Digital Forensics: Advanced

Instructors:
Christopher (Cal) Lee
University of North Carolina at Chapel Hill

Kam Woods
University of North Carolina at Chapel Hill

December 13-14, 2021
About These Slides

Authors
Cal Lee, Kam Woods

Description
These are the slides from Cal Lee and Kam Woods’s “Advanced Digital Forensics” class. There are a number of hands-on exercises included. The sample data referenced in these slides is available here: https://github.com/BitCurator/bcc-dfa-sample-data/

Learning object type
Lesson plan/materials

Learning objectives
This learning object might be used in a lesson to satisfy the following learning objectives:
• Identify the appropriate tools to: safely acquire born-digital materials from storage media and other modes of transfer; assist in the appraisal of born-digital materials; scan for sensitive information in born-digital materials; and package born-digital materials for preservation and access.
• Practice using tools in the BitCurator Environment.
Digital Archives Specialist (DAS)

Curriculum and Certification Program offered by SAA:

- **Foundational Courses**—must pass 4
- **Tactical and Strategic Courses**—must pass 3
- **Tools and Services Courses**—must pass 1
- **Transformational Courses**—must pass 1

- Course examinations are administered online.
DAS Core Competencies Addressed

• Understand the nature of records in electronic form, including the functions of various storage media, the nature of system dependence, and the effect on integrity of records over time.
• Formulate strategies and tactics for appraising, describing, managing, organizing, and preserving digital archives.
• Integrate technologies, tools, software, and media within existing functions for appraising, capturing, preserving, and providing access to digital collections.
• Curate, store, and retrieve original masters and access copies of digital archives.

https://www2.archivists.org/prof-education/das-curriculum-structure
Agenda

Day 1

- Welcome and introductions
- Motivation and overview
- Technical fundamentals
- Data acquisition considerations
- Potential elements of your own digital forensics lab
- Bit-level treatment of individual files
- Creating and extracting forensic metadata
- BitCurator reporting features
- Other BitCurator environment tools
- Preview of Day 2

Day 2

- Day 1 Postmortem – Questions, Concerns and Insights
- Command-line operations in Linux
- FIDO as an example of a command-line tool
- Regular expressions
- Extracting data from specific types of files: images, office files, email
- Windows artifacts (including the Registry)
- End user access (logistics and technical approaches)
- Incorporating digital forensics into archival workflows
- Challenges, Ethical/legal issues, and donor agreements
- Wrap up and evaluations
Personal Introductions

• Who’s teaching you?
• What about you?
  • Who are you (name, institution, job title)?
  • Why are you here (relevance to job, what you hope to get out of the workshop)?
  • What have you done so far to apply digital forensics methods in your institution?
Software You Should Have Installed for the Exercises

- VirtualBox
- VirtualBox Extensions
- BitCurator Virtual Machine
- Exiftool (or on the web via https://exif.tools/)
- Sample data: https://distro.ibiblio.org/bitcurator/samples/saa-dfa-sample-data.zip
- Visit the link above and download this now if you have not already done so!
- Additional tools for Windows users:
  - FTK Imager (Windows only)
  - OSFMount (Windows only)
  - RegRipper (Windows GUI, can also run at the command line in BitCurator environment)
Discussion Scenario

• You’ve been charged with taking care of data from a prominent community leader who has died unexpectedly
• Her materials include some paper and lots of digital data (on floppies, CDs, and a laptop hard drive)
• What should you do with the floppies?
• CDs?
• Hard drive?
Goals When Acquiring Born-Digital 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 to Apply

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

# Digital Forensics Can Help Archivists to Fulfill Their Principles

<table>
<thead>
<tr>
<th>Provenance</th>
<th>Identify, extract and save essential information about context of creation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Order</td>
<td>Reflect original folder structures, files associations, related applications and user accounts</td>
</tr>
<tr>
<td>Chain of Custody</td>
<td>Documentation of how records were acquired and any transformations to them</td>
</tr>
<tr>
<td></td>
<td>Use well-established hardware and software mechanisms to ensure that data haven’t been changed inadvertently</td>
</tr>
<tr>
<td>Identifying Sensitive Information</td>
<td>Identify personally identifying information, regardless of where it appears</td>
</tr>
<tr>
<td></td>
<td>Flag for removal, redaction, closure or restriction</td>
</tr>
</tbody>
</table>
Archivists need to apply many more processes to born-digital records (e.g. integrity checks, metadata extraction, audit trails, characterization)

The good news is that most of these processes can be performed by software
Digital Forensics in Archives

• In recent years, archivists 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
Need for Adaptation of Digital Forensics Tools and Tasks for Archivists

• Existing digital forensics tools provide valuable functionality, but they don’t always fit well into primary workflows of archives.

• For example, archives 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

After this class, you should be able to:

• Install and operate the BitCurator environment as a virtual machine in VirtualBox
• Explain and recognize different types of metadata stored in common filesystems
• Identify file types based on magic numbers (file signatures)
• Determine potential hardware options for acquiring data from various types of storage media
• Apply common Linux commands at the command line and compose basic regular expressions
• Evaluate disk image format options based on needs and priorities of your institution and collections
• Generate BitCurator reports and use bulk_extractor to identify potentially sensitive data
• Extract and interpret EXIF metadata
• Capture and analyze Windows Registry artifacts using RegRipper
• Determine essential points in your institution’s workflows where it will be beneficial to incorporate forensics tools and methods
• Make and justify decisions of professional ethics that emerge when caring for born-digital records
• Recognize technical strategies for providing access
Caveats and Such

• Advanced doesn’t mean "everything we didn’t cover in the Fundamental class"

• There’s much more about digital forensics that we won’t be addressing

• Selective hands-on experience with specific applications

• A license to learn more in the future
## Digital Resources - Levels of Representation*

<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>

<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>

Levels where digital forensics methods and tools can provide a lot of assistance

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)
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
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)
  - Using “bootstrapping” installation scripts to turn any Ubuntu Linux machine into a BitCurator Environment
  - Self-contained Linux environment in a virtual machine using e.g. VirtualBox 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: https://bitcurator.net/
Hands-On Familiarization with VirtualBox and the BitCurator VM*

*For a detailed walk-through of the steps we’re following, see the Quickstart Guide:

Get the software
Documentation and technical specifications
Screencasts
Google Group
People
Project overview
Publications
News

https://bitcurator.net/

Twitter: @bitcurator
Most tasks we will cover in this course are explained in the Quick Start Guide

BitCurator Consortium

• Continuing home for hosting, stewardship and support of BitCurator tools and associated user engagement
• Administrative home: Educopia Institute
• Funding based on membership dues
• Software and documentation are free and open source, but membership provides benefits (e.g. support, training, development priority)

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

How our members are using BitCurator
Technical Fundamentals
Checksums – Compact Representations of Bitstreams

- A given bitstream, fed into an algorithm, will generate a short string of characters that is extremely unlikely to be generated by a different bitstream fed into that same algorithm.
- Most common = MD5, SHA-1
- Can determine:
  - If bits have changed after a transfer
  - If bits have flipped within a storage environment
  - Whether two different files are identical bitstreams
- A library of hash values can identify “known and notable” (EnCase terminology) files
  - Known – files that can be ignored (e.g. software listed in National Software Reference Library)
  - Notable – specific bitstreams that you’re trying to find
In BitCurator environment: Right Click on File or Directory and Calculate MD5
The MD5 hash of the selected file:

```
keb2622125be1231b0fc9babe27942d /home/bcadmin/Pictures/bitcurator-grub.png
```
Note on MD5/SHA1 - Potential Collisions

- From a security perspective, MD5 has been “broken” since 2005
- SHA-1 was broken in February 2017
- Someone with malicious intent can create two different bitstreams that result in the same hash (i.e. hash collisions)
Hash Collisions (The Poisoned Message Attack): "The Story of Alice and her Boss"

Being an intern, Alice does not have any access to secret documents. Not enough for her ...

... tricky Alice decides to fool Caesar. Because Caesar is still relying on the widely used MD5 hash function, she implements the attack from Wang and Yu [WY05] to find MD5 collisions. When she receives her letter of recommendation (on paper), she prepares **two postscript files with the same MD5 hash**:

- One to display the letter of recommendation, and
- a second one, an order from Caesar to grant Alice some kind of a security clearance.

*Stefan Lucks and Magnus Daum,  
http://th.informatik.uni-mannheim.de/people/lucks/HashCollisions/  
Wayback link:  
May, 22, 2005

Order:

Alice Falbala is given full access to all confidential and secret information about GAUL.

Sincerely,

Julius Caesar

May, 22, 2005

To Whom it May Concern:

Alice Falbala fulfilled all the requirements of the Roman Empire intern position. She was excellent at translating roman into her gaul native language, learned very rapidly, and worked with considerable independence and confidence.

Her basic work habits such as punctuality, interpersonal deportment, communication skills, and completing assigned and self-determined goals were all excellent.

I recommend Alice for challenging positions in which creativity, reliability, and language skills are required.

I highly recommend hiring her. If you’d like to discuss her attributes in more detail, please don’t hesitate to contact me.

Sincerely,

Julius Caesar
Alternatives to MD5?
SHA (Secure Hash Algorithm)

- Originally developed by the NSA
- Several variants: SHA-0, SHA-1, SHA-2 family
- Early variants (SHA-0, SHA-1) known to be compromised
- Most commonly used is now SHA-256. Can be used to process bitstreams (“messages”) up to $3.4 \times 10^{38}$ bits (very large!)
- Disadvantage: more time, computing power required to produce

---

<table>
<thead>
<tr>
<th>Algorithm and variant</th>
<th>Output size (bits)</th>
<th>Internal state size (bits)</th>
<th>Block size (bits)</th>
<th>Max message size (bits)</th>
<th>Word size (bits)</th>
<th>Rounds</th>
<th>Operations</th>
<th>Collisions found?</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHA-0</td>
<td>160</td>
<td>160</td>
<td>512</td>
<td>$2^{54} - 1$</td>
<td>32</td>
<td>80</td>
<td>add, and, or, xor, rotate, mod</td>
<td>Yes</td>
</tr>
<tr>
<td>SHA-1</td>
<td>256/224</td>
<td>256</td>
<td>512</td>
<td>$2^{64} - 1$</td>
<td>32</td>
<td>64</td>
<td>add, and, or, xor, rotate, mod, shift</td>
<td>Theoretical attack ($2^{60}$)</td>
</tr>
<tr>
<td>SHA-2</td>
<td>512/384</td>
<td>512</td>
<td>1024</td>
<td>$2^{128} - 1$</td>
<td>64</td>
<td>80</td>
<td>add, and, or, xor, rotate, mod, shift</td>
<td>No</td>
</tr>
</tbody>
</table>

Implications of MD5/SHA1 Being "Broken"

- Rarely a concern when hash is used for integrity checks on known items (e.g. verifying that a file was transferred correctly to a repository or that files in storage are still intact)

- Can be a concern if one is relying on a hash as proof of record authenticity – risks can include cases of internal tampering

- There are more robust hash algorithms to address this (SHA-2 family, including SHA-256) so good practice is to generate one of them along with the MD5

- MD5 is still widely used, because it is fast to calculate and still widely supported
Question:

Can you use a cryptographic hash to determine specifically what any given file contains?

