Enabling Digital Forensics Practices in Libraries,
Archives and Museums: The BitCurator Experience

By

Christopher Lee and Kam Woods

Presented At

The Digital Forensic Research Conference

DFRWS 2014 USA Denver, CO (Aug 3rd - 6th)

Enabling Digital Forensics Practices in Libraries, Archives and Museums: The BitCurator Experience

Cal Lee and Kam Woods

School of Information and Library Science

University of North Carolina, Chapel Hill

Digital Forensics Research Workshop

August 3-6, 2014

Denver, CO
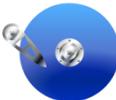
Goals of Libraries, Archives and Museums (LAMs) When Acquiring Materials

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

Fundamental Archival Principles

Provenance
- Reflect "life history" of records
- Records from a common origin or source should be managed together as an aggregate unit

Original Order
Organize and manage records in ways that reflect their arrangement within the creation/use environment

Chain of Custody
- "Succession of offices or persons who have held materials from the moment they were created"[1]
- Ideal recordkeeping system would provide "an unblemished line of responsible custody"[2]

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 Can Help to Fulfill Archival Principles

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

Identifying Sensitive Information
- Identify personally identifying information, regardless of where it appears
- Flag for removal, redaction, closure or restriction

From Bitstreams to Heritage:

Putting Digital Forensics into Practice in Collecting Institutions
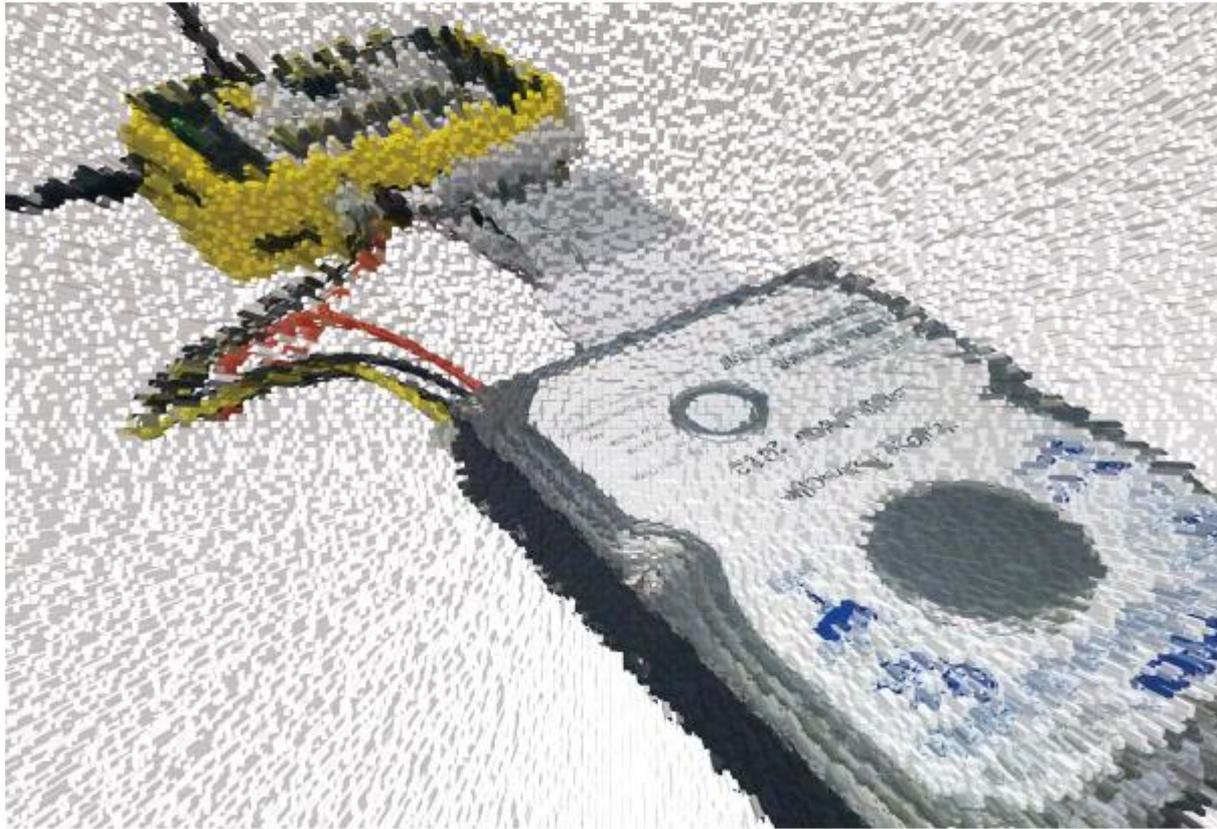
Christopher A. Lee, Kam Woods, Matthew Kirschenbaum, and Alexandra Chassanoff

http://www.bitcurator.net/docs/bitstreams-to-heritage.pdf

Digital Forensics Lab @ UNC School of Information and Library Science

Digital Forensics in LAMs

- In recent years, LAMs have discovered the value of applying various digital forensics methods, for example:
  - use of write blockers
  - generation of disk images
  - applying cryptographic hashes to files
  - capture of Digital Forensics XML (DFXML)
  - scanning bitstreams for personally identifying information

Need for Adaptation of Tools and Tasks for LAM Users

- While existing digital forensics tools provide valuable functionality, they don't always fit well into primary workflows of LAMs.

