Digital Forensics for Libraries and Archives:
Introduction to Using BitCurator

Code4Lib 2024 | Ann Arbor, MI

Instructors: Jesse A. Johnston, Elena Colón-Marrero, and Max Eckard

May 16, 2024

Agenda

9:00 - Part I: Introductions & Background: Digital Forensics for Archives

9:50 - break

10:00 - Part II: Imaging

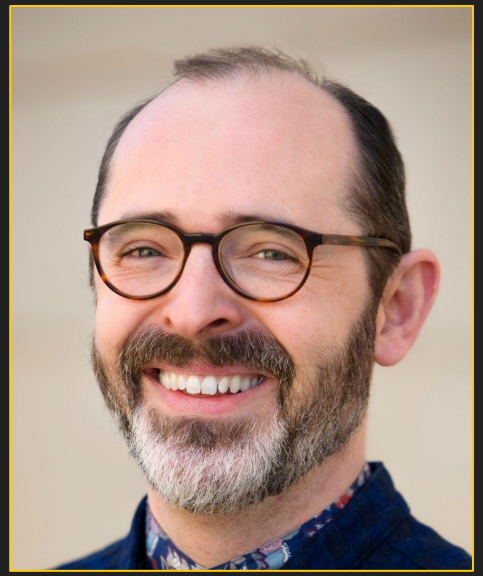
10:50 - break

11:00 - Part III: Reporting

A Bit about the Instructors

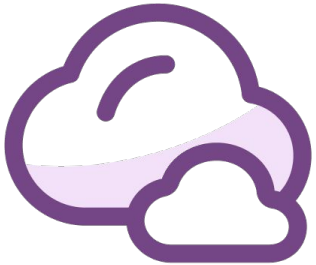
Elena Colón-Marrero,
Bentley Historical Library

Max Eckard,
Bentley Historical Library



Jesse Johnston,
U-M School of Information

slido



Icebreaker: in a word or phrase, what makes you most interested in BitCurator?

Introductions

- What was your icebreaker word/phrase?
- Name and preferred pronouns
- What brings you to the workshop?

Digital Forensics for Cultural Heritage

Much of the content in these slides about digital forensics is provided by the BitCurator.edu project, particular thanks to Cal Lee, Kam Woods, & Jess Farrell

Consider this 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?

Many information professionals know how to process this stuff:

Source: The Processing Table: Reflections on a manuscripts internship at the Lilly Library. https://processingtable.wordpress.com/tag/archival-processing/

How about processing this stuff?

Source: “Digital Forensics and creation of a narrative.” *Da Blog: ULCC Digital Archives Blog*. http://dablog.ulcc.ac.uk/2011/07/04/forensics/

Same Goals as When Acquiring Analog Materials

- Ensure integrity of materials
- Allow users to make sense of materials and understand their context
- Prevent inadvertent disclosure of sensitive data

"Machines" (i.e., computers) are pretty good at creating data, so

- 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 **automatically performed by software**

Different media, but similar underlying principles to cultural heritage

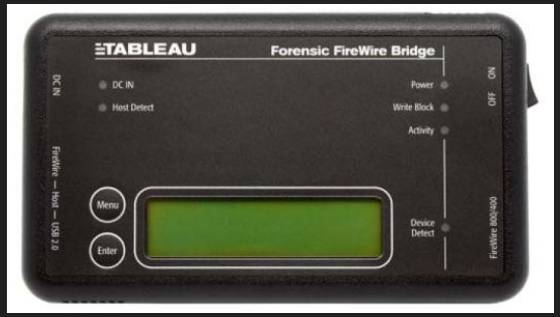
Provenance	• Reflect "life history" of records • Records from a common origin or source should be managed together as an aggregate unit
Original order	• Organize and manage records in ways that reflect their arrangement within the creation/use environment
Chain of custody	• "Succession of offices or persons who have held materials from the moment they were created"[1] • Ideal recordkeeping system would provide "an unblemished line of responsible custody"[2]

1. Pearce-Moses, Richard. A Glossary of Archival and Records Terminology. Chicago, IL: Society of American Archivists, 2005.
2. Hilary Jenkinson, A Manual of Archive Administration: Including the Problems of War Archives and Archive Making (Oxford: Clarendon Press, 1922), 11.

Digital Forensics

- Ensure authentic and trustworthy extraction & transfer of digital information, particularly developed to create evidence in prosecuting “cyber crime”
- Developing in 1980s & 1990s
- Emphases:
  - ensure traceability of evidence (chain of custody),
  - reliability of evidence (documentation),
  - determination of intent/causation (individual-created vs. computer-generated traces),
  - identify system logs and traces (metadata)
- Continually updating to accommodate new technologies & media

In addition, you may need some extra equipment



So many cables!!!
Adapters!!! Dongles 🤦

Why to adopt digital forensics techniques?

- Memory maintainers are often responsible for acquiring or helping others access materials on removable storage media
- Information is often not packaged nor described as one would hope
- Information professionals must extract whatever useful information resides on the medium, while avoiding the accidental alteration of data or metadata

How does digital forensics help in a cultural heritage context?

Provenance	• Identify, extract and save essential information about context of creation
Original order	• Reflect original folder structures, files associations, related applications and user accounts
Chain of custody	• Documentation of how records were acquired and any transformations to them • Use well-established hardware and software mechanisms to ensure that data haven't been changed inadvertently
Identify sensitive information	• Identify personally identifying information, regardless of where it appears • Flag for removal, redaction, closure or restriction

Digital forensics in digital curation, LAM, etc

- 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 information

Representing digital resources

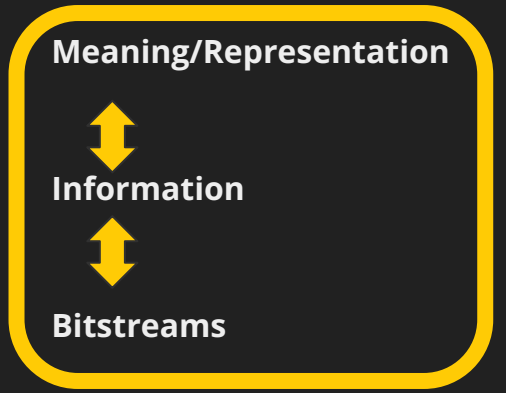


Chart by Cal Lee (via BitCurator.edu project)

Level	Label	Explanation
8	Aggregation of objects	Set of objects that form an aggregation that is meaningful encountered as an entity
7	Object or package	Object composed of multiple files, each of which could also be encountered as individual files
6	In-application rendering	As rendered and encountered within a specific application
5	File through filesystem	Files encountered as discrete set of items with associate paths and file names
4	File as “raw” bitstream	Bitstream encountered as a continuous series of binary values
3	Sub-file data structure	Discrete “chunk” of data that is part of a larger file
2	Bitstream through I/O equipment	Series of 1s and 0s as accessed from the storage media using input/output hardware and software (e.g. controllers, drivers, ports, connectors)
1	Raw signal stream through I/O equipment	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)
0	Bitstream on physical medium	Physical properties of the storage medium that are interpreted as bitstreams at Level 1

Representing digital resources

Levels where digital forensics tools and methods can help

Level	Label	Explanation
8	Aggregation of objects	Set of objects that form an aggregation that is meaningful encountered as an entity
7	Object or package	Object composed of multiple files, each of which could also be encountered as individual files
6	In-application rendering	As rendered and encountered within a specific application
5	File through filesystem	Files encountered as discrete set of items with associate paths and file names
4	File as “raw” bitstream	Bitstream encountered as a continuous series of binary values
3	Sub-file data structure	Discrete “chunk” of data that is part of a larger file
2	Bitstream through I/O equipment	Series of 1s and 0s as accessed from the storage media using input/output hardware and software (e.g. controllers, drivers, ports, connectors)
1	Raw signal stream through I/O equipment	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)
0	Bitstream on physical medium	Physical properties of the storage medium that are interpreted as bitstreams at Level 1

Chart by Cal Lee (via BitCurator.edu project)

Digital curators & digital forensics: concepts & tools

Really, it is kind of figuring out as you go.