If not, what could you use?
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)
<table>
<thead>
<tr>
<th>Name</th>
<th>Operating System(s) Using it as Native File System [often other OSs can also recognize it]</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAT12, FAT16</td>
<td>MS-DOS</td>
</tr>
<tr>
<td>FAT32 (VFAT)</td>
<td>Windows 95, 98</td>
</tr>
<tr>
<td>exFAT</td>
<td>Windows XP SP2 and later (primary use: USB drives, SD cards)</td>
</tr>
<tr>
<td>MFS</td>
<td>Macintosh System 1-3</td>
</tr>
<tr>
<td>HFS (Hierarchical File System)</td>
<td>Macintosh System 4-8</td>
</tr>
<tr>
<td>HFS+</td>
<td>Macintosh System 8.1 – 9, OS X 10.0 – 10.11</td>
</tr>
<tr>
<td>APFS</td>
<td>macOS 10.12</td>
</tr>
<tr>
<td>ext, ext2, ext3, ext4 (Extended File System)</td>
<td>Linux</td>
</tr>
<tr>
<td>XFS</td>
<td>Linux, typically Enterprise variants (RHEL)</td>
</tr>
<tr>
<td>HPFS (High Performance File System)</td>
<td>OS/2</td>
</tr>
<tr>
<td>ISOFS (ISO 9660)</td>
<td>Any OS that reads data from a CD</td>
</tr>
<tr>
<td>JFS1 (Journaled File System)</td>
<td>AIX (IBM)</td>
</tr>
<tr>
<td>ReiserFS</td>
<td>Several Linux distributions</td>
</tr>
<tr>
<td>UFS (Unix File System) aka FFS (Fast File System)</td>
<td>Various flavors of Unix</td>
</tr>
<tr>
<td>Name</td>
<td>Operating System(s) Using it as Native File System [often other OSs can also recognize it]</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>FAT12, FAT16</td>
<td>MS-DOS</td>
</tr>
<tr>
<td>FAT32 (VFAT)</td>
<td>Windows 95, 98</td>
</tr>
<tr>
<td>exFAT</td>
<td></td>
</tr>
<tr>
<td>NTFS</td>
<td></td>
</tr>
<tr>
<td>MFS</td>
<td></td>
</tr>
<tr>
<td>HFS (Hierarchical File System)</td>
<td></td>
</tr>
<tr>
<td>HFS+</td>
<td></td>
</tr>
<tr>
<td>APFS</td>
<td></td>
</tr>
<tr>
<td>ext, ext2, ext3, ext4 (Extended File System)</td>
<td></td>
</tr>
<tr>
<td>XFS</td>
<td>Linux</td>
</tr>
<tr>
<td>HPFS (High Performance File System)</td>
<td>OS/2</td>
</tr>
<tr>
<td>ISOFS (ISO 9660)</td>
<td>Any OS that reads data from a CD</td>
</tr>
<tr>
<td>JFS1 (Journaled File System)</td>
<td>AIX (IBM)</td>
</tr>
<tr>
<td>ReiserFS</td>
<td>Several Linux distributions</td>
</tr>
<tr>
<td>UFS (Unix File System) aka FFS (Fast File System)</td>
<td>Various flavors of Unix</td>
</tr>
</tbody>
</table>

Filesystems you’re most likely to encounter
NTFS vs. FAT File System Attributes

- Two disk images are in your zip file, and can also be found at: https://digitalcorpora.s3.amazonaws.com/corpora/scenarios/2009-m57-patents/usb/terry-work-usb-2009-12-11.E01

- Load each disk image into a separate instance of FTK Imager (run them side by side to compare what you see) – if you don’t have a Windows computer, look on with a partner

- Look at the properties of some files*

- What differences do you notice?

*Properties are shown in the bottom left corner. If you don’t see them, go to the View menu at the top and select “Properties.” You may need to drag the top of the properties window up to see all of the values.
Connecting a Device vs. Mounting a Filesystem
HFS+ volume is visible through Windows Disk Management after it’s connected.
But it's not visible through Windows Explorer, because Windows doesn't know how to mount the file system
Seeing Attached Devices (Whether or Not They're Mounted) - In Windows

- Control Panel
- Administrative Tools
- Computer Management
- Disk Management
Data Acquisition Considerations
Ports and Connectors

- Adoption comes and goes with changes in the industry
- Those that have had wide industry adoption have usually lasted a decade or more
- Important distinction: hardware protocol standards vs. shape of connectors (related but not always the same)

Different USB connectors. From left to right: male Micro USB B-Type, UC-E6 proprietary (not USB), male Mini USB (5-pin) B-type, female A-type, male A-type, male B-type. Shown with a centimeter ruler. Female A-type connector (4th from left) is "upside down" to show the pins.

Adapter Examples

- Micro SD to SD
- MicroSATA to SATA
- NVMe to USB
- SATA to IDE
- Ethernet to USB
- SATA to Molex Power
A Note Regarding Modern Laptops (and other devices)

- Some recent laptops, including all Apple laptops post-2019, the Microsoft Surface series of machines, and certain others, have both RAM and storage soldered to the motherboard; these drives cannot* be removed for imaging.

*Except in specialized laboratories.

M.2 (also known as "Next Generation Form Factor")

- Current most common way to connect a mass storage device (e.g. solid state storage) to the inside of many desktop and laptop computers. Cable-less, connects via an edge slot.

- Internal M.2 devices mount to the motherboard or a PCI expansion card. External M.2 devices typically housed in USB enclosures
Serial Advanced Technology Attachment (SATA)

- The most common way to connect a mass storage device (e.g. hard drive, solid state drive) to the inside of a computer from 2003 through the mid-2010’s.

- eSATA used for external mass storage devices

Data Cables:
SATA on the left, eSATA on the right

SATA Power Cable
Integrated Drive Electronics (IDE) / AT Attach

- Precursor to SATA
- No longer widely used, but you’ll still find it on older drives
- More than one drive can share the same cable, with one being the *master* (device 0) and the other being the *slave* (device 1) – these are set with *jumpers*

40-pin IDE Ribbon Cable

https://static.daniweb.com/images/attachments/0/maxtor_jumper.jpg
Floppy Disks

- Physical storage is similar to hard drives (magnetic charges in a spinning disk)
- Various types and sizes, e.g. high density, double density, 3.5 inch, 5.25 inch, 8 inch
- 3.5 inch floppies are relatively easy to read using a USB drive, but older ones are more complicated...
Floppy Controller Hardware

Kryoflux¹

FluxEngine²

Disc Ferret³

FC 5025⁴

Disk2FDI⁵

SuperCard Pro⁶

1. https://www.kryoflux.com/
Common Floppy Formats (Physical)

Many variations over time, often to increase storage density

Table source: https://en.wikipedia.org/wiki/List_of_floppy_disk_formats

<table>
<thead>
<tr>
<th>Size</th>
<th>Density</th>
<th>Tracks</th>
<th>tpi</th>
<th>bpi</th>
<th>Coercivity</th>
<th>Unformatted capacity per side</th>
</tr>
</thead>
<tbody>
<tr>
<td>2½-inch</td>
<td>Single</td>
<td>16[16][17]</td>
<td>48[16]</td>
<td></td>
<td></td>
<td>64 KB[16][17]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80</td>
<td>135</td>
<td>8717</td>
<td>600-665 Oe</td>
<td>500 KB</td>
</tr>
<tr>
<td>3½-inch</td>
<td>High</td>
<td>80</td>
<td>135</td>
<td>17434</td>
<td>720-750 Oe</td>
<td>1000 KB</td>
</tr>
<tr>
<td></td>
<td>Extended</td>
<td>80</td>
<td>135</td>
<td>34868</td>
<td>900 Oe</td>
<td>2000 KB</td>
</tr>
<tr>
<td>5¼-inch</td>
<td>Single/Double</td>
<td>40</td>
<td>48</td>
<td>5876</td>
<td>300 Oe</td>
<td>250 KB</td>
</tr>
<tr>
<td></td>
<td>Double</td>
<td>80</td>
<td>62.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quad</td>
<td>77</td>
<td>100</td>
<td>300 Oe</td>
<td></td>
<td>500 KB (Micropolis-compatible)</td>
</tr>
<tr>
<td></td>
<td>Quad</td>
<td>80</td>
<td>96</td>
<td>5922</td>
<td>300 Oe</td>
<td>500 KB</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>80</td>
<td>96</td>
<td>9646</td>
<td>600 Oe</td>
<td>833 KB</td>
</tr>
<tr>
<td>8-inch</td>
<td>Single/Double</td>
<td>77</td>
<td>48</td>
<td>300 Oe</td>
<td></td>
<td>1000 KB</td>
</tr>
</tbody>
</table>

*Coercivity is how resistant the medium is to being demagnetized (larger number means it requires a larger magnetic field to be demagnetized)
Actual bits on most IBM PC-type floppies are encoded using "modified frequency modulation" (MFM)

<table>
<thead>
<tr>
<th>IBM PC compatibles[19]</th>
<th>8-inch</th>
<th>5½-inch</th>
<th>3½-inch (90 mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>128</td>
<td>1024</td>
<td>512</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>77</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>360</td>
<td>300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MFM</td>
<td>MFM</td>
</tr>
</tbody>
</table>

Actual bits on Apple (Macintosh and earlier) may be encoded using "group code recording" (GCR) or "modified frequency modulation" (MFM)

<table>
<thead>
<tr>
<th>Apple</th>
<th>Format</th>
<th>Type</th>
<th>Capacity</th>
<th>Density</th>
<th>Density</th>
<th>Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple II</td>
<td>5½-inch</td>
<td>Double</td>
<td>256</td>
<td>13</td>
<td>35</td>
<td>113.75 KB</td>
</tr>
<tr>
<td></td>
<td>3½-inch (90 mm)</td>
<td>Double</td>
<td>512</td>
<td>Variable (8-12)</td>
<td>80</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>512</td>
<td>18</td>
<td>80</td>
<td>1440 KB</td>
</tr>
<tr>
<td>Apple Lisa</td>
<td>5½-inch FileWare</td>
<td>Double</td>
<td>512</td>
<td>Variable (15-22)</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>Apple Lisa 2/Macintosh XL</td>
<td>3½-inch (90 mm)</td>
<td>Double</td>
<td>512</td>
<td>Variable (8-12)</td>
<td>80</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>512</td>
<td>18</td>
<td>80</td>
<td>1440 KB</td>
</tr>
</tbody>
</table>

Issues with Common Floppy Formats

Why is this important?

- Floppy disk drives write the actual bitstream using either GCR or MFM (or other, less common encodings)
- Drive won’t read disks encoded in a non-supported format without specialized hardware (e.g. KryoFlux)
- Other factors:
  - Some devices (Apple in particular) use “CLV” (constant linear velocity), adjusting rotation speed depending on where the head is writing/reading
  - Others vary angular velocity of the disk based on zones (“zoned CAV”).
  - Again, need specialized hardware such as KryoFlux to read these disks on a modern drive
Additional Factors: File Systems Used on Floppies

- **Common IBM PC compatible floppies:**
  - Typically FAT12 with 512 byte clusters (although there are many less common variations)

- **Common Apple floppies (Macintosh and previous):**
  - Apple II: 13 sector disk (5.25”, ProDOS 3.2, 113.75K)
  - Apple II: 16 sector disk (5.25”, ProDOS 3.3, 140K)
  - Apple Macintosh:
    - Double-density 3.5” - Macintosh File System (MFS)
    - Double-density 3.5” - Hierarchical File System (HFS)
    - High-density 3.5” - Hierarchical File System (HFS)
## Potential Problems with ISO 9660 Media (e.g. CD-ROMs)

<table>
<thead>
<tr>
<th>Factor</th>
<th>Complications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical damage</td>
<td>May not be visible to the naked eye</td>
</tr>
<tr>
<td>Bad length in volume header</td>
<td>Older CD writing tools sometimes miscalculated sector count, so header metadata doesn’t match actual length</td>
</tr>
<tr>
<td>Incorrect block size</td>
<td>CDs can misidentify as having 512-byte sectors</td>
</tr>
<tr>
<td>File truncation</td>
<td>Filesystem may allow you to navigate to files that subsequently appear damaged or won’t render at all (files could be truncated or never fully written to disk)</td>
</tr>
<tr>
<td>One sector short when written Track at Once (TAO)</td>
<td>TAO disks often represent a length (in volume header) that is one sector short from the actual length</td>
</tr>
</tbody>
</table>

Not all tools are designed to recognize or address these issues
Note that there are structures, such as the Path Table, that are not generally used, but may contain metadata. For more info on issues in managing optical media, see: https://kamwoods.net/publications/woodsbrownarch09.pdf
Creating Disk Images of CDs

- Cdrdao – primarily for ripping audio CDs (addressing issues such as the TAO one discussed above)

- Often sufficient to use `dd` form CD-ROMs
  - In the BitCurator VM, CD-ROM drive should appear as a device called `/dev/sr0`
  - Command to acquire:
    ```bash
    bcadmin@ubuntu:~$ dd if=/dev/sr0 of=FILENAME1.iso
    ```
Dealing with Disk Images from CDs

To modify the file system, you can:

- Mount the disk image
- Use mkisofs to create a new ISO file system -
  `bcadmin@ubuntu:~$ mkisofs -r -o FILENAME2.iso /media/sr0`

If you created an EWF image of a CD:

- BitCurator mounting scripts can’t determine whether the underlying disk image data is a raw (dd) image or an ISO
- To mount the image, first use ewfexport (command-line tool) to pull out the raw (dd) image, then rename it as an ISO
Two Important Considerations for Internal Media that are Used as External Media

- **Power** - internal drive needs different connector (often Molex), not the kind that plugs into the wall.