- For example, LAMs are particularly concerned with:
  - structure and persistence of metadata
  - provisions for providing public access to data
 - support for older technologies (e.g. floppy disks, HFS)

BitCurator

- Funded by Andrew W. Mellon Foundation
  – Phase 1: October 1, 2011 – September 30, 2013
  – Phase 2 – October 1, 2013 – September 30, 2014
- Partners: 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

Core BitCurator Team

- Cal Lee, PI
- Matt Kirschenbaum, Co-PI
- Kam Woods, Technical Lead
- Porter Olsen, Community Lead
- Alex Chassanoff, Project Manager
- Sunitha Misra, Software Developer (UNC)
- Kyle Bickoff, GA (MITH)
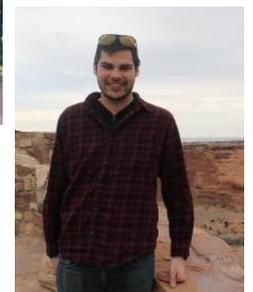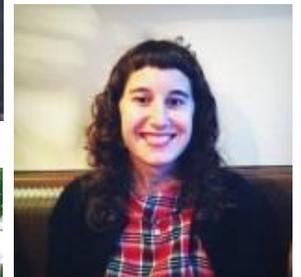
Two Groups of Advisors

| Professional Experts Panel	Development Advisory Group
• Bradley Daigle, University of Virginia Library • Erika Farr, Emory University • Jennie Levine Knies, University of Maryland • Jeremy Leighton John, British Library • Leslie Johnston, US National Archives and Records Administration • Naomi Nelson, Duke University • Erin O'Meara, Gates Archive • Michael Olson, Stanford University Libraries • Gabriela Redwine, Beinecke, Yale University • Susan Thomas, Bodleian Library, University of Oxford	• Barbara Guttman, National Institute of Standards and Technology • Jerome McDonough, University of Illinois • Mark Matienzo, Digital Public Library of America • Courtney Mumma, Artefactual Systems • David Pearson, National Library of Australia • Doug Reside, New York Public Library • Seth Shaw, University Archives, Duke University • William Underwood, Georgia Tech
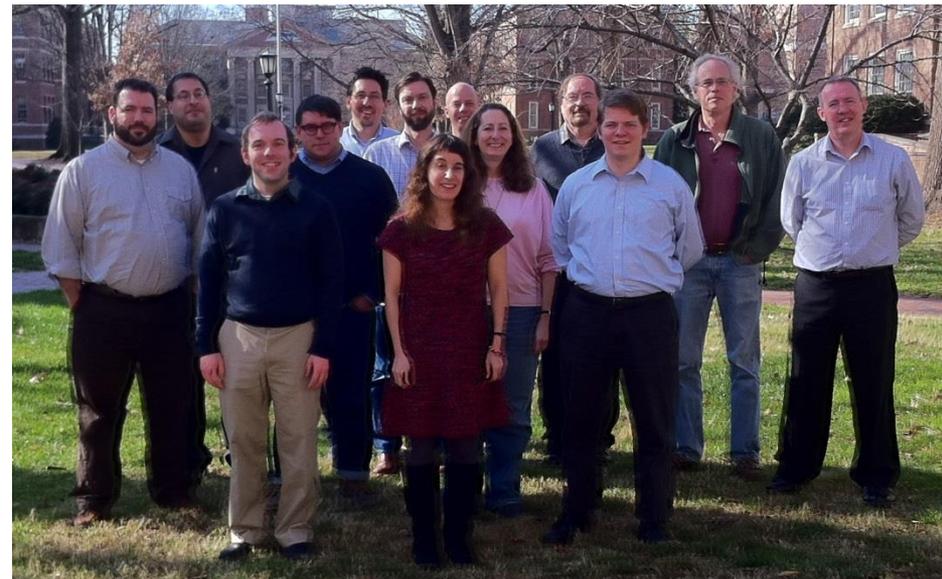
BitCurator Environment*

- Bundles, integrates and extends functionality of open source software: fiwalk, bulk_extractor, Guymager, The Sleuth Kit, sdhash and others
- Can be run as:
  - Self-contained environment (based on Ubuntu Linux) running directly on a computer (download installation ISO)
  - Self-contained Linux environment in a virtual machine using e.g. Virtual Box or VMWare
 - As individual components run directly in your own Linux environment or (whenever possible) Windows environment

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

BitCurator-Supported Workflow Elements

Source media

Additional supported output formats

Split raw   E01

Acquisition

Imaging (aimage, guymager)

AFF packaged image — Capture and image metadata → Log, device info

Filesystem metadata extraction (fiwalk)   Filesystem report, DFXML

File <-> disk block map

Image analysis (private, sensitive info)   Bulk Extractor   Human-readable reporting scripts / plug-in

Annotated feature file

Reports: file distribution, sensitive info, hashes, etc.
Exported as PDF, .xlsx, plain text as appropriate

Reporting

Analysis (accounts, filesystem activity)   reg2xml (and similar)   XML dump of registry

Reports: user accounts, device usage, environment
Exported as PDF, .xlsx, plain text as appropriate

Analysis (file similarity, deduplication)   sdhash

Reports: locations, status of similar files
Exported as PDF, .xlsx, plain text as appropriate