This media is too old for it to have much commercial value, so you are essentially re-engineering not only the creation context, but also the technology environment!

Photo: Angela Lansbury & Kasi Lemmons in “The Survivor” episode from *Murder, She Wrote* (1993); via IMDB

Some digital forensics techniques/concepts

Disk image - a bit-perfect sequence of all the bits on a particular physical device; in other words, a complete bitstream (as defined by the physical limits of a storage device).

- You may have seen `.dmg`, or `.iso` files - these are images (like a thumb drive, CD, diskette)
- We will work with "forensic images," specifically the "Expert Witness" format (aka `.E01` or EWF), which is a complete sequence of a physical drive, does not allow any modifications

Bitstreams vs **files** - chunks of bits (sequence of all bits on a drive, vs logically linked sequences for a purpose)

File characterization - identify/confirm the type of file

Mount / Unmount - drives or images are not accessible to the file system until "mounted", and in some cases this can modify file access, or modification dates, other info

Writeblocker - software or hardware device that

Checksums - algorithmically generated value representing a bit sequence (e.g., MD5, SHA family, CRC)

BitCurator Environment

BitCurator project

- Create a born-digital processing environment that supports integration of digital forensics tools into digital curation workflows, particularly cultural heritage collections context
- Led by U North Carolina School of Information & Library Science (SILS), U of Maryland Institute for Technology & Humanities (MITH)
- Cal Lee, Matt Kirschenbaum, Kam Woods, et al
- Funding from Mellon Foundation (multiple grants since 2011)

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, consulting)
- U Michigan is a member

See https://bitcuratorconsortium.org/

BC 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

BC Environment

- Bundles, integrates and extends functionality of open source software
- Can be run in various ways:
  - Self-contained environment running directly on a computer (download installation ISO)
  - Add to any Ubuntu Linux machine by importing BitCurator scripts
  - As individual components run directly in your own Linux environment or (whenever possible) Windows environment
  - **Self-contained Linux environment in a virtual machine using, e.g. VirtualBox**

To read about and download the environment, see
https://github.com/BitCurator/bitcurator-distro

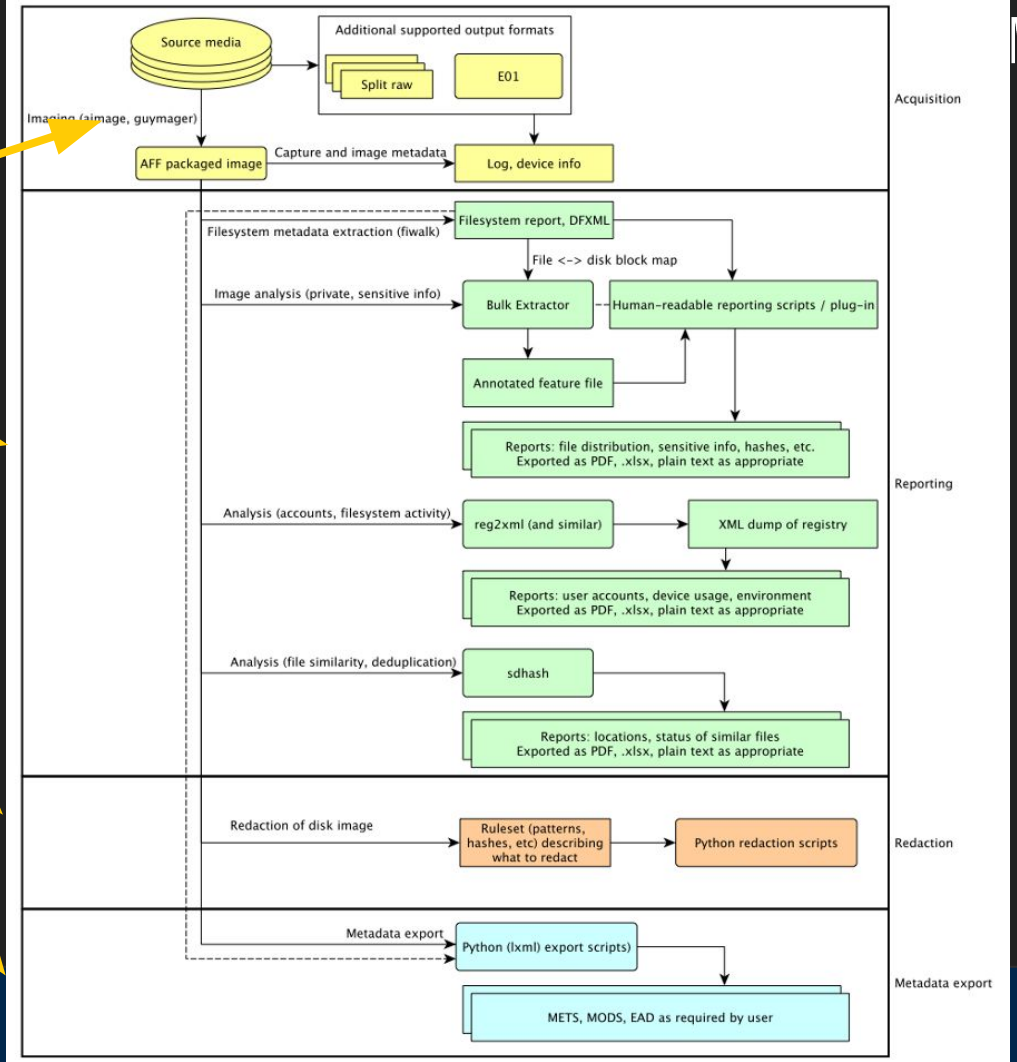
Installation resources

- Oracle VirtualBox (https://www.virtualbox.org/wiki/Downloads)
- The “virtual appliance” file from BitCurator (https://github.com/BitCurator/bitcurator-distro/wiki/Releases) - may take a while to download

Follow the instructions at: https://bitcurator.github.io/documentation/

BC Workflow

- Acquisition
- Reporting
- Redaction
- Metadata export



What sorts of removable media are you working with?

Break 1 - if you have questions during the break, please use the question on the next slide . . . for discussion during the break or right after

slido

During the break: at this point, what are your questions about the BitCurator environment?

