitCurator Environment
Exporting Files from a Disk Image
Explore raw and forensically-packaged (.E01 and .AFF) disk images in a web browser. Supported file systems include FAT, ExFAT, NTFS, HFS+, EXT2/3/4, ISO 9660 (CD-ROM), and YAFFS2 (Android). Groups of images currently registered with the system are listed below.

**Image Groups**

- **All Images**
  All images included recursively.
  - Images: 12

- **ISO test**
  Set of ISO test disk images.
  - Images: 2

- **Mixed test**
  Set of mixed-format test disk images.
  - Images: 10

[https://github.com/BitCurator/bitcurator-access-webtools](https://github.com/BitCurator/bitcurator-access-webtools)
### File Analysis for 2013-02-20_AAFS.pdf

#### File Details
- **Extension**: .pdf
- **Size**: 8476327
- **SHA1**: 0364598548ca19deb1d4f89990a4f21e8f44e5b9
- **MIME**: application/pdf

#### Full Text

**AAFS Digital & Multimedia Sciences Section**  
Thursday, February 21, 2013 / 3:45 p.m. - 4:05 p.m.

**Bulk Data Analysis With Optimistic Decompression and Sector Hashing**

Simson L. Garfinkel, Kristina Foster, Joel Young  
Naval Postgraduate School  
Kevin Fairbanks, Johns Hopkins Applied Physics Lab  
http://simson.net/

**Bulk Data Analysis With Optimistic Decompression and Sector Hashing**  

<table>
<thead>
<tr>
<th>DATE</th>
<th>TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thursday, February 21, 2013</td>
<td>3:45 p.m. - 4:05 p.m.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GPE</th>
<th>Simson L. Garfinkel</th>
<th>Kristina Foster</th>
<th>Joel Young</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORG</td>
<td>Naval Postgraduate School</td>
<td>Johns Hopkins Applied Physics Lab</td>
<td></td>
</tr>
</tbody>
</table>

http://simson.net/
Identifying Potentially Sensitive Data using Bulk Extractor - Scanning Options

See: [http://www.forensicswiki.org/wiki/Bulk_extractor](http://www.forensicswiki.org/wiki/Bulk_extractor)
Histogram of Email Addresses (Specific Instances in Context on Right)
Generating BitCurator Reports

The BitCurator Reporting Tool is used to generate reports. You can select options such as Run All, Fiwalk XML, Annotated Features, Reports, and File Access. The tool produces Office Open XML and PDF reports to assist in image analysis.

Examples of generated reports include:
- Fiwalk XML File: `/home/bcadmin/Desktop/SampleData/sampleimage.xml`
- Annotated Feature Files Directory: `/home/bcadmin/Desktop/SampleData/annotated-features`
- Output Directory For Reports (new): `/home/bcadmin/Desktop/SampleData/bc-reports`
- Config File (optional): `/Path/To/file`

Command Line Output:

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`
6) `/home/bcadmin/Desktop/SampleData/bc-reports/bc_reports/bc_format_bargraph.pdf`
7) `/home/bcadmin/Desktop/SampleData/bc-reports/format_table.pdf`
8) `/home/bcadmin/Desktop/SampleData/bc-reports/FiwalkReport.pdf`
9) `/home/bcadmin/Desktop/SampleData/bc-reports/FiwalkDeletedFiles.pdf`
10) `/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

```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>
  </creator>
  <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>
  <source>
    <image_filename>/home/bcadmin/Desktop/SampleData/sampleimage.E01</image_filename>
  </source>
  <volume_offset>0</volume_offset>
  <partition_offset>0</partition_offset>
  <block_size>2048</block_size>
  <ftype>2048</ftype>
  <ftype_str>iso9660</ftype_str>
  <block_count>36839</block_count>
</dfxml>
```
Operationalizing Original Order - Filesystem Metadata Output from fiwalk*

```xml
<fileobject>
  <parent_object>
    <inode>102</inode>
  </parent_object>
  <partition>1</partition>
  <id>901</id>
  <name_type>r</name_type>
  <filesize>100857</filesize>
  <alloc>1</alloc>
  <used>1</used>
  <inode>6783</inode>
  <meta_type>1</meta_type>
  <mode>511</mode>
  <nlink>1</nlink>
  <uid>0</uid>
  <gid>0</gid>
  <mtime prec="2">2009-11-17T19:35:10</mtime>
  <atime prec="86400">2009-12-10T05:00:00</atime>
  <crttime prec="2">2009-12-10T19:34:11</crttime>
  <libmagic>PDF document, version 1.4</libmagic>
  <byte_runs>
    <byte_run file_offset="0" fs_offset="56621568" img_offset="56653824" len="100857"/>
  </byte_runs>
  <hashdigest type="md5">eb60256dabfa67cef7211bcba659815</hashdigest>
  <hashdigest type="sha1">e56f606877f10daf91dc0304ea120b35452bd36e</hashdigest>
</fileobject>
```

*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.