- **Cooling** – when pulled from the computer, you’ve also separated the drive from the fan, so you should often add an external one to ensure cooling.
Kryoflux installed and running in a mini jukebox

*Adapted from a Mini JukeBox setup designed by the National Library of Australia*
Write Blocking – One-Way Street 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
Hard Disk Write Blocker

A WiebeTech SATA/IDE write-blocker
This end connects to your computer using USB 2.0/3.0, eSATA, or FireWire. The cables are in the box.
Hard Disk Write Blocker

This end connects to the drive you want to image using SATA or IDE. There are power cables in the box that can connect to either drive.
Hard Disk Write Blocker

This end powers the write blocker itself. There’s a power supply in the box.
Hard Disk Write Blocker

The WiebeTech Combodock V5 supports both USB 2.0 and USB 3.0. USB 3 is faster, but may not always be compatible with all computers.
The ComboDock has been connected to an IDE drive, a host computer, and powered on. It will allow you to select write-blocked or read/write. “Enter” enables write-blocking.
Hard Disk Write Blocker

The drive is now fully powered, and you can use the dock to examine some metadata...
This drive has just been powered on, and it’s registering as 14 degrees Celsius (57 Fahrenheit). This is well below room temperature, but it will quickly rise without a fan.
Hard Disk Write Blocker

This drive is indicating a raw capacity of 156 thousand megabytes (about 150GB).
The name of the manufacturer is embedded in the metadata. Why might this be important?
Hard Disk Write Blocker

The model number is next. This number is *not* unique, but common to many drives.
The serial number, however, *is* unique.
A USB Write Blocker
USB Write Blocker

Above is a USB write blocker manufactured by Tableau.
USB Write Blocker

This end has a power connector for the blocker, and a USB ‘b’ port to connect to the host. The FireWire port is just for updating the write blocker firmware. It has no other use.
USB Write Blocker

This end connects to the USB device. This particular write-blocker will *not* recognize USB floppy drives, only USB flash drives and USB hard disks. And look, a power switch.
Here’s what it looks like when everything is plugged in. Plug everything in before turning the write-blocker on.
USB Write Blocker

Once you turn it on, the “Host Detect” and other lights will light up. If “Host Detect” doesn’t light up, the write blocker can’t see your computer. Something has gone wrong.
USB Write Blocker

Different write blockers may expose different metadata. This write blocker displays the product name, in addition to the manufacturer.
Host-Based Write Blockers
FRED UltraBay

- The “UltraBay” on the FRED provides write blocking for a range of interfaces.
  - USB (top left, next to power switch)
  - SCSI (right of USB) [no longer included in newer FREDs]
  - PATA/IDE (below SCSI)
  - SATA (left of IDE)
- It also includes a MOLEX power connector. There’s a cable in the toolbox that converts this to SATA power, if required.
- The ports in the white box on the right (top picture) are NOT write-blocked.
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
Other Potential System Changes to Reduce Risk of Writing to File Systems

- **Macintosh:** Disk Arbitrator “will block the mounting of file systems to avoid mounting as read-write and violating the integrity of the evidence”
  https://github.com/aburgh/Disk-Arbitrator

- **Windows:** See “Digital Forensics: How to configure Windows Investigative Workstations”
  https://www.sans.org/blog/digital-forensics-how-to-configure-windows-investigative-workstations/
Potential Elements of your own Digital Forensics Lab
FRED Options from Digital Intelligence

FRED
The Forensic Standard
Learn more »  Buy »

FRED DX
Dual CPU Performance
Learn more »  Buy »

FRED SR
High Capacity
Learn more »  Buy »

FRED L
Forensic Laptop
Learn more »  Buy »

https://digitalintelligence.com/products/overview
Mini Jukebox

Configuration #2 for Manuscripts

3 x Double DVD Drive units (MyBorg-006> MyBorg-007)

- 1 x Storage Tower (black) with 4SATA Multilane connector installed ST4SAML-B, 4 bay aluminium unit

- 2 x Plextor PX-800A DVD-RW Super Multi Drive Drives
  - Drive Letter Mapping W & X

- 1 x 3.5” floppy disk Drive Letter Mapping A or B

- 1 x Western Digital WD5001ABYS CAVIAR RE2/ 500GB Hard Drive Letter Mapping Z

Stanford University Libraries and Academic Information Resources (SULAIR)
The British Library (UK)
School of Information and Library Science at UNC Chapel Hill, North Carolina
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.

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.

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.

Weibetech 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. Weibetech charges $250 for the UltraDock Hardware Write Protector.

Outfitting a Born-Digital Archives Program
(Ben Goldman, Penn State University)

https://practicaltechnologyforarchives.org/issue2_goldman/
Useful Resource - Mediapedia

Item Details

Product Name: CompacTape™ (a.k.a. DLT™ Tape) TK-50 ½" Data Cartridge
Name of Holotype: DLT tape (formerly CompacTape)
Product Code/Number: SKU 375705
Manufacturer: 3M - Scotch - Imation
Genre(s): Data
Carrier Type: tape cartridge
Process Type: magnetic

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.
Creating a Disk Image in Guymager
Examples of Disk Image Formats

- RAW and Split RAW (RAW stored across multiple files)
- Advanced Forensics Format (AFF) [no longer recommended]
- EnCase Evidence File (.E01)
- ISO (for CD-ROM)
- IMG (floppy or sometimes CD-ROM)
RAW (dd)

- Copies of the raw media data. Often split into smaller chunks to make them more manageable and so that the resulting images can fit onto limited filesystems and media such as FAT or DVD/CDROM.

  - **Advantages:**
    - Very simple, use simple tools to manipulate the image.
    - Image can be easily split for storage and transport on removable media
    - Output can be piped to other applications for immediate processing

  - **Disadvantages:**
    - Can be very large (no compression). Zipped raw images cannot be operated on directly with regular tools (efficiently perform arbitrary seeks).
    - Often too large to store on FAT formatted media
    - No metadata other than filenames, no hashes.
    - No checksumming on files – not robust
      - Missing segments (for example from scratched CD/DVD – can sometimes be overwritten with 0’s).
      - Overwritten data (unrecoverable – no checksums on small blocks in file).
Expert Witness Format (Encase)

- Evidence file consists (in order) of: Acquisition information, Data Block, CRC (cyclic redundancy check), acquisition hash (MD5)
- Can be split for storage, transport
- CRC computed for every 32K block; balance between integrity and speed, also makes it very difficult to tamper with the evidence file (1 in 4 billion chance of collision)
- Cannot be manipulated with simple (open source UNIX) tools; support reverse engineered in libewf
- Previously limited to 2GB size
- Largely proprietary
- Has been reverse engineered by Joachim Metz in libewf (used in open source tools that read EWF) - https://github.com/libyal/libewf
ISO Images (.iso extension) for CD-ROM and DVD

- Similar to raw, but can’t contain
  - multiple tracks
  - audio or video tracks
- Don’t contain control headers or error correction fields (raw can include these)
- Filesystem usually will be either ISO 9660 (CD-ROM) or UDF (DVDs)
Accessing Data in Disk Images

- Virtualization and emulation
- Mounting the original filesystem
- Accessing (but not mounting) disk images using forensics software

Two options discussed later for end user access:
- Remote, dynamic access to disk image contents
- Cross-drive analysis
Emulation as a Service

https://eaasi.gitlab.io/program_docs/intro-emulation-workshop/06-eaasi/index.html
What's the difference between the two options shown in FTK Imager below?
Mounting a Disk Image to Browse the Contents
Mounting a Disk Image to Browse the Contents
Exporting Selected Files from a Disk Image
Exercise: Multiple Views into Disk Image Files

Resources we’ll be using (also in the zip file you downloaded):

2. IMG file – [https://distro.ibiblio.org/bitcurator/lab/something.img](https://distro.ibiblio.org/bitcurator/lab/something.img)
3. OSFMount (Windows only)
4. FTK Imager (Windows only)
5. BitCurator Environment

- Step 1 – Mount the ISO and IMG files using **OSFMount**
- Step 2 – Find the drives using **Windows Explorer** and investigate their contents
- Step 3 – Open **FTK Imager** and add both images as evidence items, and explore what we see in the drives
- Step 4 – Use the **BitCurator environment** to mount the disk images [Right click on image file, then select: Scripts > Mount Disk Image]
- Step 5 – Use the **BitCurator environment** to select files within the images to export [Use Forensics Tools > BitCurator Disk Image Access]
Bit-Level Treatment of Individual Files
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
Hex Dump Tools

- Many free or inexpensive tools available for download, e.g. Cygnus Hex Editor, Hex Workshop, HexAssistant, MiniDumper, Hex Fiend (Mac), GHex (Linux)*

- BitCurator environment has a built-in hex editor (GHex)

Online tool: https://hexed.it/

* See https://en.wikipedia.org/wiki/Comparison_of_hex_editors
In the BitCurator environment:
Identifying File Types

- Magic numbers and file signatures
- File extensions
- Metadata stored in file system
- MIME types
Magic Numbers and File Signatures

- Distinct string or pattern that is found within files of a given type (most often in the header)
- Most effective searches for magic numbers often involve regular expressions (e.g. grep) in order to indicate multiple variations of a pattern
- Utilities that use this: file (Unix), TrID, DROID, FITS
- Examples:

<table>
<thead>
<tr>
<th>File Format</th>
<th>Hex</th>
<th>ASCII</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOC</td>
<td>D0 CF 11 E0 A1 B1 1A E1</td>
<td>ÐÏà¡±á</td>
</tr>
<tr>
<td>JPG</td>
<td>FF D8 FF</td>
<td>ÿØÿ</td>
</tr>
<tr>
<td>PDF</td>
<td>25 50 44 46 2D 31 2E</td>
<td>%PDF-1.</td>
</tr>
<tr>
<td>ZIP</td>
<td>50 4B 03 04</td>
<td>PK..</td>
</tr>
</tbody>
</table>
FITS “identifies, validates, and extracts technical metadata for various file formats. It wraps several third-party open source tools, normalizes and consolidates their output, and reports any errors. FITS was created by the Harvard University Library Office for Information Systems for use in its Digital Repository Service (DRS).”