ⓘ Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

Part II: Imaging

Part II

- The BitCurator Environment & Tools
- To image or not to image?
- Different types of images
- Creating and mounting images

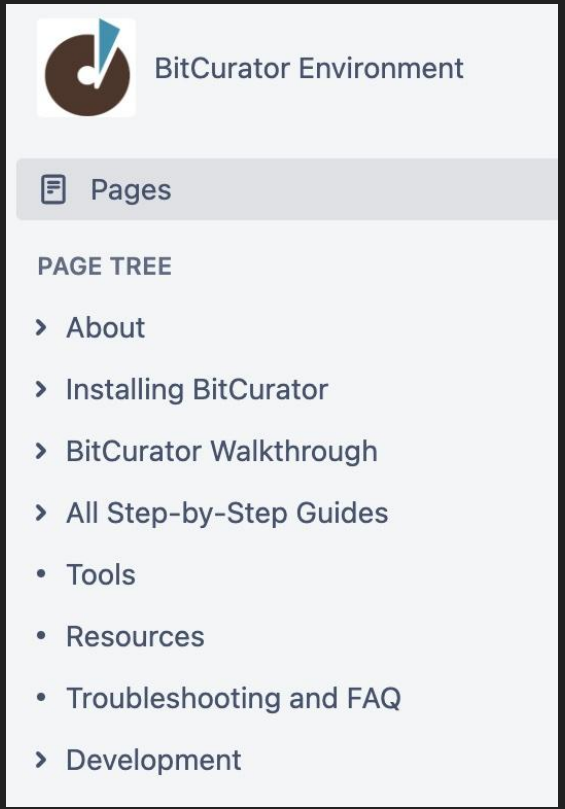
BitCurator tour

Tools/software of note in BitCurator

- Nautilus (like Windows Explorer or Mac Finder)
- “Nautilus scripts” (right click on a file to get the menu)
- Software writeblocker
- Guymager (create EWF disk images)
- Bulk_extractor (identify sensitive information)
- Brunnhilde (characterize & report on file formats)
- fiwalk (create DFXML map)
- BitCurator reporting tools

Resources of Note

- Most of the tools & tasks we will cover are also explained in the BC QuickStart Guide (https://github.com/BitCurator/bitcurator-distro/wiki/Releases#quickstart-guide)
- There is also the BC wiki (https://github.com/BitCurator/bitcurator-distro/wiki)



Imaging media in BC

To image or not to image?

- Choosing between extracting files and/or chunks of content
- Collection considerations:
  - What is your collecting purpose?
  - What is the role of the device(s)?
- Device considerations:
  - What devices have an OS (that means lots of redundant & proprietary files)?
  - If it's a storage device, may have deleted/unintended files (these are captured by forensic imaging approaches)
- DANNNG! Disk Imaging Decision Factors

Creating a disk image = creating an exact copy of the data on a medium

- Getting an “image” of a storage medium involves working at a level below the file system
- Can “see” file attributes and deleted files not visible through higher-level copy operations

Some disk image formats you may see

- 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 format (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) - http://sourceforge.net/projects/libewf/files/