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

Redaction

Metadata export   Python (lxml) export scripts

METS, MODS, EAD as required by user

Metadata export

- **Acquisition**
- **Reporting**
- **Redaction**
- **Metadata Export**

See: http://bitcurator.net

Mounted Devices set to Read-Only by Default*
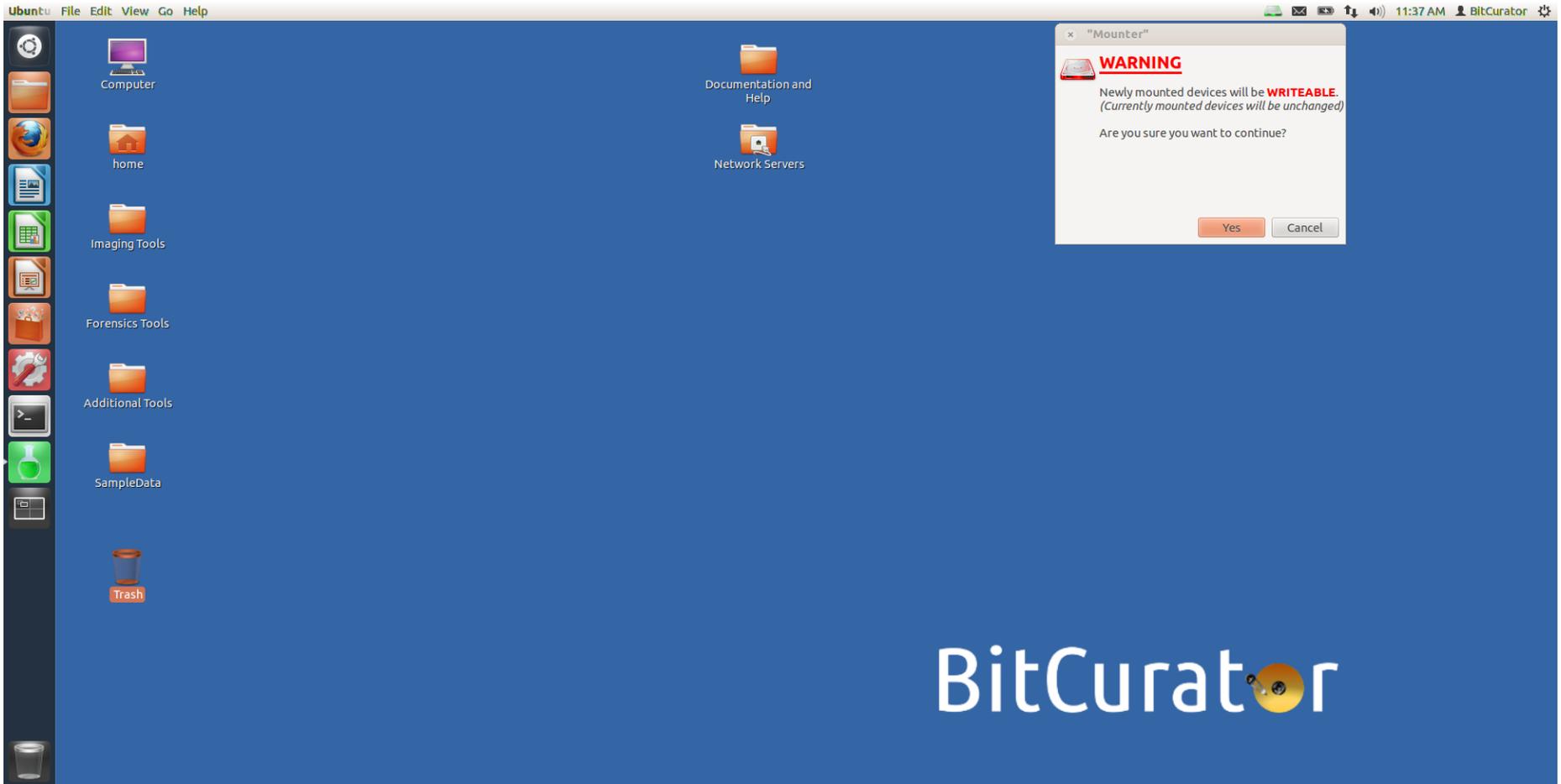
*Not to replace hardware-based write blocking, but useful for various purposes