Tools currently bundled into it:

- Jhove
- Exiftool
- National Library of New Zealand Metadata Extractor
- DROID
- FFIdent
- File Utility (windows)

FITS may no longer be available as a part of BitCurator due to JRE incompatibilities
Siegfried

http://www.itforarchivists.com/siegfried/

- Signature-based file format identification tool
  - PRONOM file format signatures (National Archives of UK) (default)
  - MIME-info file format signatures (freedesktop.org)
  - FDD file format signatures (Library of Congress)
- Unlike FITS, does not have validation built in, and fewer extraction tools, but much lighter weight.
- Has a lot of customization for output
  - CSV
  - YAML (text)
  - DROID CSV
  - JSON
  - stdout
Brunnhilde

https://github.com/tw4l/brunnhilde

- Reporting companion for Siegfried
  - Requires Siegfried (but running Brunnhilde also runs Siegfried)
  - Command-line and GUI (we’ll be using the CLI version later)
- Reports generated
  - HTML (human readable)
  - Siegfried CSV
  - Directory tree
  - Other CSVs extracted from Siegfried logs (e.g., warnings, unidentified files)
- Can run other processes too, but not required
  - Virus scan
  - bulk_extractor
  - Disk image processing
File Extensions

- Changing file extension usually changes default application OS uses to open (i.e. associates with) the file
- The "8.3" (eight characters, followed by three-character extension) limit in the past – based on FAT – resulted in many creative uses of the extension part of the file name (e.g. reports1.994, april-94.rpt)
- Convention is often still to use only three letters
- No authority for standardizing use, so three-letter extensions are often shared by many formats
- Security risks associated with trusting the file extension to be accurate – malicious code masquerading as another type of file (e.g. viruses sent as email attachments)
MIME types ("Content-type", "internet media type")

- Widely adopted and recognized by applications
- Based on two-level hierarchy (e.g. text/html, application/octet-stream, image/tiff)
- Major advantage is official registration of MIME types through a central authority
## MIME types

<table>
<thead>
<tr>
<th>Name</th>
<th>MIME Type / Internet Media Type</th>
<th>File Extension</th>
<th>More Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D Crossword Plugin</td>
<td>application/vnd.hzn-3d-crossword</td>
<td>.x3d</td>
<td>IANA: 3D Crossword Plugin</td>
</tr>
<tr>
<td>3GP</td>
<td>video/3gpp</td>
<td>.3gp</td>
<td>Wikipedia: 3GP</td>
</tr>
<tr>
<td>3GP2</td>
<td>video/3gpp2</td>
<td>.3g2</td>
<td>Wikipedia: 3G2</td>
</tr>
<tr>
<td>3GPP MSEQ File</td>
<td>application/vnd.mseq</td>
<td>.mseq</td>
<td>IANA: 3GPP MSEQ File</td>
</tr>
<tr>
<td>3M Post It Notes</td>
<td>application/vnd.3m.post-it-notes</td>
<td>.pwn</td>
<td>IANA: 3M Post It Notes</td>
</tr>
<tr>
<td>3rd Generation Partnership Project - Pic Large</td>
<td>application/vnd.3gpp.pic-bw-large</td>
<td>.plb</td>
<td>3GPP</td>
</tr>
<tr>
<td>3rd Generation Partnership Project - Pic Small</td>
<td>application/vnd.3gpp.pic-bw-small</td>
<td>.psb</td>
<td>3GPP</td>
</tr>
<tr>
<td>3rd Generation Partnership Project - Pic Var</td>
<td>application/vnd.3gpp.pic-bw-var</td>
<td>.pvb</td>
<td>3GPP</td>
</tr>
<tr>
<td>3rd Generation Partnership Project - Transaction Capabilities Application Part</td>
<td>application/vnd.3gpp2.tcap</td>
<td>.tcap</td>
<td>3GPP</td>
</tr>
<tr>
<td>7-Zip</td>
<td>application/x-7z-compressed</td>
<td>.7z</td>
<td>Wikipedia: 7-Zip</td>
</tr>
</tbody>
</table>

[https://www.iana.org/assignments/media-types/media-types.xhtml](https://www.iana.org/assignments/media-types/media-types.xhtml)
The NSRL (https://www.nsrl.nist.gov) includes a library of hashes of files associated with a large number of software tools developed over the past few decades. See the product list at:

https://www.nsrl.nist.gov/RDS/rds_2.41/ProdList.txt

Various third-party tools can be used to access the NSRL. There’s a web interface available at:

https://www.hashsets.com/home/

(Navigate to Hash Set Engines > National Software Reference Library > SEARCH BY NAME / MD5). But it often generates invalid results, so the following instructions are based on running a command-line tool instead.
Using NSRL Hash Sets to Investigate System Files

- Find a directory from your computer that contains system files.
  - For Windows, a good place to look is in Computer > Local Disk (C:) > Program Files. For example, you could select Program Files > 7-Zip.
  - On a Mac, look in /Applications/ and select a specific folder
- Move the contents of the directory to a new folder called system-files on your host computer’s desktop.
- Navigate to your shared folders [Desktop > Shared Folders and Media] in the BitCurator environment and copy the folder system-files to the desktop of the BitCurator environment
- Use md5deep to create a set of md5 hashes of the files in the system-files folder, then pipe the output into nsrllookup to generate lists of known and unknown hashes:
  - `md5deep -r ~/Desktop/system-files | nsrllookup -s nsrllookup.com -K known-hashes.txt -U unknown-hashes.txt`
- What is the above command doing?
- Look at the contents of the two files:
  - `type known-hashes.txt`
  - `type unknown-hashes.txt`
For Your Reference: Running NSRL Lookup in Windows

- Visit: [https://rjhansen.github.io/nsrllookup/](https://rjhansen.github.io/nsrllookup/) and download the Windows binary (64-bit).
- Open the .zip file and extract the executable to your desktop.
- Visit: [https://github.com/jessek/hashdeep/releases](https://github.com/jessek/hashdeep/releases) and download md5deep-4.4.zip.
- Open the .zip file and extract md5deep64.exe to your desktop.
- Open a command prompt window (in the start box, type “cmd”). Navigate to your desktop (cd Desktop).
- Type: nsrllookup –help
- Same commands as in previous slide but *use quotation marks around the file path in the command.*
Exercise: Using PRONOM

The PRONOM technical registry contains information about a wide variety of file formats, including versioning information. You can find it at [https://www.nationalarchives.gov.uk/PRONOM/Default.aspx](https://www.nationalarchives.gov.uk/PRONOM/Default.aspx). PRONOM has an online search feature that can be used to view the registry.

Click on “Search PRONOM” and navigate to the “File Format” tab. Clicking on the first search button (under “1. File Formats”) will allow you to view all of the entries in the registry.

DROID incorporates information from PRONOM. It also uses file magic and file format extensions to provide a “best effort” at identifying file types. If you’d like to know more about DROID, you can find a quick demonstration video at: [https://vimeo.com/24718678](https://vimeo.com/24718678)

Note: We’ll see DROID output in the Siegfried exercise later.
Creating and Extracting Forensic Metadata
High-Level View of Metadata Generation and Reporting

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
<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>

*Developed by Simson Garfinkel*
PREMIS (Preservation) Metadata Generated from Running BitCurator Tools – Recorded as PREMIS Events
This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<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 -X /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">
    <partition_offset>0</partition_offset>
    <block_size>2048</block_size>
    <ftype>2048</ftype>
    <ftype_str>iso9660</ftype_str>
    <block_count>36839</block_count>
  </volume>
</dfxml>
```
Identifying “Features” of Interest in Disk Images or Directories

Bulk Extractor
Bulk Extractor* – Identifying Potentially Sensitive Information


*Developed by Simson Garfinkel
bulk_extractor scan completed. See Status, below, for details.

Options

bulk_extractor has completed.
Report bulk_extractor-output has been opened and is ready for viewing.

Status

Elapsed time: 0.4985 sec.
Overall performance: 2.958 MBytes/sec.
Total email features found: 0
Done.

Close
Histogram of Email Addresses (Specific Instances in Context on Right)

your credit card number, so this information can only be viewed by Motorola. Motorola uses Secure Sockets Layer (SSL) encryption on technology, the highest level of security on the Internet. The SSL protocol provides server authentication, data integrity, and privacy on the Web. This security measure helps ensure that no imposters, eavesdroppers, or vandals get your personal information. SSL not only encrypts your personal and financial information transmitted, including credit card information, but also verifies the identity of the server and that the original message arrives safely at its destination. However, no data transmission over the Internet can be guaranteed to be 100% secure. As a result, while we strive to protect your personal information, Motorola cannot ensure or warrant the security of any information you transmit to us or from our Web site, and therefore you use our site at your own risk. Once we receive your transmission, we use our best effort to ensure its security on our systems. 000020 000021 000022 000023 As a global company Motorola has international sites and users all over the world. When you give Motorola personal information, that information may be sent electronically to servers outside of the country where you originally entered the information. In addition, that information may be used, stored and processed outside of the country where you entered that information. Whenever Motorola handles personal information, regard less of where this occurs, it takes steps to ensure that your information is treated securely and in accordance with the relevant Terms of Use and this Privacy Policy. How can I correct or change my personal information? If you would like to review, correct or change personal information you have provided, or remove your name from our mailing list, please e-mail us at privacy@motorola.com. If you have established a "user profile" on a Motorola website, you may change the information you provided at an
BitCurator Reporting Tool
Various Specialized BitCurator Reports
### Detail: Specialized BitCurator Reports

<table>
<thead>
<tr>
<th>File</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>bc_format_bargraph.pdf</td>
<td>histogram of file formats found on the volume</td>
</tr>
<tr>
<td>bulk_extractor_report.pdf</td>
<td>high-level overview of feature locations on disk</td>
</tr>
<tr>
<td>fiwalk_deleted_files.pdf</td>
<td>shows paths to any deleted materials found in a given partition</td>
</tr>
<tr>
<td>fiwalk-output.xml.xlsx</td>
<td>Excel converted DFXML output (file system metadata)</td>
</tr>
<tr>
<td>fiwalk_report.pdf</td>
<td>high-level overview of file system characteristics</td>
</tr>
<tr>
<td>format_table.pdf</td>
<td>long-form file format names for formats shown in bar graph</td>
</tr>
<tr>
<td>premis.xml</td>
<td>PREMIS preservation metadata</td>
</tr>
</tbody>
</table>
Nautilus Scripts

- Scripts that can be run using Nautilus (GNOME file manager)
- Most provide more convenient access (right click and menu selection) to functions performed by applications that could also be run directly
Right-click on file or directory and create MD5
Quick access to a hex view
## Other functionality to meet 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 (I addition to Guymager)</td>
<td>dd, dcfldd, ddrescue, cdrdao (for audio CDs)</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 user needs

<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>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>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>
Command Line Operations

- Opens up many more possibilities, such as:
  - stringing tools together
  - performing batch operations
  - changing parameters from their default values
  - using tools that are only available through the command line (no GUI)
Some Considerations

- Role of pipes – feed output from one process into another process
- Switches – settings that can be applied to a command (e.g. -a, -r)
- Argument – a specific piece of data that is processed by a program (e.g. a variable or fixed value)
- Regular expressions – used to find patterns (more on this later)
- Text created in Windows and Unix, even though they’re both ASCII, will encode new lines differently, so you may need to translate using a tool such as dos2unix or unix2dos.
Some Important Commands and Tasks

- **mkdir** – make a directory
- **cd** – change the directory that you’re in [“cd ..” goes to the parent of the current directory]
- **ls** – list contents of a directory
- **md5sum** – generate cryptographic hashes
- **cat** – output content of a text file (can be concatenation of contents of two files)
- **file** – determine file types based on magic numbers
- **strings** – matches patterns in the text (ASCII) parts of a file (file can be binary)
- **diff** – compare two files
- **hexdump** – very basic (non-GUI) hex viewer
General Unix/Linux CLI Tips

- **man** – manual page that explains how to run a command or some other technical information (e.g. ascii page)
- **control-z** – quit currently running program
- **clear** – clear the screen (hide text from previous commands)
- **Up arrow** – cycles through previous commands, so you can rerun (or adapt) them
- **Tab** – hit this key after you’ve started typing a string that the operating system can fill in for you (e.g. a long file name). Hitting tab multiple times will cycle through available options.
Exercise: Basic Linux Commands

- The saa-dfa-sample-data.zip file you downloaded earlier contains a folder of sample files named `file_ident_ex`
- If you haven’t done this already, add shared folder to BitCurator VM, pointing to the desktop of the host

- Move the `file_ident_ex` folder to the BitCurator VM desktop
Exercise: Basic Linux Commands

Open a terminal in the BitCurator environment (using the Terminal icon in the dock)
## Exercise: Basic Linux Commands

<table>
<thead>
<tr>
<th>Command</th>
<th>Reason/Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>pwd</code></td>
<td>Show the directory you’re currently in</td>
</tr>
<tr>
<td><code>ls</code></td>
<td>List the contents of the current directory</td>
</tr>
<tr>
<td><code>cd Desktop</code></td>
<td>Change the current directory to Desktop</td>
</tr>
<tr>
<td><code>ls</code></td>
<td>List the contents of the current directory</td>
</tr>
<tr>
<td><code>unzip files.zip</code></td>
<td>Decompress and unpack content of files.zip</td>
</tr>
<tr>
<td><code>ls</code></td>
<td>List the contents of the current directory</td>
</tr>
<tr>
<td><code>cd files</code></td>
<td>Change the current directory to files</td>
</tr>
<tr>
<td><code>ls</code></td>
<td>List the contents of the current directory</td>
</tr>
<tr>
<td><code>md5sum [file name of first file] &gt; firsthash</code></td>
<td>Create a hash of a file and output it to a text file</td>
</tr>
<tr>
<td><code>less firsthash</code></td>
<td>Display the content of the output to the screen</td>
</tr>
<tr>
<td>Control-z</td>
<td>Stop the “less” program</td>
</tr>
<tr>
<td><code>md5sum [file name of second file] &gt; secondhash</code></td>
<td>Create a hash of a second file and output it to a text file</td>
</tr>
<tr>
<td><code>cat firsthash secondhash &gt; bothhashes</code></td>
<td>Combine the context of the two output files</td>
</tr>
<tr>
<td><code>more bothhashes</code></td>
<td>Display the content of the output to the screen</td>
</tr>
<tr>
<td><code>most bothhashes</code></td>
<td>Display the content of the output to the screen (follow instructions for adding it), then run this command again</td>
</tr>
</tbody>
</table>

**Gives you the right administrative permissions**

**Uses Advanced Packaging Tool to get the program**