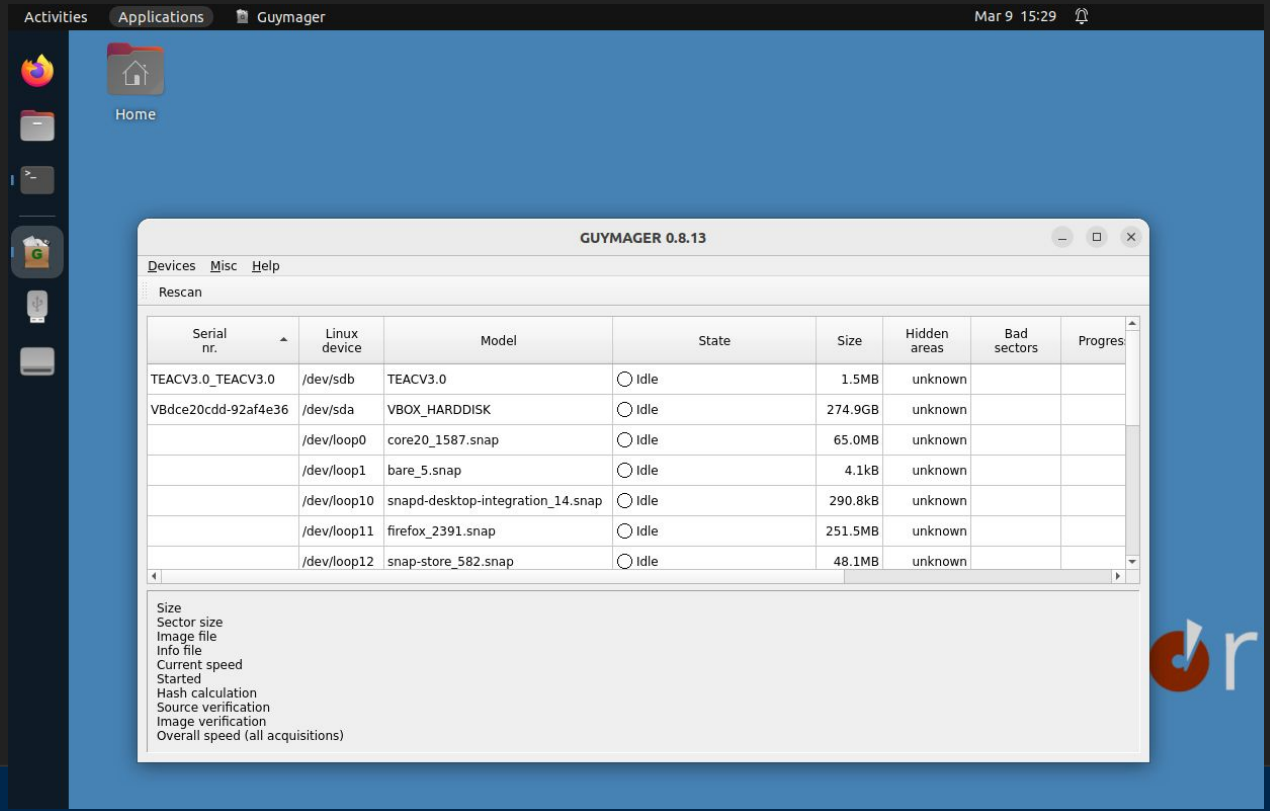
ISO (.img) for CD-ROM, DVD

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

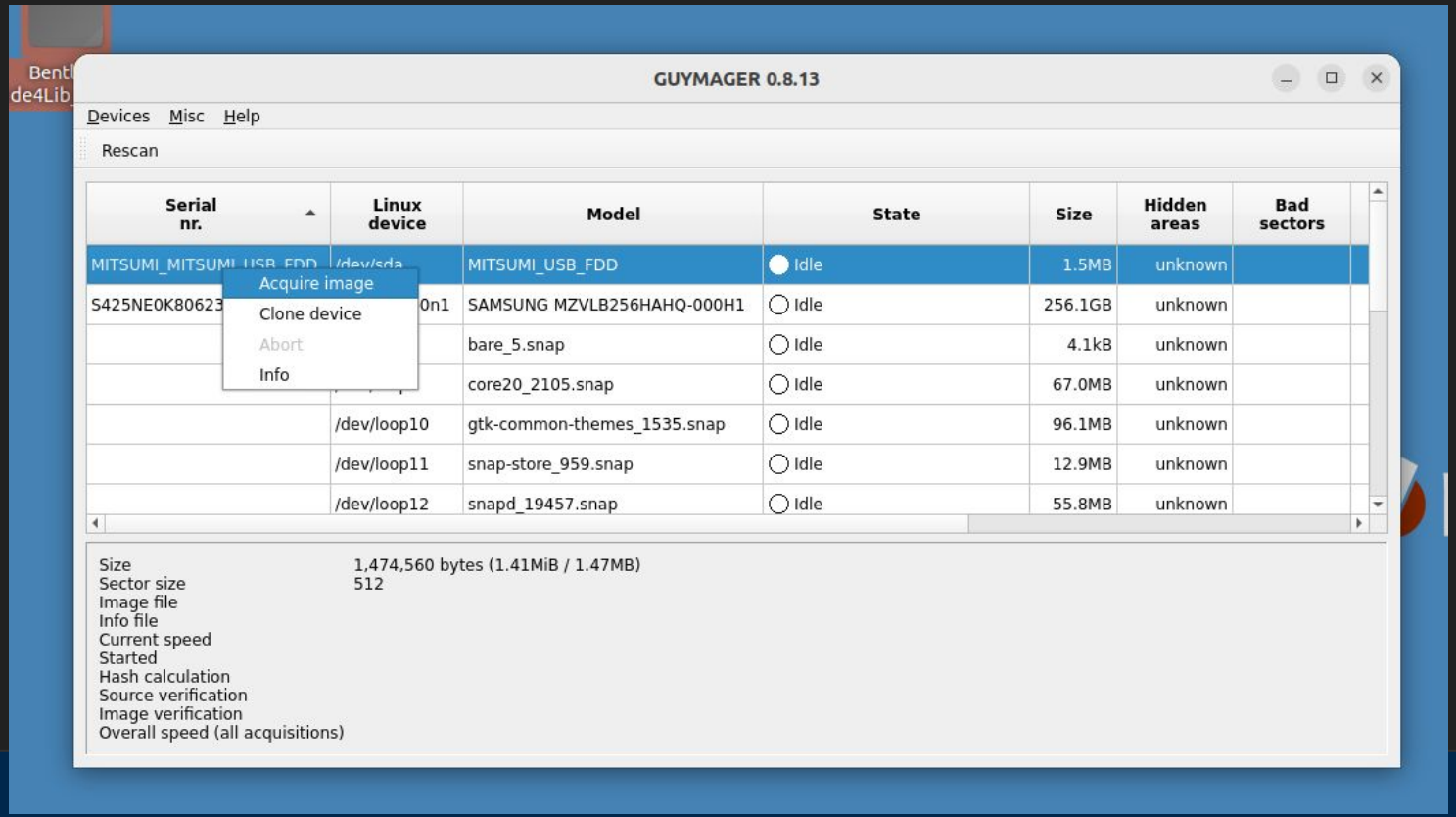
Accessing disk images

- Virtualization and emulation
- Mounting the original filesystem
- Accessing (but not mounting) disk images using forensics software
- For end user access:
  - Remote, dynamic access to disk image contents (via server, virtual environment)
  - Cross-drive analysis

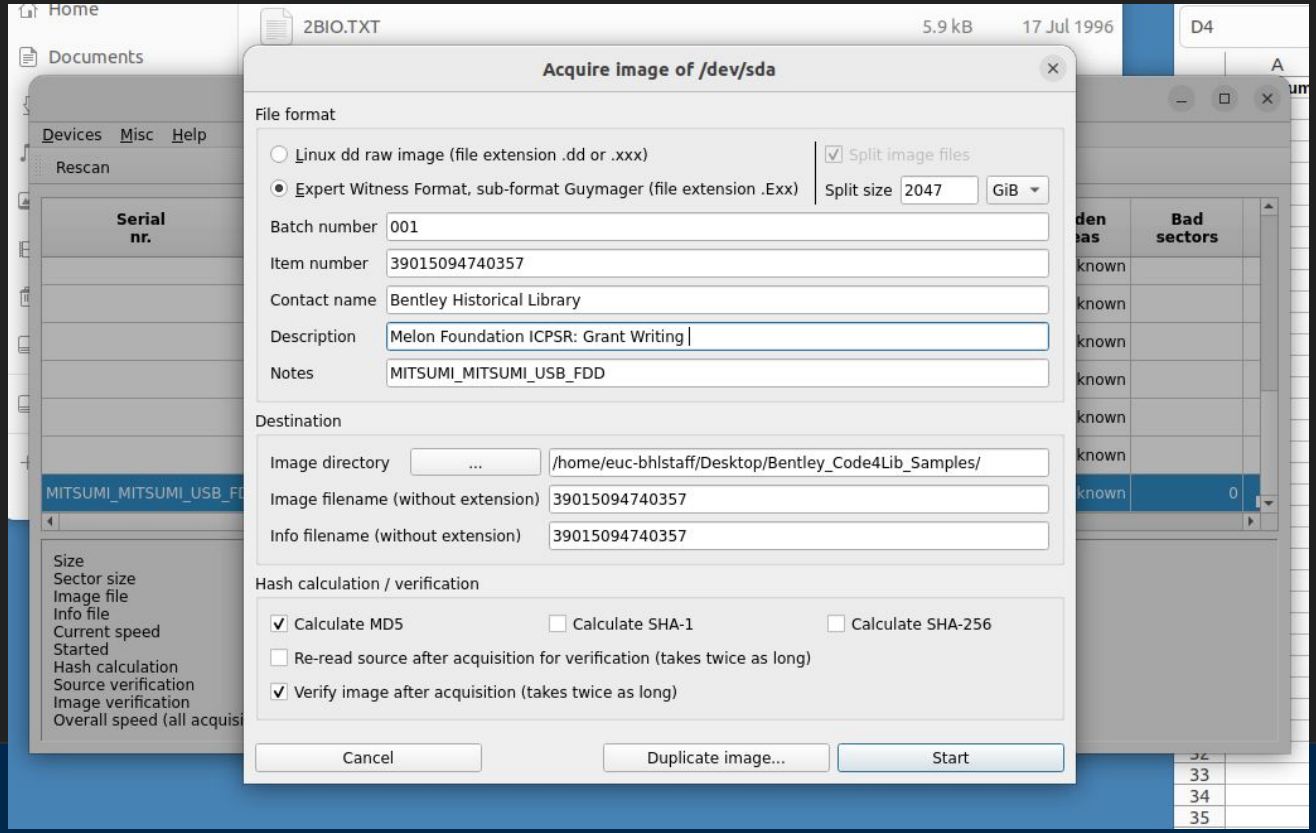
Creating a disk image in BC: Using Guymager