Creating a Disk Image in Guymager*
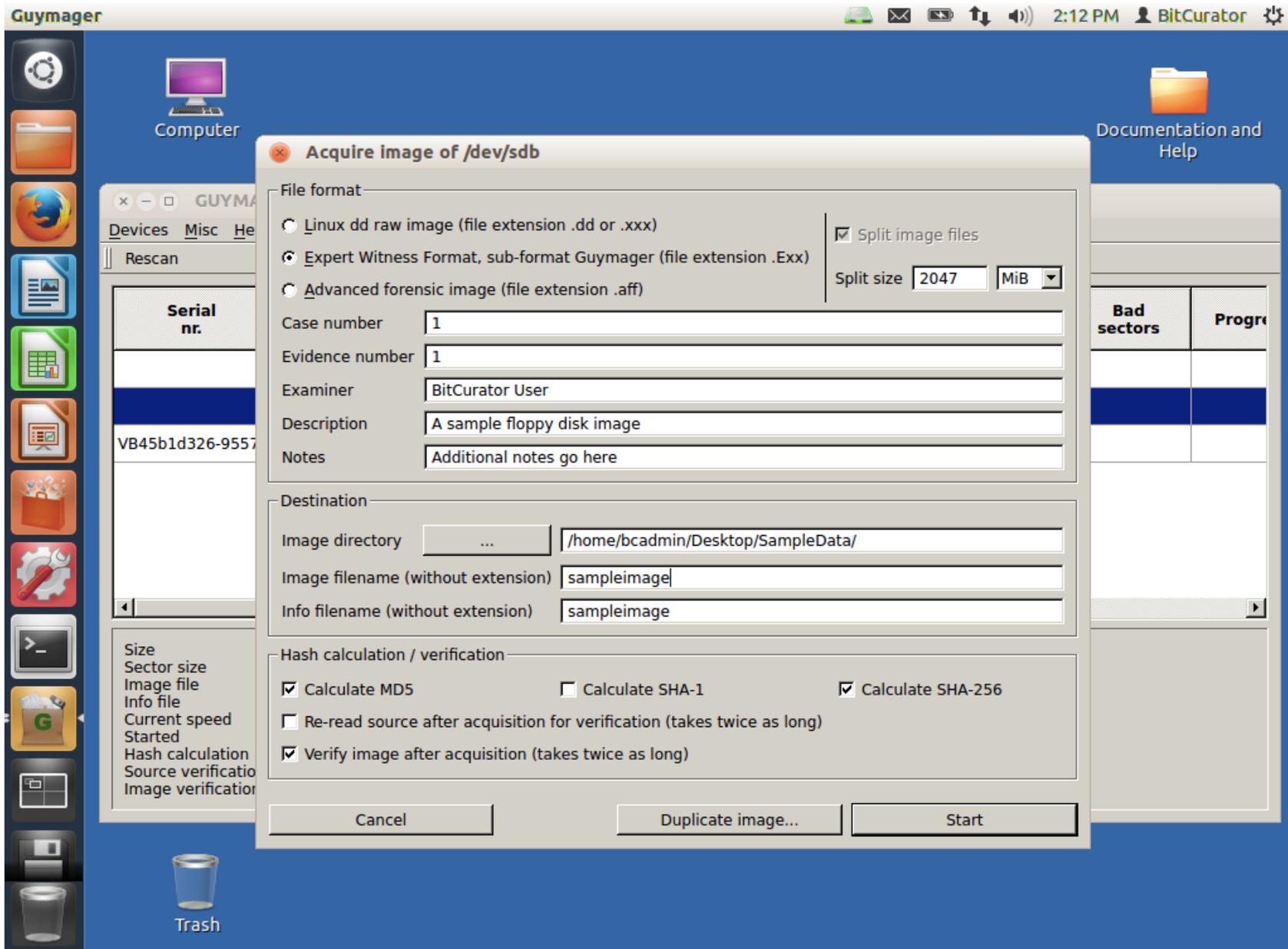
*Developed by Guy Voncken

Mounting a Disk Image to Browse the Contents
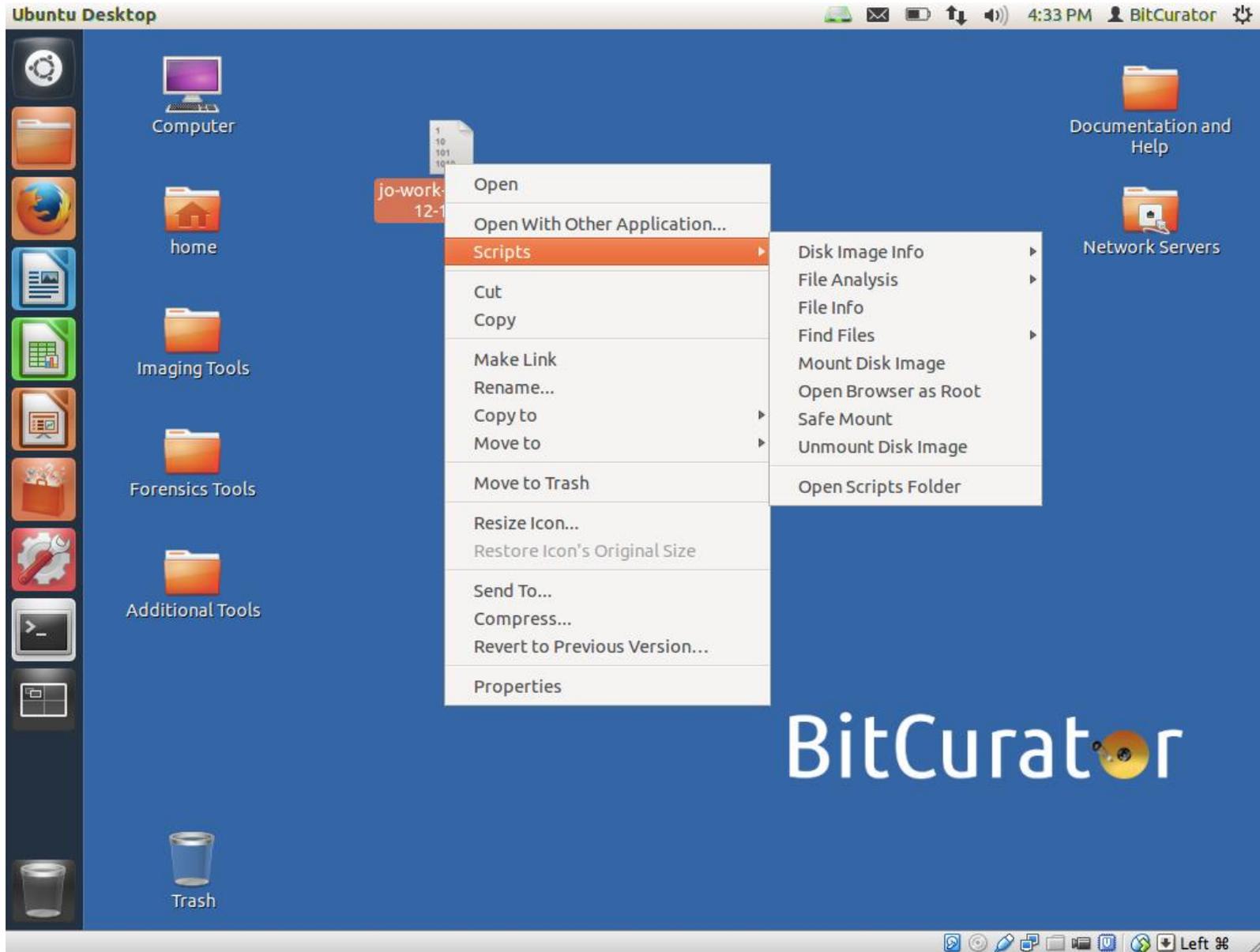
Mounting a Disk Image to Browse the Contents