```
sudo apt-get install most
```
# Exercise: Basic Linux Commands

<table>
<thead>
<tr>
<th>Command</th>
<th>Reason/Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>rm firsthash</td>
<td>Delete (remove) firsthash file</td>
</tr>
<tr>
<td>rm secondhash</td>
<td>Delete (remove) secondhash</td>
</tr>
<tr>
<td>ls</td>
<td>List the contents of the current directory</td>
</tr>
<tr>
<td>hexdump [file name] -C</td>
<td>less</td>
</tr>
<tr>
<td>Use up and down arrows</td>
<td>Navigate within the hex view of the file's content</td>
</tr>
<tr>
<td>:q</td>
<td>Quit the “less” program</td>
</tr>
</tbody>
</table>
Exercise: Siegfried and Brunnhilde

- Over the next few slides, we will run Siegfried and Brunnhilde in over the same set of files in several different ways

- Goals
  - Generate characterization and related technical metadata
  - Illustrate how the data can be configured for different uses
  - Identify decision points when data are unclear

- Source files to analyze: file_ident_ex directory in the zip file you downloaded and extracted earlier
Installing Siegfried and Brunnhilde

- Start up the BitCurator VM (if it’s not already running)
- Open a Terminal Window
Installing Siegfried and Brunnhilde

Enter the following commands in Terminal:

- `wget -qO- https://bintray.com/user/downloadSubjectPublicKey?username=bintray | sudo apt-key add`
- `echo "deb http://dl.bintray.com/siegfried/debian wheezy main" | sudo tee -a /etc/apt/sources.list`
- `sudo apt-get update && sudo apt-get install Siegfried`
- `sudo pip install brunnhilde`

Note: there is a text file in the sample files that includes these commands if you want to copy/paste them
Running Siegfried and Brunnhilde I

- Ensure you have extracted the `saa-dfa-sample-data.zip` file
- Create a shared folder (we’ll step you through this)
- Start up the BitCurator VM
- Create “sieg-out” and “brunn-out” folders on the desktop
- Drag “file_ident_ex” folder from the shared `saa-dfa-sample-data` directory to the Desktop
Running Siegfried and Brunnhilde II

- Open a Terminal window
- At the prompt, enter the following command:
  - `sf ~/Desktop/file_ident_ex/ > ~/Desktop/sieg_out/sieg_out.yaml`
- Open the sieg_out directory and look around
  - What did the command do?
  - What does the file tell you?
  - How would you characterize the data presented in the file?
  - Does anything strike you as odd? Particularly useful?
Running Siegfried and Brunnhilde III

- In the same Terminal window, enter the following commands:
  - `sf -droid ~/Desktop/file_ident_ex/ > ~/Desktop/sieg_out/sieg_out-droid.csv`
  - `sf -json ~/Desktop/file_ident_ex/ > ~/Desktop/sieg_out/sieg_out-json.json`

- Open the sieg_out directory and look around
  - What did the commands do?
  - How do these files differ from the one created on the previous slide?
  - Between the three output files, which do you think is most useful presentation of the data? (Hint: there may be more than one answer)
  - Do any of these files strike you as particularly useful? Particularly worthless?
Running Siegfried and Brunnhilde IV

- In the same Terminal window, enter the following command:
  - `brunnhilde.py -w ~/Desktop/file_ident_ex/ ~/Desktop/brunn_out/DAS-FileIdent`

- Open the `brunn_out` directory and look around
  - Did Brunnhilde perform any tasks over and above Siegfried?
  - How do the Brunnhilde output files differ from those generated by Siegfried?
  - Inspect `csv_reports`. How would you characterize what you see here?
  - Are the files here that Brunnhilde and Siegfried found problematic? What conclusions might you draw from them?
  - Is there information that Brunnhilde highlighted that you missed in Siegfried’s output?
Regular Expressions

- What is a regular expression, or regex?
  - Simply a pattern for matching bits of text.

- What are regex’s useful for?
  - Three things:
    1. *Matching*: Does this text contain a pattern?
    2. *Replacement*: Replace some part of the text with other text
    3. *Extraction*: Yanking out a bit of the text to use somewhere else.
Regular Expressions

- Regular expressions may contain ordinary letters, numbers, and *a few special symbols* that allow you to match a wide range of patterns with a small amount of syntax.

- A regex needs to be interpreted by a program (such as a Perl or Python script) or by an application (such as the Forensic Toolkit).

- Regular expression syntax may look different in different languages and programs.
## Regular Expressions – Special Characters

<table>
<thead>
<tr>
<th>Character</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>.</td>
<td>Match anything</td>
</tr>
<tr>
<td>+</td>
<td>Match one or more occurrences</td>
</tr>
<tr>
<td>*</td>
<td>Match zero or more occurrences</td>
</tr>
<tr>
<td>^</td>
<td>Match only at the start of the text</td>
</tr>
<tr>
<td>$</td>
<td>Match only at the end of text</td>
</tr>
<tr>
<td>\w</td>
<td>Match an alphanumeric word</td>
</tr>
<tr>
<td>\d</td>
<td>Match a number</td>
</tr>
<tr>
<td>\s</td>
<td>Match any whitespace</td>
</tr>
<tr>
<td>\S</td>
<td>Match anything except whitespace</td>
</tr>
</tbody>
</table>
## Regular Expressions - Examples

<table>
<thead>
<tr>
<th>Expression</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>r+</td>
<td>Match the letter 'r' one or more times</td>
</tr>
<tr>
<td>t*</td>
<td>Match the letter 't' zero or more times</td>
</tr>
<tr>
<td>\d</td>
<td>Match any single digit (shorthand for the expression [0-9])</td>
</tr>
<tr>
<td>\dd</td>
<td>Match any pair of digits (alternatively, [0-9][0-9])</td>
</tr>
</tbody>
</table>
What's "wrong" with the second and third regex patterns shown here? What else might they match?
Regular Expressions in FTK

- Tools for building regular expressions one part at a time
- Various default regular expressions that you can use or adapt
For Fun on Your Own – Regex Golf!

Try it at https://alf.nu/RegexGolf
Extracting Data From Specific Types of Files
**Exchangeable Image File Format (EXIF)**

- Possible tags:
  
  [https://exiftool.org/TagNames/EXIF.html](https://exiftool.org/TagNames/EXIF.html)

<table>
<thead>
<tr>
<th>Camera manufacturer</th>
<th>Canon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera model</td>
<td>Canon EOS 1200D</td>
</tr>
<tr>
<td>Author</td>
<td>Praveen. P</td>
</tr>
<tr>
<td>Exposure time</td>
<td>1/60 sec (0.016666...</td>
</tr>
<tr>
<td>F-number</td>
<td>f/11</td>
</tr>
<tr>
<td>ISO speed rating</td>
<td>200</td>
</tr>
<tr>
<td>Date and time of data generation</td>
<td>22:29, 22 November 2018</td>
</tr>
<tr>
<td>Lens focal length</td>
<td>41 mm</td>
</tr>
</tbody>
</table>
Example of EXIF Metadata from a JPEG File (Generated Using exiftool*)

---- ExifTool ----
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
Y Cb Cr Sub Sampling : YCbCr4:2:0 (2 2)
---- GPS ----
GPS Img Direction : 83
GPS Img Direction Ref : Magnetic North
GPS Latitude Ref : North
GPS Latitude : 35 deg 55' 2.24"
GPS Longitude Ref : West
GPS Longitude : 79 deg 2' 57.55"
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
Y Cb Cr 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 Method : 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 :
Exif Image Width : 2592
Max Aperture Value : 2.6
Sharpness : Normal
Exif Image Height : 1944
Exif Image Height : 1944
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
Thumbnail Image : (Binary data 7164 bytes, use -b option to extract)
Focal Length : 3.4 mm
Light Value : 6.7

*https://exiftool.org/TagNames/EXIF.html (Also available through the BitCurator environment)
ExifTool Exercise in BitCurator

- Start up the BitCurator VM
- Download one or more pictures to your desktop that you’d like to examine
- Options for viewing EXIF:
  1. **PyEXIFToolGUI**:
     - Navigate to Desktop > Forensics Tools > PyEXIFToolGUI
     - Open the tool and select File > Load Images
     - Let’s also add some GPS coordinates: Select Edit Data, enter the values, then select Save to Selected Image(s) [Make sure that the image is selected]
  2. **File info menu**:
     - Navigate to the image file
     - Right click on it and select Scripts > File Info > Meta Information [Pick EXIF Data]
  3. **exiftool at the command line**:
     - Open a command prompt window
     - Navigate to where you stored the image (e.g. cd Desktop)
     - Type: `exiftool [Filename]`
     - Note: You can scroll up and down by using Shift + Page Up/Page Down, or you can invoke the command as `exiftool [Filename] | less` (type “q” to quit)
Optional Exiftool Exercise (Windows or Mac)

- Download one or more pictures to your desktop that you’d like to examine
- Download and unzip the latest Windows executable (or Mac package) from: https://exiftool.org/
- Save exiftool(-k).exe to your desktop
- Change file name to: exiftool(-k -a -u -g1 -w txt).exe [NOTE: This is changing the parameters for running the software – same as if you were to add these switches at the command line. This trick might not work on a Mac, but you can always issue the commands directly.]

- -k = pause the program before terminating
- -a = allow extraction of duplicate tags
- -u = extract unknown tags
- -g1 = organize output by tag group
- -w = write output text file

- Drag and drop pictures onto the exiftool icon and examine the results
- Change file name to: exiftool(-X -k -a -u -g1 -w xml).exe
- Drag and drop pictures onto the exiftool icon and examine the results
- For more about exiftool, see: https://github.com/exiftool/exiftool
### Stripping Metadata From Images

[Table showing results of stripping metadata from different social media sites]

<table>
<thead>
<tr>
<th>Social Media site/system</th>
<th>Summary</th>
<th>Displays correctly?</th>
<th>Displays 4Cs?</th>
<th>Save As embedded?</th>
<th>Download embedded?</th>
</tr>
</thead>
<tbody>
<tr>
<td>500ps - <a href="http://www.500ps.com">www.500ps.com</a></td>
<td>Some embedded metadata fields are shown, all correctly, but not the rights-relevant 4Cs. Metadata preserved in SaveAs file. Compared to 2013: SaveAs preserves metadata now = improvement</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>BEHANCE - <a href="http://www.behance.net">www.behance.net</a></td>
<td>All rights-relevant fields and more are shown, all correctly, embedded metadata is preserved in the SaveAs and the downloaded image file. Compared to 2013: not tested then</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Dropbox - <a href="http://www.dropbox.com">www.dropbox.com</a></td>
<td>No embedded metadata shown. Embedded metadata only preserved in the downloaded image file but not in the SaveAs. Compared to 2013: also SaveAs files preserved metadata then = decline</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Eyevan - <a href="http://www.sysm.com">www.sysm.com</a></td>
<td>No embedded metadata shown. SaveAs file was downscaled and all metadata was stripped off. Compared to 2013: not tested then</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Facebook - <a href="http://www.facebook.com">www.facebook.com</a></td>
<td>No embedded metadata shown. SaveAs file preserved Copyright, Notice and Creator in IIM, anything else is stripped off. Surprises 2: IIM fields contain data generated by Facebook. Compared to 2013: at least 2 fields in IIM survive now = slight improvement</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Flickr - FREE account - <a href="http://www.flickr.com">www.flickr.com</a></td>
<td>Some embedded metadata fields are shown, all correctly, but not all rights-relevant 4Cs. Embedded metadata is stripped off SaveAs files but preserved in downloaded files. Compared to 2013: plus - any downloaded file preserves metadata now; minus - even high resolution SaveAs file does not preserve it now.</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Google Photos - photos.google.com</td>
<td>Some embedded metadata fields are shown, all correctly, but not all rights-relevant 4Cs. SaveAs works only for downscaled files - only EXIF metadata is preserved. Downloaded files preserved all metadata. Compared to 2013: Google + photos: SaveAs file sets IIM and XMP metadata stripped off now = decline</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Imply - <a href="http://www.imply.com">www.imply.com</a></td>
<td>No embedded metadata shown. Embedded metadata is preserved in the high resolution/original size SaveAs image file but stripped off in a downscaled file. Compared to 2013 the loss of metadata in downscaled images was not tested in 2013.</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Instagram - instgram.com</td>
<td>Tested using the Instagram iOS app v 6.4.11. No embedded metadata fields are shown. No retrieval of image files possible. Compared to 2013: then SaveAs was possible - with stripped off metadata</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Joomeo - <a href="http://www.joomeo.com">www.joomeo.com</a></td>
<td>Some embedded metadata fields are shown, all correctly, but not the rights-relevant 4Cs. Embedded metadata preserved in the downloaded image file. Compared to 2013: more embedded metadata were shown than, including 4Cs = slight decline</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>LINKED IN 2015 - <a href="http://www.linkedin.com">www.linkedin.com</a></td>
<td>No embedded metadata shown. Only embedded EXIF fields are preserved in SaveAs files. Compared to 2013: not tested then.</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Pictify - <a href="http://www.pictify.com">www.pictify.com</a></td>