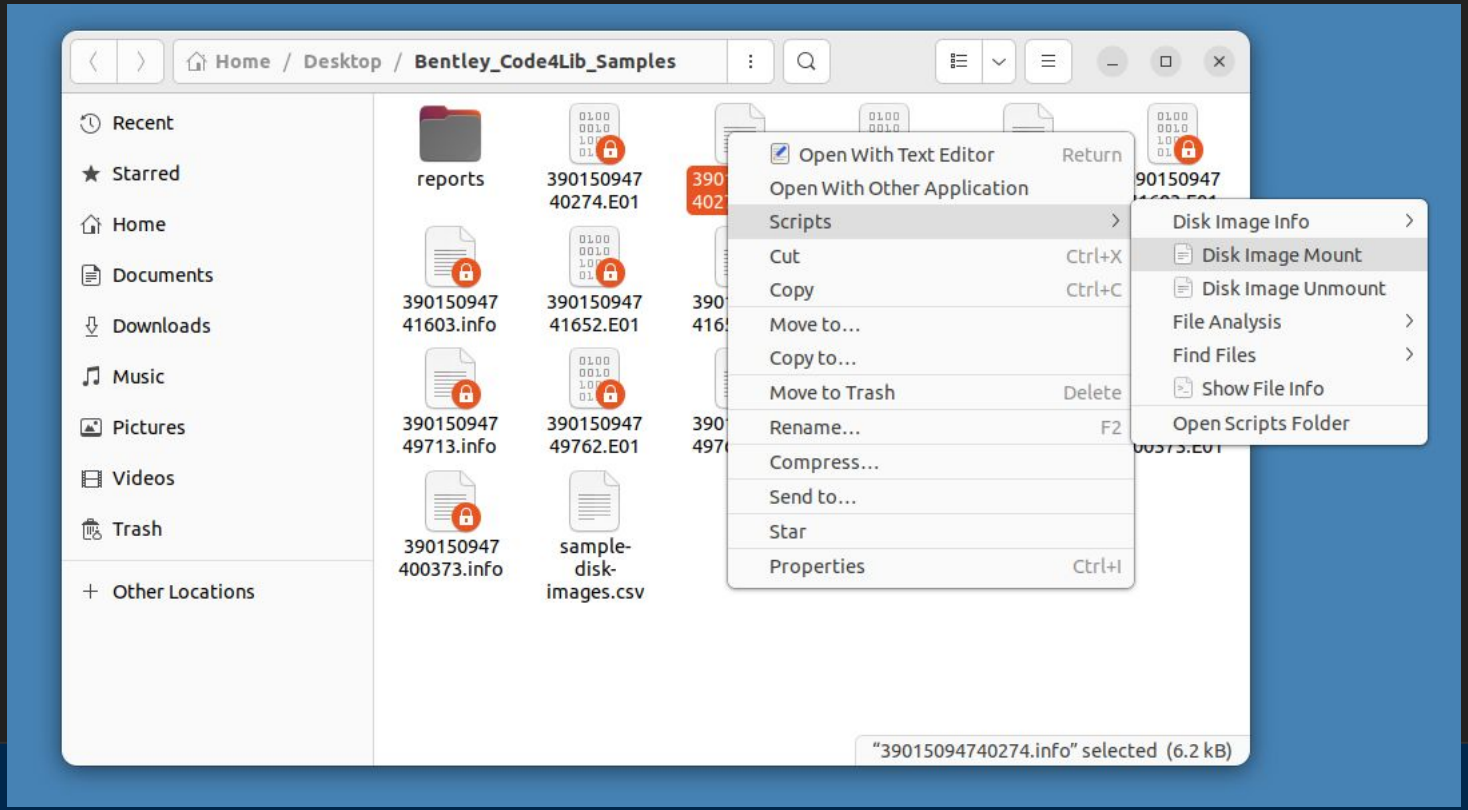
Creating a disk image in BC: Using Guymager



Creating a disk image in BC: Using Guymager

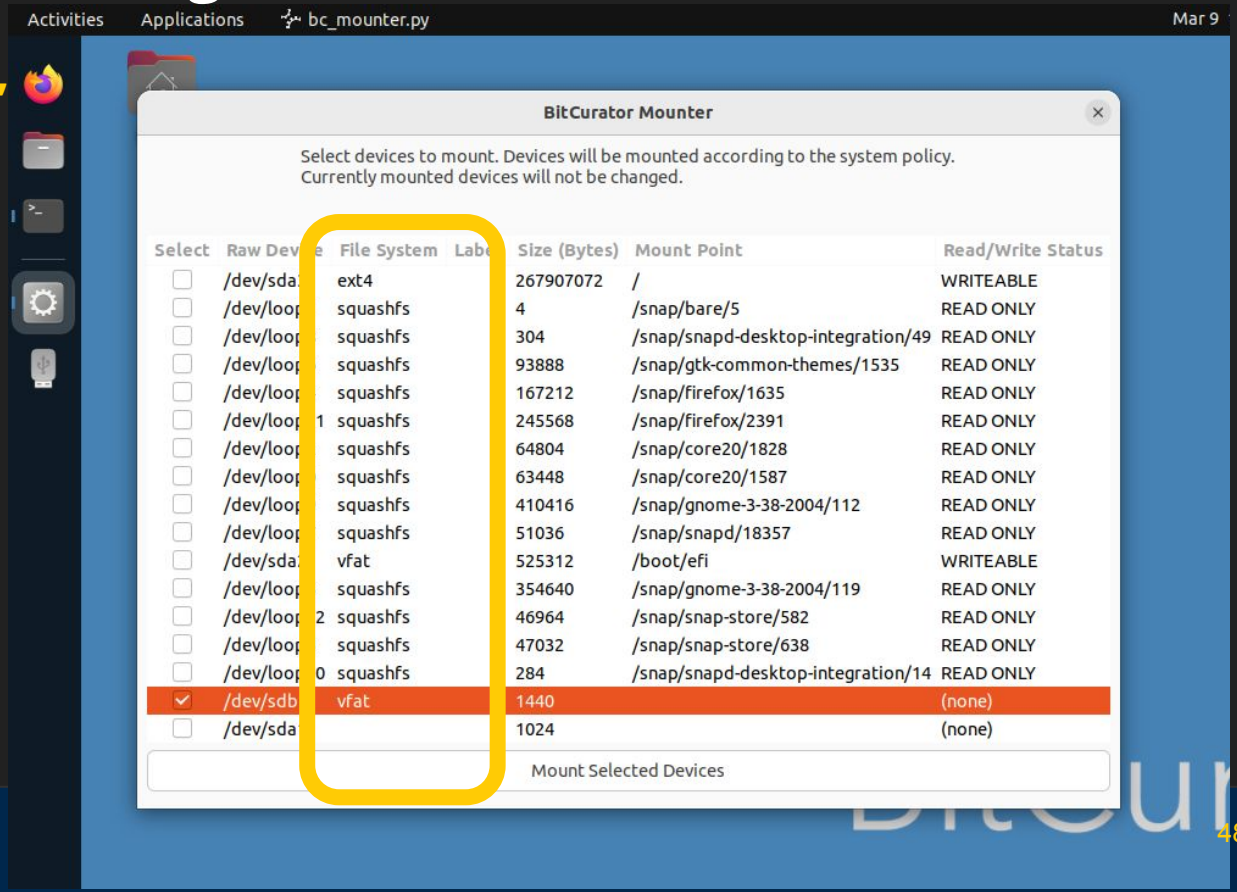


Mount disk image: Using BitCurator Mounter/Scripts



Mount disk image: Using BitCurator Mounter

- **Note: "File System" Column**
- **You will need this information for reporting!**



Demo: mounting & viewing an EWF disk image in BC

- Using a disk image from the workshop files (linked at https://drive.google.com/drive/folders/1UQKnuwDyv8rEe2-5aFAEKkvFgYHBW7Lo)
- Mount the image, and open it in BC
- Observations? What information do you see?