Bulk Extractor* – Identifying Potentially Sensitive Information

See: http://www.forensicswiki.org/wiki/Bulk_extractor
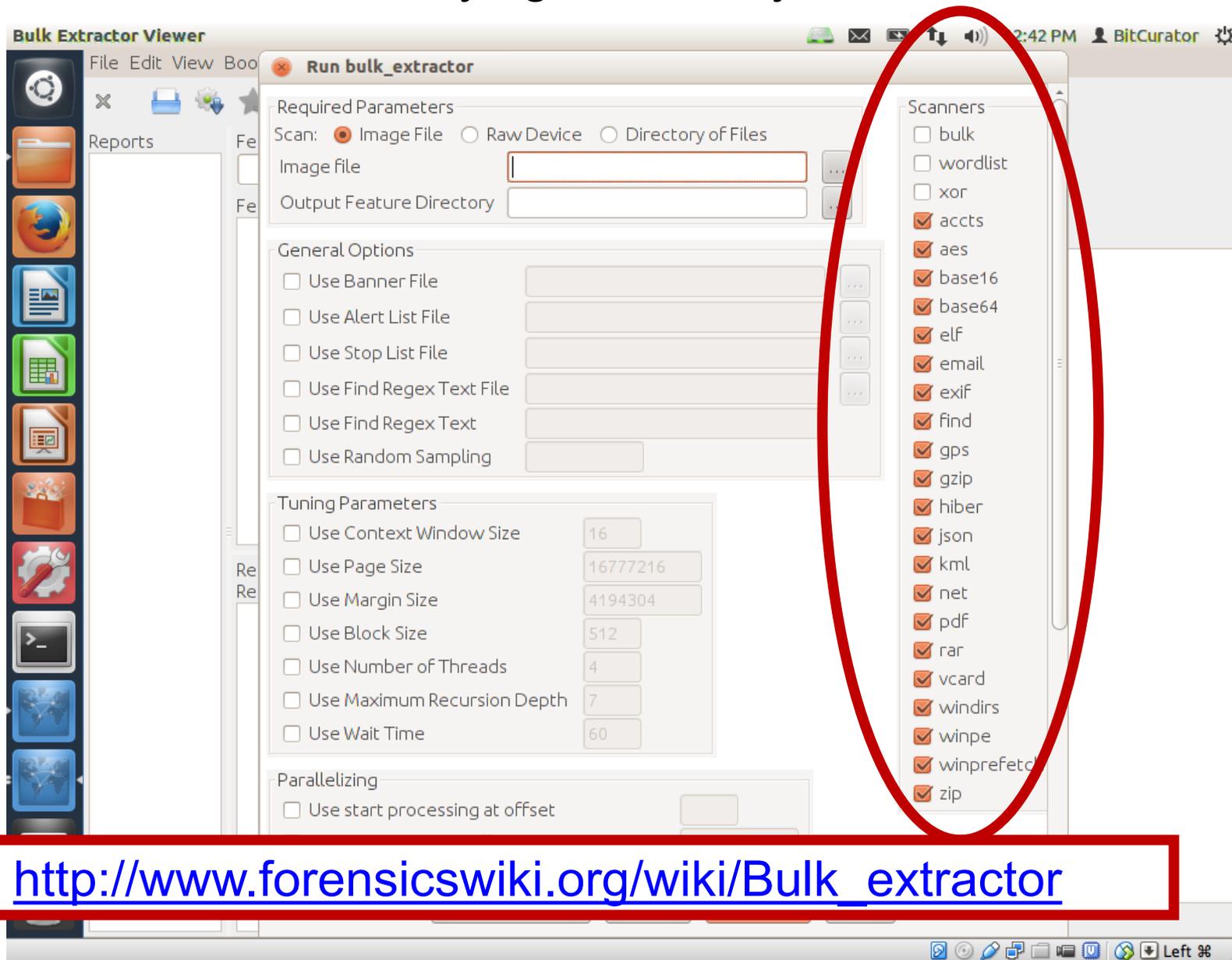
*Developed by Simson Garfinkel

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

BitCurator Reporting Tool

Various Specialized BitCurator Reports

Look Familiar? Filesystem Metadata - Output from fiwalk*

```xml
<fileobject>
     <filename>Documents and Settings/All Users/Documents/
               My Pictures/Sample Pictures/Blue hills.jpg
     </filename>
     ...
     <filesize>28521</filesize>
     <alloc>1</alloc>
     <used>1</used>
     <inode>6245</inode>
     ...
     <uid>0</uid>
     <gid>0</gid>
     <mtime>1208174400</mtime>
     <ctime>1257729636</ctime>
     <atime>1257729636</atime>
     <crtime>1257729636</crtime>
     <seq>2</seq>
     <libmagic>JPEG image data, JFIF standard 1.02</libmagic>
     <byte_runs>
      <run file_offset='0' fs_offset='0' img_offset='363200512'
         len='0'/>
     </byte_runs>
     <hashdigest type='MD5'>
         6fb2a38dc107eacb41cf1656e899cf70
     </hashdigest>
     <hashdigest type='SHA1'>
         4eee44b18576e84de7b163142b537d2fe6231845
     </hashdigest>
</fileobject>
```