<td>No embedded metadata shown. No retrieval of image file possible. Compared to 2013: then SaveAs was possible - with stripped off metadata.</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Pinterest - <a href="http://www.pinterest.com">www.pinterest.com</a></td>
<td>No embedded metadata shown. Embedded metadata preserved in high resolution/original size images, but IIM and XMP metadata is</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>✅</td>
</tr>
</tbody>
</table>

[Link to the full test results page: https://www.embeddedmetadata.org/social-media-test-results.php]
Are the “new” office formats (ODF and OOXML) better or worse for forensics?

What kinds of information can you get out of them?

What sorts of approaches might you take to view and/or extract the information?
Office Documents – PPTX File Example


- Change the file extension to .zip
- Open it with 7-Zip or WinZip
- Extract all the files
- Examine the contents of the resulting directory
  - Can you find a thumbnail of the first slide?
  - Where are the slides stored?
  - Where are embedded images stored?
  - Can you determine who created the file?
Jonathan Larson Fast Save Example

https://www.nypl.org/blog/2011/04/22/no-day-today-look-jonathan-larsons-word-files
Hidden Data Exercise – Using a Hex Editor


- Open the file in HxD (or upload to https://hexed.it)
- Go to offset 73B90 (use Search > Goto or just Control+G)
- What do you see there?
- What does it tell you about the document?
Hidden Data Exercise – Inspection in MS Word

- Do the following:
  - Open it in Word – what is it?
  - If prompted to do so at the top, select “Enable Editing”
  - Select: File > Options > Trust Center > Trust Center Settings...
  - Then Privacy Options > Document Inspector > Inspect
Hidden Data Exercise – Inspection in MS Word

- Are you prompted with this?
- Why do you think this is?
- If you see this, click OK, then save the document
- Run Document Inspector again
- What does it tell you?
Email

- What’s in an email header?
- Which parts of the header would be of most interest to you as someone responsible for managing and preserving a collection that includes email?
- Which parts of the header would be of most interest to future researchers?
Windows Artifacts
Desktop Operating System Market Share

https://gs.statcounter.com/os-market-share/desktop/worldwide
Let’s make sure you can see all the files on your computer.
After you've opened some files, we'll show the most recent ones here.
File Explorer

Navigation pane:
- Desktop
- Downloads
- Documents
- Pictures
- Music
- Videos
- OneDrive
- This PC
- Network

Panes:
- Desktop
- Downloads
- Documents
- Pictures
- Music
- Videos
- OneDrive
- This PC
- Network

Layout:
- Desktop
- This PC
- Downloads
- This PC
- Documents
- This PC
- Pictures
- This PC
- Music
- This PC
- Videos
- This PC

Recent files (0)

After you've opened some files, we'll show the most recent ones here.

6 items
You have chosen to display protected operating system files (files labeled System and Hidden) in File Explorer.

These files are required to start and run Windows. Deleting or editing them can make your computer inoperable.

Are you sure you want to display these files?

Options:
- Hide protected operating system files (Recommended)
- Launch folder windows in a separate process
- Restore previous folder windows at logon
- Show drive letters
- Show encrypted or compressed NTFS files in color
- Show popup description for folder and desktop items
- Show preview handlers in preview pane
- Show status bar
- Show sync provider notifications

Restore Defaults  OK  Cancel  Apply
Windows Registry

- Information about:
  - Applications installed
  - Application settings
  - Hardware installed
  - Hardware settings
  - User interface and system preferences
  - User accounts
  - Locations of files and recent activities, e.g. Most Recently Used (MRU)
  - Lots of online activities, e.g. usernames and passwords, browsing and search query history
# Five Main Registry Files

<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTUSER.DAT</td>
<td>One for each user account, includes information such as Most Recently Used (MRU) file lists, desktop settings, default application behaviors</td>
</tr>
<tr>
<td>SAM (Security Accounts Manager)</td>
<td>User account information (including passwords) and security settings</td>
</tr>
<tr>
<td>SECURITY</td>
<td>User and group security policies, e.g. which accounts can load device drivers, get remote access to the machine</td>
</tr>
<tr>
<td>SOFTWARE</td>
<td>Information about all install programs, including settings and directory paths</td>
</tr>
<tr>
<td>SYSTEM</td>
<td>Windows systems settings, such as drive letter mappings, storage volume information, system boot profile, last known good configuration, system name, Windows setup information, hardware profile information</td>
</tr>
</tbody>
</table>
Where are They Located?
Registry Hives

Structure:

Hive
  • Key
    • Subkey
      • Value

Example:
HKEY_CURRENT_USER\Software\Microsoft\Windows\CurrentVersion\Explorer\RecentDocs

What do you think this is?
## Registry Hives

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HKEY_CLASSES_ROOT</td>
<td>Just points to HKEY_LOCAL_MACHINE\Software\Classes</td>
</tr>
<tr>
<td>HKEY_CURRENT_USER</td>
<td>User setting information, which is generated dynamically from HKEY_USERS when a user logs into Windows</td>
</tr>
<tr>
<td>HKEY_LOCAL_MACHINE</td>
<td>Hardware and software settings that are specific to this computer but shared across users (generated at startup from SYSTEM.DAT)</td>
</tr>
<tr>
<td>HKEY_USERS</td>
<td>Information about each of the user accounts on the computer, e.g. desktop settings, default software behaviors - generated at startup from NTUSER.DAT files, and when user logs out of applications or out of Windows, data are written back to the ntUSER.DAT files</td>
</tr>
<tr>
<td>HKEY_CURRENT_CONFIG</td>
<td>Just points to HKEY_LOCAL_MACHINE\Config</td>
</tr>
</tbody>
</table>

Question: Where would you find these registry hives on a disk image? (Hint: This is a trick question)
## Registry Hive Value Types

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>REG_BINARY</td>
<td>Raw binary data displayed as hexadecimal*</td>
</tr>
<tr>
<td>REG_DWORD</td>
<td>32-bit unsigned integer (4 bytes)</td>
</tr>
<tr>
<td>REG_EXPAND_SZ</td>
<td>Variable-length string, usually in UTF-16 (Unicode)</td>
</tr>
<tr>
<td>REG_FULL_RESOURCE_DESCRIPTOR</td>
<td>Series of nested arrays used by a hardware device, binary data displayed as hexadecimal*</td>
</tr>
<tr>
<td>REG_LINK</td>
<td>Symbolic link to another registry key (Unicode)</td>
</tr>
<tr>
<td>REG_MULTI_SZ</td>
<td>Ordered list of strings (multi-string value), usually in UTF-16</td>
</tr>
<tr>
<td>REG_NONE</td>
<td>No specific type – displayed as hexadecimal*</td>
</tr>
<tr>
<td>REG_QWORD</td>
<td>64-bit integer (8 bytes)</td>
</tr>
<tr>
<td>REG_RESOURCE_LIST</td>
<td>Series of nested arrays used by a hardware device, binary data displayed as hexadecimal*</td>
</tr>
<tr>
<td>REG_RESOURCE_REQUIREMENTS_LIST</td>
<td>Series of nested arrays used by a hardware device, binary data displayed as hexadecimal*</td>
</tr>
<tr>
<td>REG_SZ</td>
<td>Fixed-length text string, usually in UTF-16</td>
</tr>
</tbody>
</table>

* Can be opened and viewed in a hex editor
Security ID (SID)

- One assigned to each user account
- Associated with various resources, including files, folders and Recycling Bins
SID Example

S-1-5-21-1180590209-877416012-3186324384-1002
Always an “S”, indicating that this is an SID.
Revision level (version of the SID specification being used).
Authority that issued the SID. Value is usually “5”, indicating NT Authority.
Domain identifier – value can be up to 500.
Account or group on a domain or local machine
Relative Identifier (RID), designating a specific user in the SAM file. Those below 1000 are default accounts (e.g. 500 = Administrator), and those 1000 or above are created for specific groups or users.
Examining an NTUSER.DAT File

- The files in registry (a folder within the zip file you downloaded earlier) were extracted from a full-drive (including the operating system) disk image

- The following is an example of how these files can be extracted using FTK Imager
- Navigate to: Partition 1 > [root] > Documents and Settings > Charlie > NTUSER.DAT
- Right click on NTUSER.DAT and select Export Files.
Then export the other four registry files from Windows\System32\config
Perform these same tasks in the BitCurator environment
RegRipper Instructions - BitCurator

- Navigate to Forensics Tools, and click on the RegRipper icon
- NOTE: IGNORE examples that it presents, because they use commands and syntax for Windows, not Linux
- Issue each of the following commands:

  perl rip.pl -r ~/Desktop/sample-data/registry/NTUSER.DAT > ~/Desktop/ntuser-report -f ntuser
  perl rip.pl -r ~/Desktop/sample-data/registry/SAM > ~/Desktop/sam-report -f sam
  perl rip.pl -r ~/Desktop/sample-data/registry/SOFTWARE > ~/Desktop/software-report -f software
  perl rip.pl -r ~/Desktop/sample-data/registry/SYSTEM > ~/Desktop/system-report -f system

*Enter each command in its entirety before hitting enter (line breaks above are simply to fit the text onto the slide, not ones that you should type yourself). Remember that the up arrow and tab can save you time when typing commands.
• Create a folder called regripper-exercise on your desktop
• Find the registry directory in the folder you extracted from the saa-das-sample-data zip file earlier
• Copy this folder to the regripper-exercise folder on your Desktop
RegRipper Instructions – Windows II

• Navigate to saa-dfa-sample-data\RegRipper3.0
• Run rr.exe (it may simply appear as rr) by double clicking it.
• The next set of steps will be run 5 times - once for each of the files in regripper-exercise\registry
• Next to the Hive File window, select Browse
  • Navigate to regripper-exercise\registry and select the first Hive File
  • E.g., NTUSER.DAT
• Next to Report File, select Browse
  • Create a new file in regripper-exercise that corresponds to the Hive File above
  • E.g., NTUSER_report.txt
• In the Profile dropdown, select the appropriate profile
  • E.g., ntuser profile selection is not required in RegRipper 3.0
• Select Rip It.
• Repeat the above steps for SAM, SECURITY, SOFTWARE, and SYSTEM
## RegRipper Output Questions

<table>
<thead>
<tr>
<th>Report</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>ntuser-report.txt</td>
<td>Are you able to identify files that the user recently opened? If so, what were they? Can you determine what the most recently opened files of specific types (e.g. txt) were?</td>
</tr>
<tr>
<td>sam-report.txt</td>
<td>How many accounts were there on the computer that is represented in the disk image? What is the Relative Identifier (RID) for the user account you’re examining? What other interesting information can you gain from the SAM report about this user account and how might you use that information?</td>
</tr>
<tr>
<td>security-report.txt</td>
<td>What is the Machine SID for the computer represented in the disk image? Why would you want to know this? How does it relate to the RID that you identified above?</td>
</tr>
<tr>
<td>software-report.txt</td>
<td>Identify three different applications that were installed on the computer and the file paths where the applications were stored.</td>
</tr>
<tr>
<td>system-report.txt</td>
<td>Find the devclass output. What does this output tell you? How might this information be useful?</td>
</tr>
</tbody>
</table>
RegRipper Output Discussion – ntuser-report

• Are you able to identify the files that the user recently opened? If so, what were they?
  • How did you go about finding this information?
  • What line number(s) points to this information?

• Can you determine what the most recently open files of specific types (e.g. txt) were?
  • How did you go about finding these?
  • What line numbers have this information?

• Look at lines 1109-1117 - what type of information are you looking at?

• Is there any other information you find particularly compelling in this report?

• What might you do with this information?
RegRipper Output Discussion – sam-report

- How many accounts were there on the this computer?
  - How did you go about finding this information?
  - What line number(s) points to this information?

- What was the Relative Identifier (RID) for the user account you’re examining?
  - How did you go about finding this?

- How many logins did Pat make on this machine?

- Is there any other information you find particularly compelling in this report?

- What might you do with this information?
RegRipper Output Discussion – security-report

• What is the Machine SID for the computer represented here?
  • How did you go about finding this information?
  • What line number(s) points to this information?

• Why would you want to know this information

• How does this relate to the RID in the previous report?
RegRipper Output Discussion – software-report

• Identify three different applications that were installed on this computer
  • How did you go about finding this information?
  • What line number(s) points to this information?

• Why would you want to know this information?

• How might it aid description?