Choose your own adventure section

I. Use one of the already created disc images available in the linked files, and continue looking at these, identifying different sections, or reporting
II. Create your own image, and get ready to create further documentation from the reports (next section)

Key points

- Digital forensic approaches can offer useful tools to digital curators in working with legacy removable media
- Important concepts include thinking beyond the file level and disk imaging
- BitCurator environment offers a useful bundle of tools that are of use to digital curators

Key Points

- Digital forensics developed for extracting & tracking electronic evidence, but has some useful applications for digital curators
- Born-digital content should be moved off removable media into controlled & trackable locations. This may entail extraction of specific files, or creation of disk images.
- You can do this with highly specialized software, but there are also open-source and modular tools that you can adapt in creating workflows

Part III: Reporting

Basically, this is the metadata part!

Reporting in BitCurator = metadata generation

At a high level, you will be using, and creating a workflow piecing together:

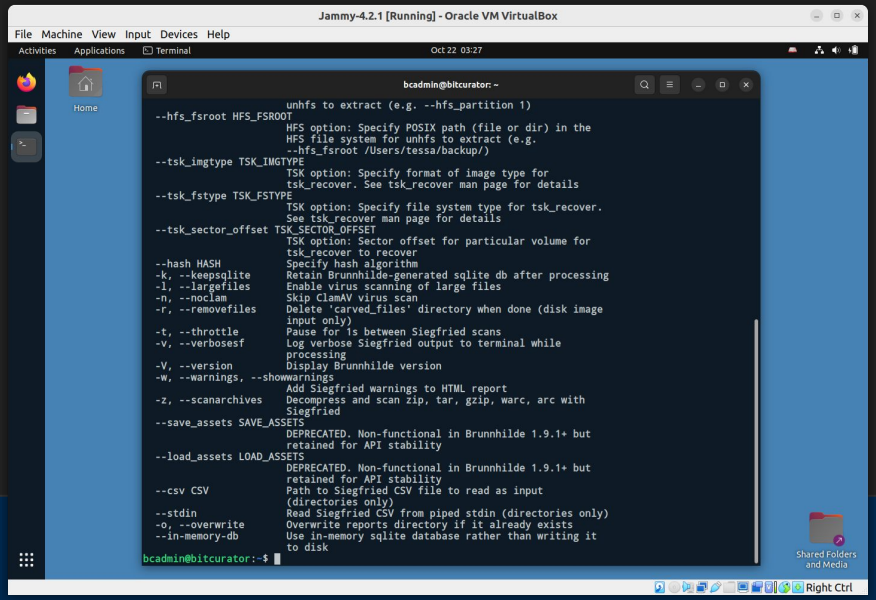
- A “map” of the disk image, which records relationships, integrity (checksums), names, timestamps, etc (this is in DFXML)
- A summary of the file types, duplicates, and other relationship information
- Tools for assessing PII & sensitive content
- Summaries of sensitive content, if discovered

One possible structure to group content & metadata

```
c4l24_bicuratorintro_barcodeID_image0XX/  ←parent directory (sample name)
│
├── reports/      ←subdirectory for detailed metadata (use mkdir)
│   ├── beout/  ←bulk extractor reports (generated by bulk_extractor)
│   ├── brunn_output/  ←brunnhilde reports (generated by brunnhilde.py)
│   └── mappedfeatures/  ←sensitive info (generated by identify_filenames.py)
│
├── c4l24_bicuratorintro_barcodeID_image0XX_dfxml.xml  ←DFXML (E01 "map"
generated by fiwalk)
├── c4l24_bicuratorintro_barcodeID_image0XX.E01  ←disk image (generated by
Guymager)
└── c4l24_bicuratorintro_barcodeID_image0XX.info  ←disk image metadata (from
Guymager)
```

First Things First

A simple way to get usage instructions for any of the following tools is to simply type their names in the terminal and press enter. E.g., `brunnhilde.py`, which is the same as as using `brunnhilde.py -h` or `brunnhilde.py --help`.



Map your image

Your goal is to create a DFXML “map” of the image. This will include: all filesystem data, checksums for integrity, and explain the relationships of elements of the disk image.

Tool: fiwalk

To run: Use fiwalk in the terminal

Command syntax:

```
fiwalk -f -X <output filename_dfxml.xml> <input image file>
```

File summaries & reports

Your goal is to create a summary of file types, duplicates, and any hard to identify files.

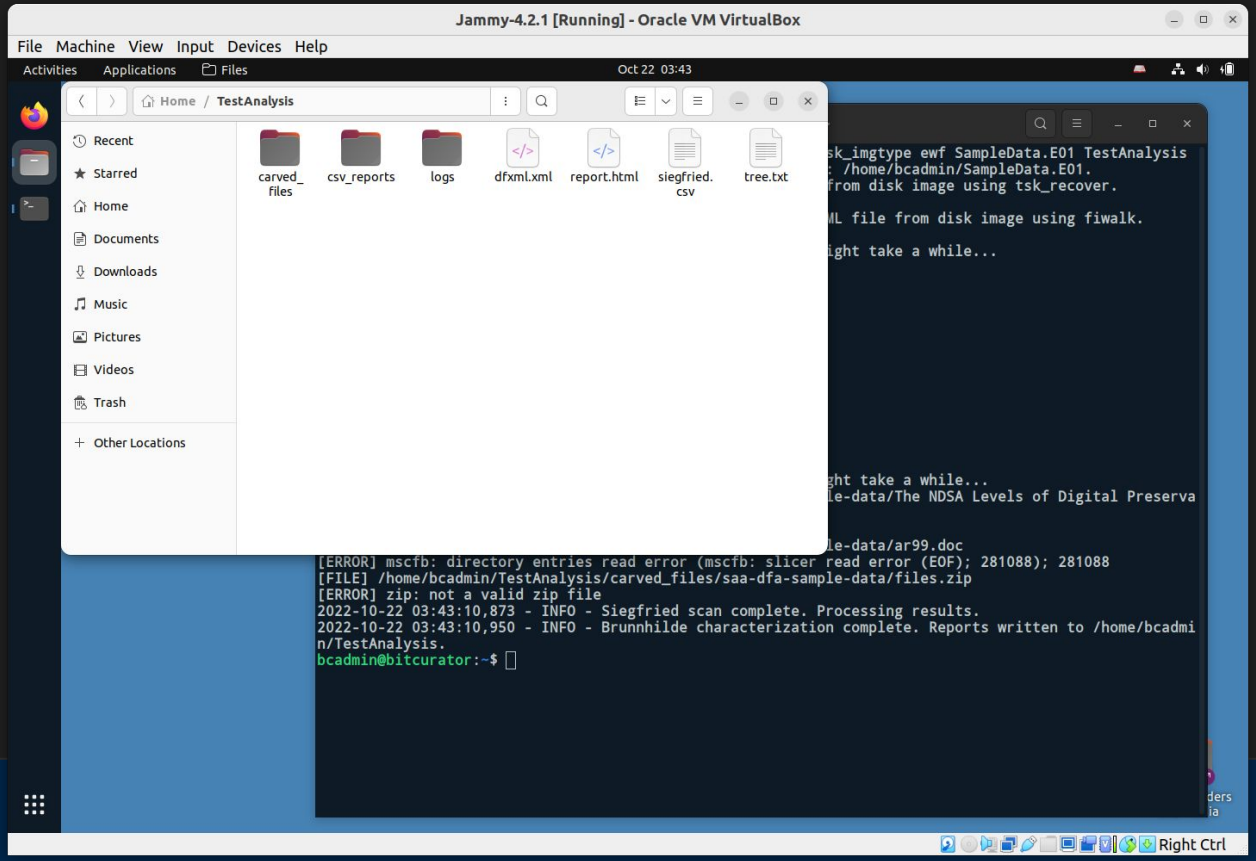
Tool: brunnhilde

To run: Use brunnhilde in the terminal

Command syntax:

```
brunnhilde.py -d -b --tsk_fstype <file system type> --tsk_imgtype
      <image type> <image input file> <output destination>
```

Brunnhilde output



Identify sensitive information

Your goal is to create reports that identify potentially sensitive information, like SSNs, emails, etc.