*Developed by Simson Garfinkel

Specialized BitCurator Reports

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

Maybe Less Familiar? PREMIS (Preservation) Metadata Generated from Running BitCurator Tools – Recorded as PREMIS Events

Exporting Files from a Disk Image
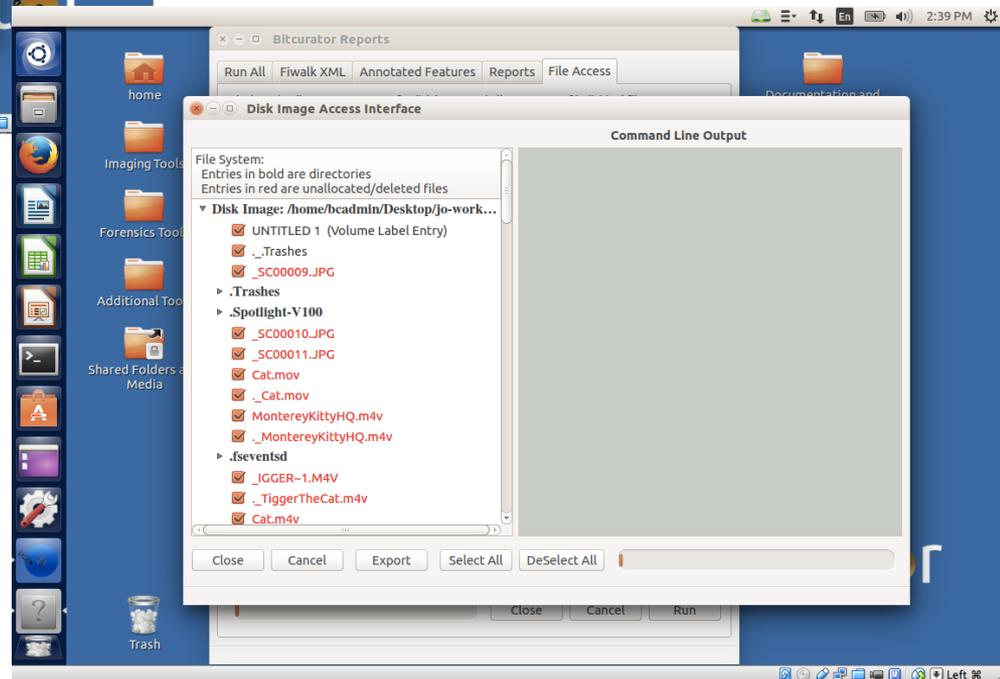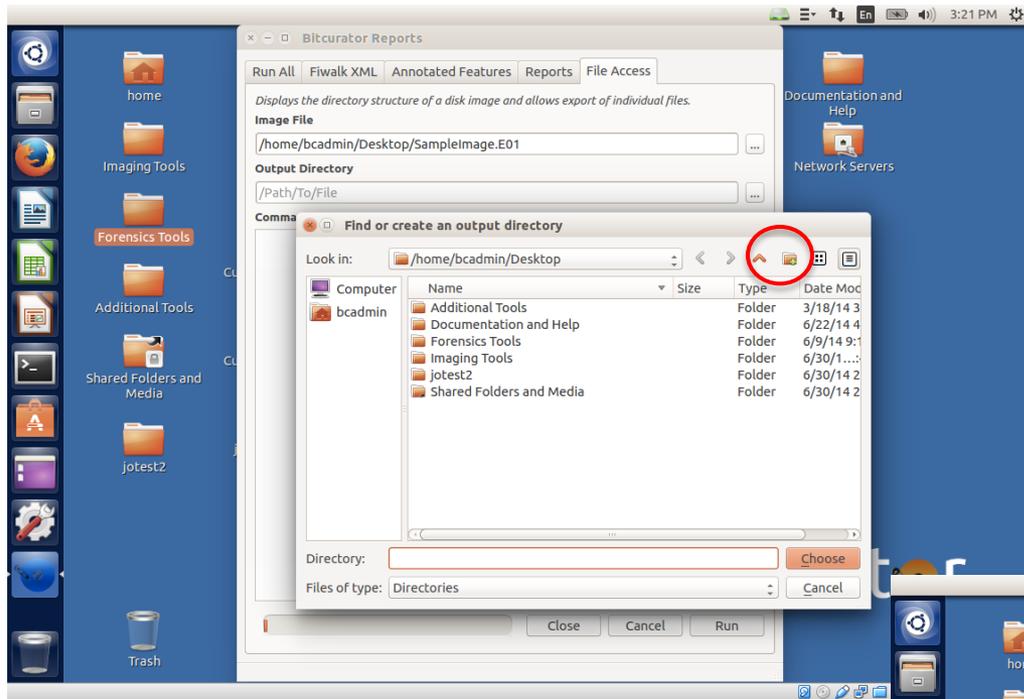
Nautilus Scripts