```
1       Launching appinitdlls v.20130425
2       appinitdlls v.20130425
3       (Software) Gets contents of AppInit_DLLs value
4
5       AppInit_DLLs
6       Microsoft\Windows NT\CurrentVersion\Windows
7       LastWrite Time Fri Nov 20 18:55:34 2009 (UTC)
8       AppInit_DLLs : {blank}
9       LoadAppInit_DLLs : 1
10      *LoadAppInit_DLLs value globally enables/disables AppInit_DLLS.
11      0 = disabled (default)
12
13      Wow6432Node\Microsoft\Windows NT\CurrentVersion\Windows not found.
14      Analysis Tip: The AppInit_DLLs value should be blank; any DLL listed
15      is launched with each user-mode process.
16      ---------------------------------
17      apppaths v.20120524
18      (Software) Gets content of App Paths subkeys
19
20      App Paths
21      Microsoft\Windows\CurrentVersion\App Paths
```
RegRipper Output Discussion – system-report

- Find the devclass output
- What does this output tell you?
- How might this information be useful?

```
ControlSet001\Control\Session Manager\AppCertDlls not found.
-------------------------------------------------------------------
appcompatcache v.20130425
(System) Parse files from System hive Shim Cache

Signature: Oxdeadbeef
WinXP, 32-bit
C:\Program Files\AVG\AVG9\avgsrcmax.exe
ModTime: Mon Dec 7 23:47:51 2009 Z
UpdTime: Tue Dec 8 17:52:13 2009 Z
Size : 361752 bytes

C:\Program Files\RealVNC\VNC4\winvnc4.exe
ModTime: Thu Oct 16 01:13:58 2008 Z
UpdTime: Tue Dec 8 01:45:23 2009 Z
Size : 439632 bytes

C:\RAM\mddbak.exe
ModTime: Sat Nov 14 01:07:38 2009 Z
UpdTime: Sun Dec 6 16:10:51 2009 Z
Size : 95104 bytes