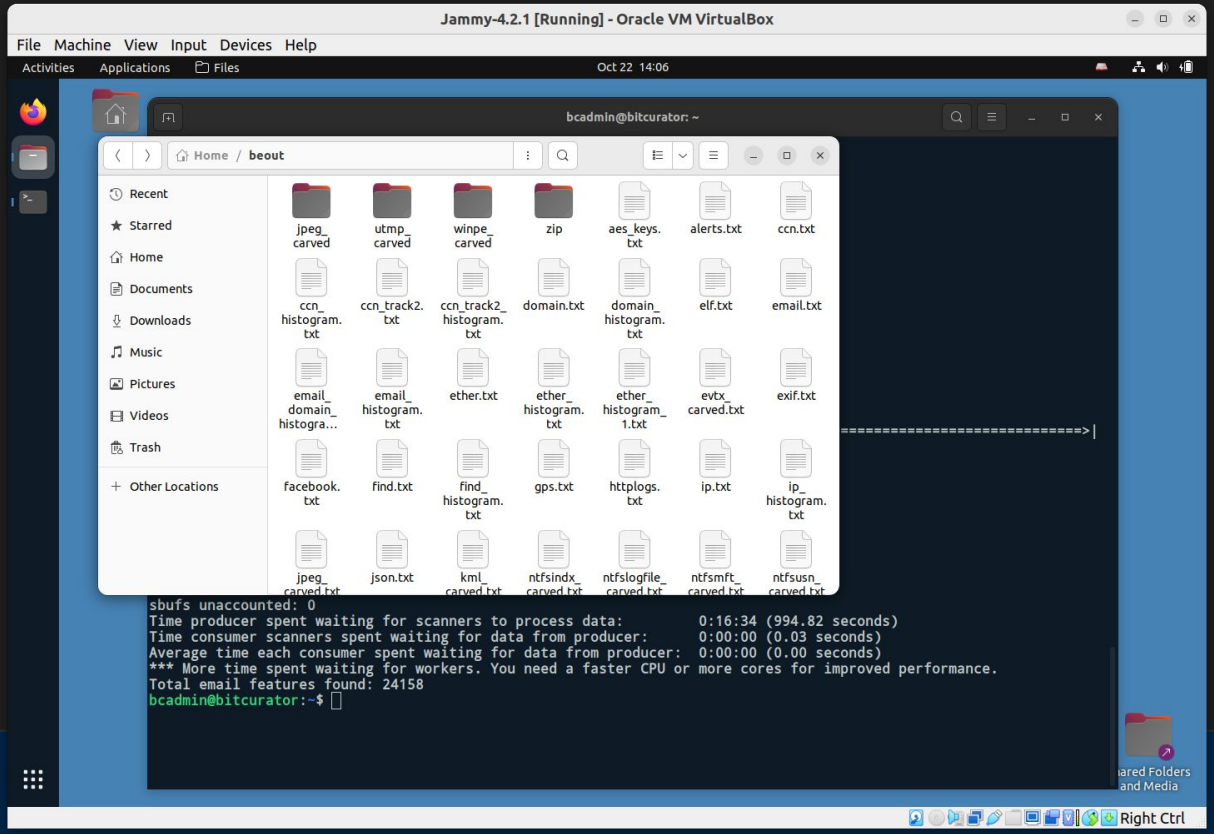
Tool: bulk_extractor

To run: use bulk_extractor in the terminal AND/OR use Bulk Reviewer

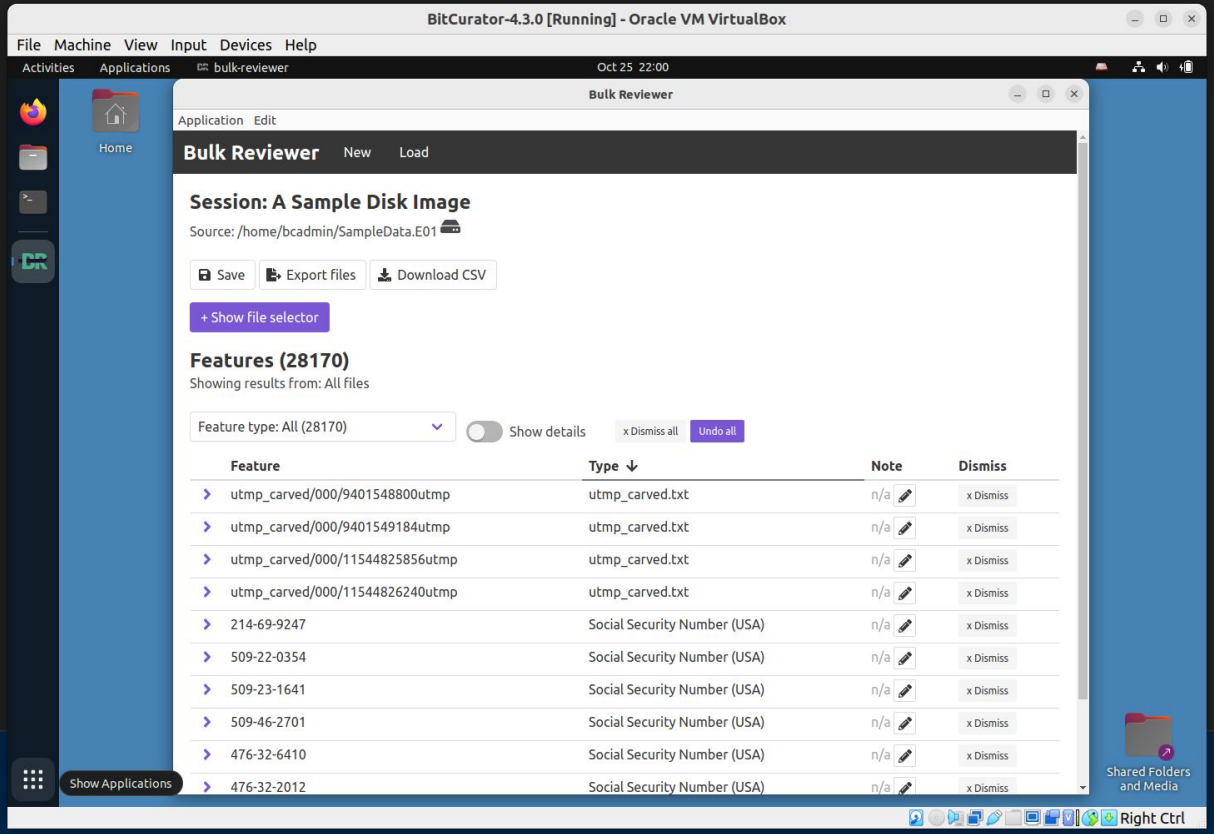
Command syntax:

```
bulk_extractor -o <output destination> <input target disk image file>
```

Brunnhilde output



Identify sensitive information (GUI) - Bulk Reviewer



Summarize sensitive information reports

Your goal is to summarize the reports on sensitive information, show main types of features, and to note what files contain the features.