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-080027f8dfe8</objectIdentifierType>
      <objectIdentifierValue>/home/bcadmin/Desktop/terry-work-usb-2009-12-11.E01</objectIdentifierValue>
    </objectIdentifier>
  </object>
  <event>
    <eventIdentifier>
      <eventIdentifierType>0d4ea1ce-b8dc-11e3-a80f-080027f8dfe8</eventIdentifierType>
      <eventIdentifierValue>E01</eventIdentifierValue>
    </eventIdentifier>
    <eventType>Capture</eventType>
    <eventDateTime>Wed Jan 19 12</eventDateTime>
    <eventOutcomeInformation>
      <eventOutcome>E01</eventOutcome>
      <eventOutcomeDetailVersion>20100226</eventOutcomeDetailVersion>
    </eventOutcomeInformation>
  </event>
  <event>
    <eventIdentifier>19882604-b8dc-11e3-93f0-080027f8dfe8</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 (for audio CDs)</td>
</tr>
<tr>
<td>(I addition to Guymager</td>
<td></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>
</tbody>
</table>
Other Functionality to Meet Identified User Needs (Continued):

<table>
<thead>
<tr>
<th>Function</th>
<th>Tool(s)</th>
</tr>
</thead>
<tbody>
<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>GTK Hash, md5deep, md5sum</td>
</tr>
<tr>
<td>Compare hashes of files to hashes in the National Software Reference Library (NSRL) of known system files</td>
<td>nsrllookup</td>
</tr>
<tr>
<td>View and edit bytestreams (hex editor)</td>
<td>Bless Hex Editor, GHex</td>
</tr>
</tbody>
</table>
Incorporating digital forensics into LAM workflows
Five Sources of Workflow Examples


http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf

Digital Sustainability Lab – Massachusetts Institute of Technology  

Workflows, BitCurator Consortium https://bitcuratorconsortium.org/workflows

OSSArcFlow Project - https://educopia.org/research/ossarcflow
Workflow

The following workflows depict the step-by-step processes BitCurator Consortium members follow to acquire, process, describe, and store the born-digital materials in their collections. Most of these resources are only accessible to members. Learn more about the benefits of membership.

If you are interested in adding a workflow to our listing, please contact us.

<table>
<thead>
<tr>
<th>Title</th>
<th>Contributor</th>
<th>Release Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processing Workflow</td>
<td>The University of Maryland, Libraries</td>
<td>2016 March 22</td>
</tr>
<tr>
<td>Princeton University Archives (Members Only)</td>
<td>Princeton University</td>
<td>2015 June 30</td>
</tr>
<tr>
<td>Penn State Born Digital (Members Only)</td>
<td>Penn State University</td>
<td>2014 July 29</td>
</tr>
<tr>
<td>Duke University Archives</td>
<td>Duke University</td>
<td>2012 August 12</td>
</tr>
<tr>
<td>Beineke Rare Books and Manuscripts Library</td>
<td>Yale University</td>
<td>2012 August 12</td>
</tr>
<tr>
<td>Maryland Institute for Technology in the Humanities</td>
<td>The University of Maryland, MITH</td>
<td>2012 August 12</td>
</tr>
<tr>
<td>University of North Carolina, Chapel Hill, Archives</td>
<td>University of North Carolina Chapel Hill, SILS</td>
<td>2012 August 12</td>
</tr>
<tr>
<td>University of Virginia Libraries</td>
<td>University of Virginia</td>
<td>2012 August 12</td>
</tr>
<tr>
<td>Yale University, Manuscripts and Archives</td>
<td>Yale University</td>
<td>2012 August 12</td>
</tr>
</tbody>
</table>

https://bitcuratorconsortium.org/workflows
Investigating, Synchronizing, and Modeling a Range of Archival Workflows for Born-Digital Content

Project Abstract
The Educopia Institute, in collaboration with the University of North Carolina at Chapel Hill School of Information and Library Science (UNC SILS), LYRASIS, and Artefactual, Inc. are investigating, synchronizing, and modeling a range of workflows to increase the capacity of libraries and archives to curate born digital content. These archival workflows will incorporate three leading open source software (OSS) platforms—BitCurator, Archivematica, and ArchivesSpace—and the project will be designed to generate findings that can be generalizable to settings that are using other platforms and applications.

This project will significantly impact curation practices by increasing our understanding of how institutions of different sizes and types may engage in OSS tool integration and workflow development. Our findings will be used to support a broad range of libraries and archives actively collecting and curating digital content. The knowledge gained by working with multiple institutions of different types and sizes will also broaden field-wide understanding of curation approaches and priorities, and how those impact the use of tools and capabilities in Archivematica, ArchivesSpace, and BitCurator. We expect the empirical findings about institutional needs, as well as formal workflow models, to contribute to digital curation research literature.

This project has been generously funded by the Institute of Museum and Library Services.

Project Outputs
Digital Dossiers

Contact:
Katherine Skinner
Additional Documents: OSSArcFlow proposal
As-Is Workflows (June 2018)

In the fall of 2017, the project team worked with partners at each institution to mock up a visual representation of their current workflow activities. Representing a "snapshot in time," these documents show how a diverse group of institutions are using OSS tools in their workflows to curate born-digital content. They also provide an essential starting point for synthesizing and comparing both the gaps and overlaps that currently exist between common OSS tools and environments.

1. Atlanta University Center, Robert W. Woodruff Library
2. District of Columbia Public Library
3. Duke University
4. Emory University
5. Kansas Historical Society
6. Massachusetts Institute of Technology
7. Mount Holyoke College
8. New York Public Library
9. Rice University
10. Stanford University
11. New York University
12. Odum Institute

[https://educopia.org/research/ossarcflow](https://educopia.org/research/ossarcflow)
For Further Information

https://bitcurator.github.io/
BitCurator
Quick Start Guide

Last updated: August 1, 2018
Release(s): 2.0.4 and later

https://github.com/BitCurator/bitcurator-distro/wiki/Releases
WEBINAR EVALUATION

We appreciate your feedback!

After you exit the webinar, you will automatically be taken to an online webinar evaluation.

Please take a couple of minutes to complete the survey and help us plan future webinars.
STAY CONNECTED & INFORMED

CoSA Website
http://www.statearchivists.org

CoSA Twitter Handle
@StateArchivists

CoSA Facebook Page
www.facebook.com/CouncilOfStateArchivists