C:\Program Files\Java\jre6\bin\jqsnotify.exe
ModTime: Sun Oct 11 12:17:34 2009 Z
```
Viewing and Copying Registry Information if You're Running the Original Environment

- What if you’re logged in to the original computer? How might you get information out of the registry?
- What if you wanted to replicate that registry information on another computer?
- Hint: There are tools built into Windows for this.
Restore Points

- Snapshots of Registry hives and some other essential system (including .EXE, .INI, .LNK) files. They’re created:
  - when there are major system changes, e.g. installing software
  - at regularly scheduled intervals
  - if the user manually creates one

- Let’s look at some restore points: Start Button > All Programs > Accessories > System Tools > System Restore [or just "System Restore" in the Start box]
Examining the Recycle Bin

1. In the start menu box, type **cmd**
2. Type: **cd c:\$recycle.bin**
   (What is this doing?)
3. Type **dir /a**
   (What is this doing?)
4. Type **dir *.* /s**
   (What is this doing?)
5. Put one or more files into the Recycle Bin
   (by moving there or by deleting)
6. Repeats steps 2-4. What do you see now?
A Brief Discussion of Mac Forensics

- No Registry, so where is all the good stuff stored?

- See: [https://forensicswiki.xyz/wiki/index.php?title=Mac_OS_X_10.9_-_Artifacts_Location](https://forensicswiki.xyz/wiki/index.php?title=Mac_OS_X_10.9_-_Artifacts_Location) but note that this is information a snapshot in time; artifact locations tend to change between versions of macOS.
Archival Importance and Role of SID

- If the volume is NTFS, you can find the SID associated with a specific file.
- If you also have registry files from the original computer (particularly SAM.DAT), you can get information associated with that SID, such as the name of the user/group, last time he/she logged in, and various other account details.
setuplog.txt

• See disk image example below: Partition 1 > [root] > WINDOWS > setuplog.txt

• What do you see in this file?
• What information could be useful for digital curation?
• When/how might you use it?
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

*Note: The first three were discussed earlier
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 Redaction Tools

- Software to redact strings and byte sequences identified in disk images
- Three types of redaction actions:
  - SCRUB (overwrite the bytes in the target with zeroes),
  - FILL (overwrite by filling with a given character),
  - FUZZ (altering the content of a binary, so it can no longer run).
- Best used through a command-line interface but also include a graphic user interface (GUI) that supports the same functions
- Python API allowing institutions to develop custom redaction facilities using open-source tools including lightgrep

https://github.com/bitcurator/bitcurator-access-redaction
• Funded by Andrew W. Mellon Foundation: October 1, 2016 – September 30, 2018
• Develop software for collecting institutions to extract, analyze, and produce reports on features of interest in text extracted from born-digital materials
• Use existing natural language processing software libraries to identify and report on those items likely to be relevant to ongoing preservation, information organization, and access activities
• May include entities (e.g. persons, places, and organizations), potential relationships among entities (e.g. appear together within documents or set of documents), and topic models to provide insight into how concepts are naturally clustered within the documents.
This is showing topic modeling, which we’ll look at in more detail soon.
Another path at this point is to feed the text into spacy for named-entity recognition.
BitCurator Access Webtools

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
BitCurator Access Webtools
BitCurator Access Webtools

fourpartusb1.E01

Format: EnCase 6
Size: 3.70B
Sectors: 7821312
Blocka/Sector: 512
MD5: 24619c765b59f6b6667a9e939f1ea1354
SHA-1: dfbaa99541f8807a3f0c3200d7f9a212550c5

Download:

Partitions

<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>File System</th>
<th>Start</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>fourpartusb1.E01</td>
<td>Win95 FAT32 (0x0b)</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>fourpartusb1.E01</td>
<td>Win95 FAT32 (0x0b)</td>
<td>1955331</td>
</tr>
<tr>
<td>11</td>
<td>fourpartusb1.E01</td>
<td>Mac OS X HFS (0x0f)</td>
<td>3910660</td>
</tr>
<tr>
<td>12</td>
<td>fourpartusb1.E01</td>
<td>Linux (0x83)</td>
<td>5867520</td>
</tr>
</tbody>
</table>

Showing 1 to 4 of 4 entries
## BitCurator Access Webtools

![Directory Listing](image)

### Directory Listing

<table>
<thead>
<tr>
<th>Type</th>
<th>Filename</th>
<th>Bytes</th>
<th>Created</th>
<th>Modified</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TESTFAT (Volume Label Entry)</td>
<td>0.0B</td>
<td>N/A</td>
<td>2013-05-02T16:01:26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.Trashes</td>
<td>4.0KB</td>
<td>N/A</td>
<td>2013-05-02T14:11:00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.RASHE=1,YEN</td>
<td>0.0B</td>
<td>N/A</td>
<td>2013-05-02T14:11:00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.Trashes</td>
<td>4.0KB</td>
<td>N/A</td>
<td>2013-05-02T14:11:00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.Spotlight-V100</td>
<td>4.0KB</td>
<td>N/A</td>
<td>2013-05-02T14:11:00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>08311397.pdf</td>
<td>2.2MB</td>
<td>N/A</td>
<td>2013-05-02T15:56:06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2013-02-20_AAFS.pdf</td>
<td>6.2MB</td>
<td>N/A</td>
<td>2013-05-02T15:56:36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.fseventsd</td>
<td>4.0KB</td>
<td>N/A</td>
<td>2013-05-02T16:01:26</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$MBR</td>
<td>512.0B</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$FAT1</td>
<td>953.0KB</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$FAT2</td>
<td>953.0KB</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$OrphanFiles</td>
<td>0.0B</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

Showing 1 to 12 of 12 entries
File Analysis for 2013-02-20_AAFS.pdf

<table>
<thead>
<tr>
<th>File Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extension: .pdf</td>
</tr>
<tr>
<td>Size: 6476327</td>
</tr>
<tr>
<td>SHA1: 0364598548ca19ddeb1d4f89990a4f21e8f44e5b9</td>
</tr>
<tr>
<td>MIME: application/pdf</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Full Text</th>
</tr>
</thead>
</table>
| 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. Garnkel, 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 |
# File Analysis for 2013-02-20_AAFS.pdf

## File Details
- **Extension**: .pdf
- **Size**: 6476327
- **SHA1**: 0364598548ca19deb1d4f89990af421e8f44e5b9
- **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. Garinkel, Kristina Foster, Joel Young*

*Naval Postgraduate School*

*Kevin Fairbanks, Johns Hopkins Applied Physics Lab*

[http://simson.net/](http://simson.net/)

---

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

*Simson L. Garinkel, Kristina Foster, Joel Young*

*Naval Postgraduate School*

*Kevin Fairbanks, Johns Hopkins Applied Physics Lab*

[http://simson.net/](http://simson.net/)
Topic Modeling in bitcurator-nlp-gentm (using pyLDAvis)
Forensic Feature Extraction and Cross-Drive Analysis

Incorporating digital forensics into archival workflows
Storage Media Acquisition and Handling Profile for Digital Repositories*

BitCurator-Supported Workflow

See: https://bitcurator.net/
Five Sources of Workflow Examples


https://dcs.library.virginia.edu/files/2013/02/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
OSSArcFlow

Investigating, Synchronizing, and Modeling a Range of Archival Workflows for Born-Digital Content

Project Abstract
The Educoopia 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

https://educopia.org/research/ossarcflow
Challenges

- Incorporation into LAM workflows, e.g. metadata conventions, connections to collection management systems
- Obsolete storage media and filesystems
- Dealing with large, internally complex data files
- Provision of public access
- Defining and implementing ethical commitments
“A community-based approach would use SWAT sites wherein a few self-selected institutions acquire and maintain the gear and expertise to read data and transfer content from particular types of obsolete media. The SWAT sites would provide transfer services for institutions that don’t have the capacity to read a particular medium (or the SWAT sites might become the likely places to deposit particular types of media).”


See also:
Legal and Ethical Issues
Ethical Dilemmas

- What ethical dilemmas related to born-digital materials have you faced or do you expect to face?
- What would the competing interests or values be?
- How would you decide?
Donor Agreements

- Donor agreements (as of 2012) tend not to address the kinds of issues raised in this class*
- What are the most important issues to resolve with creators/donors?
- What’s the right level of detail to include in donor agreements and discussions with potential donors?

Specific Guidance Documents


Text added in 2013:

“Be aware that any digital materials that you donate, including computers, computer disks, and other digital storage media, may contain passwords, web browsing history, other users’ files, and copies of seemingly deleted files. Whether or not these files are apparent to researchers will depend on the initial method of transfer and on the repository’s access policies and procedures for handling digital material, which may change over time as technology evolves. Discuss any concerns about deleted content with the archivist or curator.”

*https://www2.archivists.org/publications/brochures/deeds-of-gift
What does it mean for electronically stored information (ESI) to be “accessible”?
“The person responding need not provide discovery of electronically stored information from sources that the person identifies as not reasonably accessible because of undue burden or cost.” (Rule 45 (d)(1)(D)) (emphasis added)
Judge Shira Scheindlin:

"[t]he more information there is to discover, the more expensive it is to discover all the relevant information until, in the end, 'discovery is not just about uncovering the truth, but also about how much of the truth the parties can afford to disinter.'"

Seven-Factor Test from Zubulake v. UBS Warburg

1. extent to which the request is specifically tailored to discover relevant information
2. availability of such information from other sources
3. total cost of production, compared to the amount in controversy
4. total cost of production, compared to the resources available to each party
5. relative ability of each party to control costs and its incentive to do so
6. importance of the issues at stake
7. relative benefits to the parties of obtaining the information

(217 F.R.D. at 322)
Zubulake's Five Categories of ESI (Most to Least Accessible)*

- Active, online data
- Near-line data
- Offline storage
- Backup tapes
- Erased, fragmented or damaged data

Magistrate Judge John Facciola:

"...I am anything but certain that I should permit a party who has failed to preserve accessible information without cause to then complain about the inaccessibility of the only electronically stored information that remains"

Rights to Control Information

- Most frequently discussed in library lit is copyright
- Claims can extend far beyond intellectual property rights, as defined by law
- Cultural property, replevin and repatriation
- Right to privacy
- Protection of human subjects in research
- Privileged or protected information (e.g. client-attorney, healthcare, social services, library circulation, source – journalist)
- Right to publicity – individual's protection from unauthorized commercial use of her name, persona, or likeness
- Prevention of misappropriation (including plagiarism)
"If a forensic examiner has complete confidence in his/her conclusions, this is usually an indication that he/she is missing something – there is always uncertainty and all assertions should be qualified accordingly..."

"Investigators cannot, in general, directly observe digital data and instead they can only observe the data displayed on a monitor or other output device, which is driven by various types of hardware and software. Because the observation of the data is indirect, a hypothesis must be formulated that the actual data is equal to the observed data. Testing this hypothesis requires that the hardware and software being used are accurate and reliable. Hypotheses also need to be formulated about the data abstractions that exist and the previous states and events that occurred."

Examples of Potentially Useful Inferences (that could be wrong)

- Name embedded in a MS Word file is the document’s author
- Given IP address identifies an individual
- Presence of email addresses on different hard drives indicate correspondence patterns between individuals
- Many common MD5 values across storage locations indicate sharing of files across those locations (context-based filtering can help to address this)
- Last modified date indicates when a document was finalized
- Parts of a page available through the WayBack Machine for a given date represent the parts of the page as available on that date
"Even if a document can be traced to a particular computer and/or IP address, how can we identify who was actually at the keyboard composing the document? It is a particular problem in environments where multiple users may have access to the same computer or when users do not have to authenticate themselves to access a particular account."

Shared Computer Use in the Home*

Ethics Questions To Consider

1. When acquiring a disk as part of a collection, should you create a bit-level image of the disk, in order to ensure the potential to recreate not only the “payload” data of files but also various forms of information within and below the filesystem?

2. Should you retain “hidden” data in a Word document or only retain what you assume to be the text that the author intended?

3. You’re responsible for managing a Microsoft Outlook .pst file over time (including saved and sent messages, calendar items, draft and deleted messages, address book, and viruses). Should you retain the whole .pst file or extract messages and attachments that were sent and received?

4. If a collection documents the life of an individual, how would you determine the appropriate scope for collecting information associated with that person’s online presence (e.g. postings, affiliations, profiles, micro-contributions)?

5. If your institution routinely “normalizes” submitted files into designated file formats, are you obligated to ensure that the normalization doesn’t violate the intentions of creator or other interested stakeholders? If so, what does this obligate you to do specifically?

6. Someone cropped a set of images in order to remove sensitive parts, but the images still have pixel information and embedded thumbnail reflecting the “removed” parts. How should you approach the management of the images?
Lessons and Insights

- Digital forensics has arrived for archival processing
- Introduction of digital forensics doesn’t dictate specific policies or practices
- The disk image is the cornerstone of many forensics methods
- “Taking bitstreams seriously” can have major advantages
- Disk images afford new access scenarios
To Learn More About Available Software

Forensics Wiki. [https://forensicswiki.xyz/page/Main_Page](https://forensicswiki.xyz/page/Main_Page)

BitCurator Environment. [https://bitcurator.net](https://bitcurator.net)

BitCurator Software Overview. [https://bitcurator.github.io/](https://bitcurator.github.io/)

Community Owned digital Preservation Tool Registry (COPTR) [https://coptr.digipres.org/Main_Page](https://coptr.digipres.org/Main_Page)


Lifecycle Tools for Archival Email Stewardship. [https://docs.google.com/spreadsheets/d/1V1N22xnr5e0EbDlZWx58bjYO6rkrMrYH9wGX9-CK8c4/](https://docs.google.com/spreadsheets/d/1V1N22xnr5e0EbDlZWx58bjYO6rkrMrYH9wGX9-CK8c4/)

Online Forums

BitCurator User Group
https://groups.google.com/forum/#!forum/bitcurator-users

Digital Curation List
https://groups.google.com/forum/#!forum/digital-curation
Further Education

Training Overview

AccessData Training is conducted by Synticate, a trusted training partner championed by a team of expert trainers. Through Synticate, AccessData is able to provide Computer Forensic, Mobile Forensic, Legal and eDiscovery Training. Choose from individual training courses, annual training passes, on-demand training videos, or Custom Training to suit your team’s specific scenario and goals.

COMPUTER FORENSICS

Synticate will educate you in technology and prepare you with innovative ideas and workflows to improve and strengthen your skills to identify, respond, investigate, prosecute, and adjudicate cases. The Synticate computer forensic team focuses on how to properly collect, process, review and report case data toward successful case resolution. View Synticate’s Computer Forensic courses and syllabi, course calendar, on-demand training videos, annual training passes, custom training and certifications. Or, simply contact Synticate and we’ll set up a call to discuss our offerings and which type of training and delivery method suits your specific needs.

Courses

<table>
<thead>
<tr>
<th>Level</th>
<th>Title</th>
<th>Delivery Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introductory</td>
<td>EnCase OnDemand Forensics and Enterprise Fundamentals Training</td>
<td>Online</td>
</tr>
<tr>
<td>Introductory</td>
<td>EnCase OnDemand Forensics and Enterprise Fundamentals Training</td>
<td>Online</td>
</tr>
<tr>
<td>Introductory</td>
<td>EnCase OnDemand eDiscovery for the Legal Team</td>
<td>Online</td>
</tr>
<tr>
<td>Introductory</td>
<td>Digital Media – Acquisition and Triage</td>
<td>Online</td>
</tr>
<tr>
<td>Introductory</td>
<td>EnCase of Computer Forensics I</td>
<td>Online</td>
</tr>
<tr>
<td>Introductory</td>
<td>EnCase of Computer Forensics II</td>
<td>Online</td>
</tr>
</tbody>
</table>

DIGITAL INTELLIGENCE

17465 W. Greendale Drive
New Berlin, WI 53151
866-DIGITAL (866-344-6453)
Outside the U.S: 462-783-3332

Site Contents Copyright © 2015
www.DigitalIntelligence.com

REQUEST A QUOTE

The Digital Archivist Boot Camp is a 3 day course designed to provide the attendee with an overview of digital forensic workflows and equipment, as well as how to utilize them for forensic image and duplication/archiving. The foundation of how to build a case within forensic tools will also be discussed and covered with practical exercises.

The course will be attended by new digital forensic archivists, new forensic eDiscovery practitioners or first responders tasked with preserving various types of digital media and tasked with basic data recovery, organization of data or password cracking.

Digital Archivist Boot Camp
$1795.00
REGISTER / ORDER
Thank you!

Go forth and curate the bits!
This resource was released by the BitCuratorEdu project and is licensed under a Creative Commons Attribution 4.0 International License.

Most resources from the BitCuratorEdu project are intentionally left with basic formatting and without project branding. We encourage educators, practitioners, and students to adapt these materials as much as needed and share them widely.

The BitCuratorEdu project is a three-year effort funded by the Institute of Museum and Library Services (IMLS) to study and advance the adoption of digital forensics tools and methods in libraries and archives through professional education efforts. This project is a partnership between Educopia Institute and the School of Information and Library Science at the University of North Carolina at Chapel Hill, along with the Council of State Archivists (CoSA) and several Masters-level programs in library and information science.