Tool: identify_filenames.py

To run: use identify_filenames in the terminal

Command syntax:

```
identify_filenames.py --all --image_filename <input disk image> --xmlfile <DFXML of
   the image> <bulk extractor reports location> <destination for summary report>
```

So what?

Reports create technical and preservation metadata about directories or disk images that can accompany them in to the future and aid in later appraisal and processing for preservation and access.

- Some reports may be needed for contextualizing and using the disc images in other programs (dfxml).
- Some reports may be more for risk management and analyzing PII.
- Some may be more for preservation planning (file types).
- Some may be for general description (dates of creation, titles/names of files, users, or other topical information).

Part IV: Open

Time for discussion, more demonstration (different media etc), providing access, etc

Q&A - Discussion

[or use next slide]

slido

Other questions and feedback?

Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

Workflow

Some suggestions for how to implement BC

Documenting workflows

Workflow is “the sequence of processes through which a piece of work passes from initiation to completion”

(*Oxford English Dictionary*, Second Edition, 1989)

As a documenter, **your goal**: represent the process & tasks

Workflow as model

- Explicit, symbolic representation of the workflow
- Usually inspired by new system design or attempts to reengineer a process
- There are many different ways to model a workflow
- But the basic components tend to be similar

Parts of a workflow

- **Entities/Stages** – where something happens (e.g. data are transformed, someone makes a decision, data are captured)
  - Sequence
  - Action
  - Decision points
  - Documentation/ data gathering
- **Input(s)** – control and/or information that flows into an entity/stage
- **Output(s)** – control and/or information that flow out of an entity/stage

Two main goals, related but different, of the representation

- Describe what is **being done** now
  - To understand, analyze, audit current state of things
  - Should be explicitly tied to **how** things are currently done and **who** currently does them
- Describe what you want to **get done**
  - To design new systems, reengineer processes
  - Should focus on the purposes and objectives of a process, rather than fixating on how things are currently done and who currently does them

Identifying a process*

- **Name it**
  - Verb-noun – e.g. generate AIP, harvest web site
  - Verb-qualifier-noun – e.g. generate descriptive information, develop preservation strategy
  - Verb-noun-noun – e.g. assign file permissions, verify object integrity
- Ensure there is a clearly intended **result**
  - Test: noun-is-verbed form (e.g. AIP is generated, web site is harvested, object integrity is verified, permissions assigned)

* Sharp, Alec, and Patrick McDermott. *Workflow Modeling: Tools for Process Improvement and Applications Development*. 2nd ed. Boston, MA: Artech House, 2009. p.40

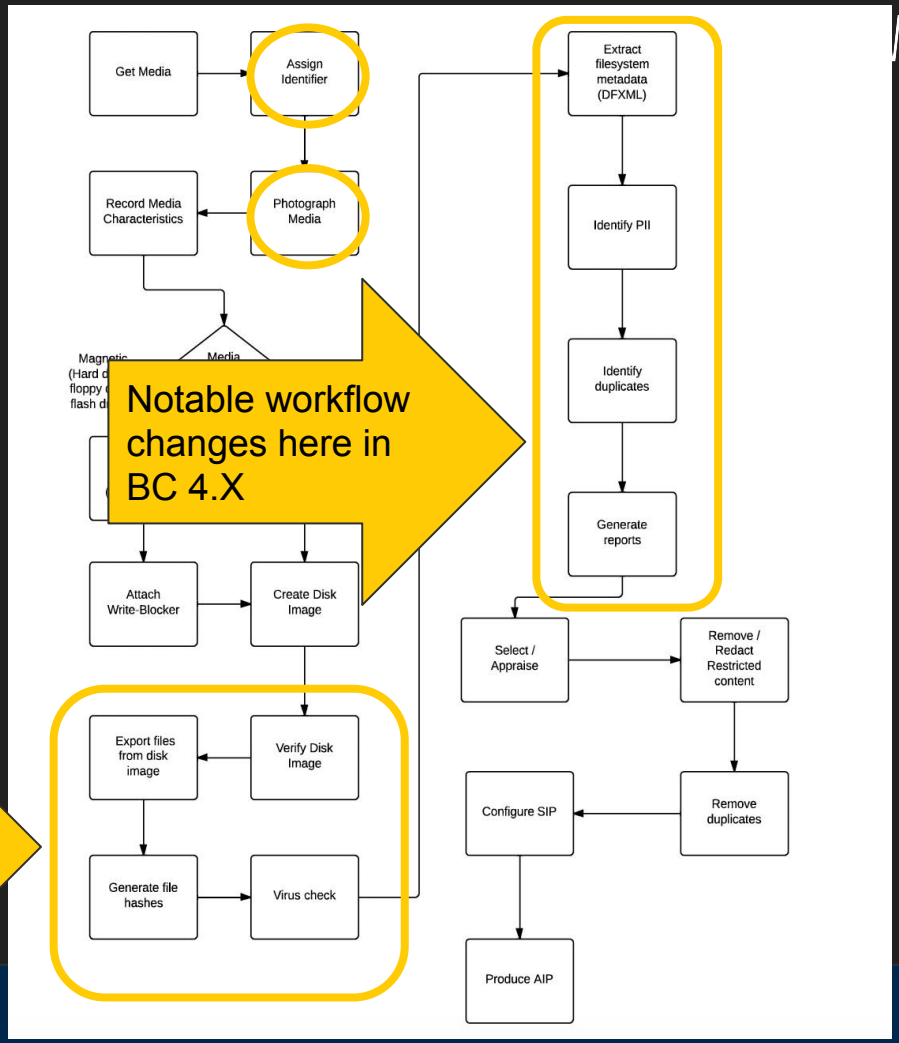
Criteria for Identified Result (Sharp et al)

- **Discrete and identifiable** – “you can differentiate individual instances of the result, and it makes sense to talk about 'one of them’”
- **Countable** – “you can count how many of that result you've produced in an hour, a day, or a week”
- **Essential** – “fundamentally necessary to the operation of the enterprise, not just a consequence of the current implementation,” i.e. “must focus on 'what, not who or how’.”

Example workflow

- How well does this follow the principles from Sharp et al?
- Questions/concerns?
- Observations?

Example workflow from Meister Chassanoff 2014.



Workflow designing

- Gather in your assignment groups - introduce yourselves if you haven't yet
- Get some post-its
- Based on what you know so far, begin identifying the basic steps and processes that you're going to undertake in making your proposal
  - What steps?
  -

Key Points - reporting & workflow mgmt

- BCE offers various tools that help to identify, record, and summarize technical metadata about files on original media
- These include identifying files and filetypes as well as capacities to identify and redact PII

Thank You!