- 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 Calculate MD5
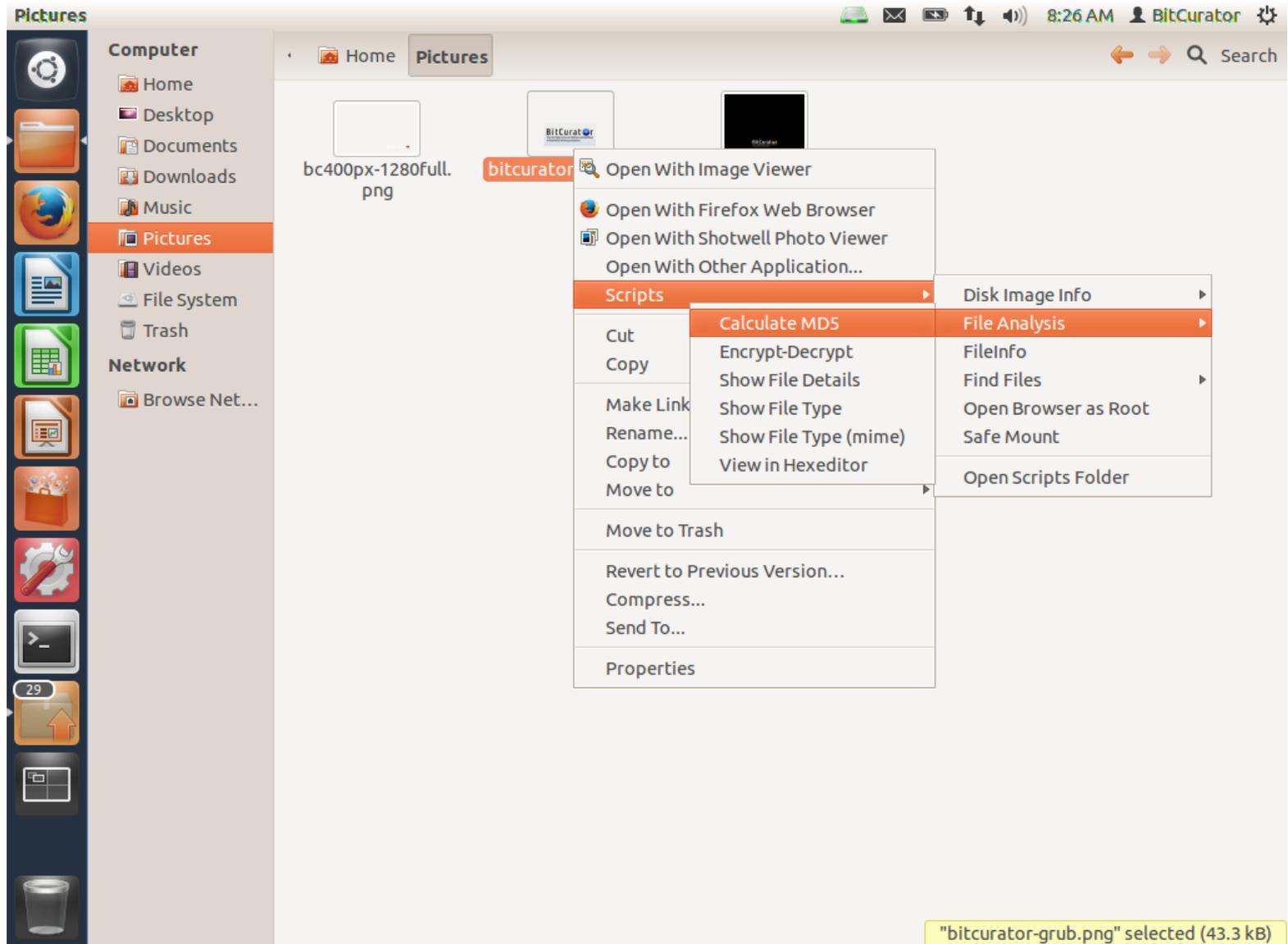
8:31 AM   BitCurator

Computer

Home
Desktop
Documents
Downloads
Music
Pictures
Videos
File System
Trash

Network

Browse Net...

Home   **Pictures**   Search

bc400px-1280full.
png

bitcurator-grub.png

bitcurator-grub-
new.png

Calculate MD5 (Files and Directories)

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

Handling
● Display on screen
○ Save to file (the selected filename + .md5 extenion)

Cancel   OK

29

"bitcurator-grub.png" selected (43.3 kB)

8:30 AM   BitCurator

Computer

Home
Desktop
Documents
Downloads
Music
Pictures
Videos
File System
Trash

Home   Pictures

Search

bc400px-1280full.
png

bitcurator-grub.png

bitcurator-grub-
new.png

Calculate MD5 (Files and Directories)

The MD5 hash of the selected file:

ceb2622125be1231b0fc9babee27942d /home/bcadmin/Pictures/bitcurator-grub.png

Cancel   OK

Network

Browse Net...

29

"bitcurator-grub.png" selected (43.3 kB)

Quick Access to a Hex View:

GHex

Pictures
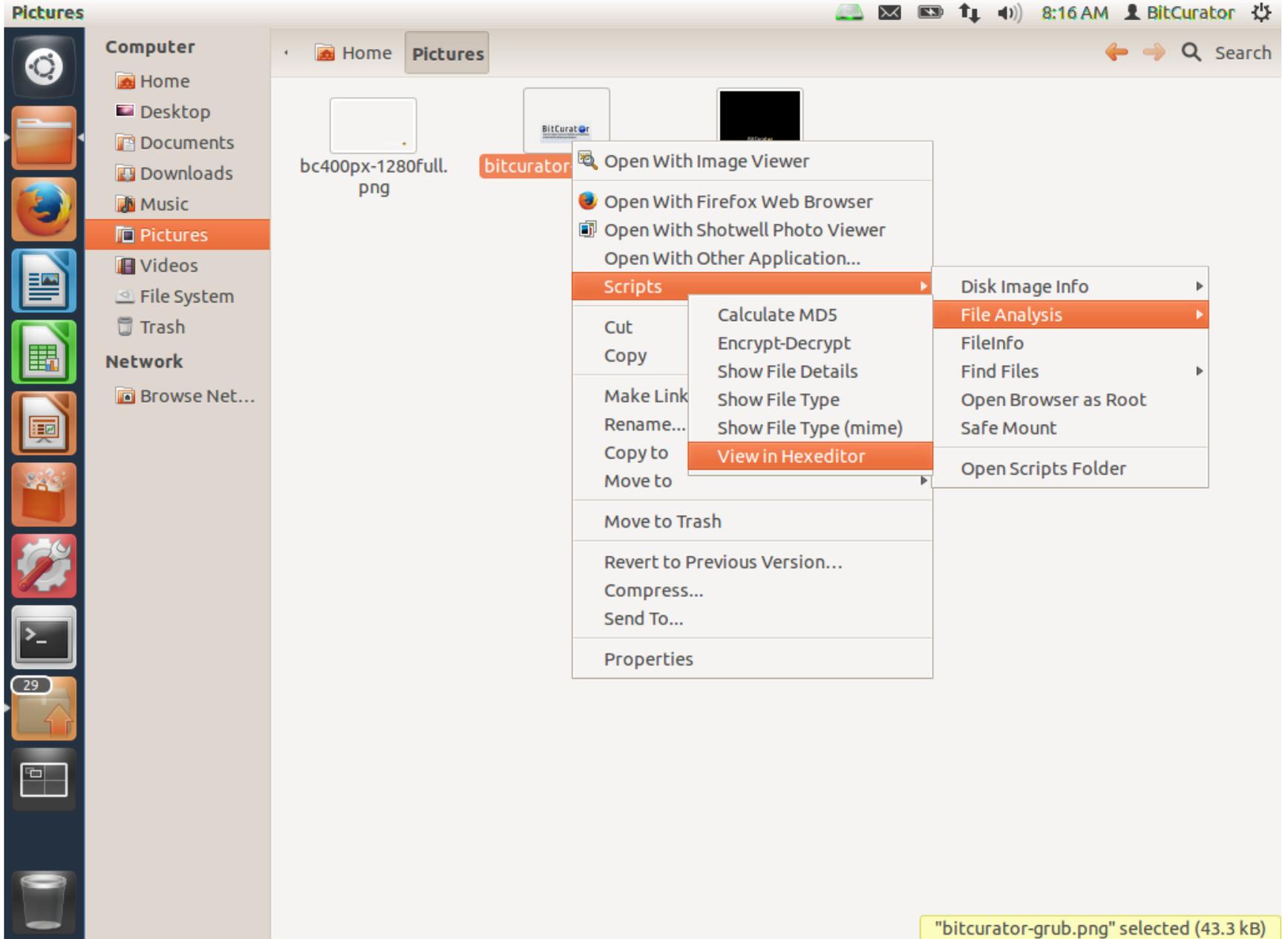
bitcurator-grub.png - GHex

```
00000000 89 50 4E 47 0D 0A 1A 0A 00 00 00 0D 49 48 44 52 00 00    .PNG........IHDR..
00000012 02 80 00 00 01 E0 08 02 00 00 00 BA B3 4B B3 00 00 00    .............K....
00000024 09 70 48 59 73 00 00 0B 13 00 00 0B 13 01 00 9A 9C 18    .pHYs............
00000036 00 00 0A 4F 69 43 43 50 50 68 6F 74 6F 73 68 6F 70 20    ...OiCCPPhotoshop
00000048 49 43 43 20 70 72 6F 66 69 6C 65 00 00 78 DA 9D 53 67    ICC profile..x..Sg
0000005A 54 53 E9 16 3D F7 DE F4 42 4B 88 80 94 4B 6F 52 15 08    TS..=...BK...KoR..
0000006C 20 52 42 8B 80 14 91 26 2A 21 09 10 4A 88 21 A1 D9 15     RB....&*!..J.!...
0000007E 51 C1 11 45 45 04 1B C8 A0 88 03 8E 8E 80 8C 15 51 2C    Q..EE..........Q,
00000090 0C 8A 0A D8 07 E4 21 A2 8E 83 A3 88 8A CA FB E1 7B A3    ......!.......{.
000000A2 6B D6 BC F7 E6 CD FE B5 D7 3E E7 AC F3 9D B3 CF 07 C0    k........>.......
000000B4 08 0C 96 48 33 51 35 80 0C A9 42 1E 11 E0 83 C7 C4 C6    ...H3Q5...B.......
000000C6 E1 E4 2E 40 81 0A 24 70 00 10 08 B3 64 21 73 FD 23 01    ...@..$p....d!s.#.
000000D8 00 F8 7E 3C 3C 2B 22 C0 07 BE 00 01 78 D3 0B 08 00 C0    ..~<<+".....x.....
000000EA 4D 9B C0 30 1C 87 FF 0F EA 42 99 5C 01 80 84 01 C0 74    M..0.....B.\.....t
000000FC 91 38 4B 08 80 14 00 40 7A 8E 42 A6 00 40 46 01 80 9D    .8K....@z.B..@F...
0000010E 98 26 53 00 A0 04 00 60 CB 63 62 E3 00 50 2D 00 60 27    .&S....`.cb..P-.`'
00000120 7F E6 D3 00 80 9D F8 99 7B 01 00 5B 94 21 15 01 A0 91    ........{..[.!....
00000132 00 20 13 65 88 44 00 68 3B 00 AC CF 56 8A 45 00 58 30    . .e.D.h;...V.E.X0
00000144 00 14 66 4B C4 39 00 D8 2D 00 30 49 57 66 48 00 B0 B7    ..fK.9.-.0IWfH...
00000156 00 C0 CE 10 0B B2 00 08 0C 00 30 51 88 85 29 00 04 7B    ..........0Q..)..{
00000168 00 60 C8 23 23 78 00 84 99 00 14 46 F2 57 3C F1 2B AE    .`.##x.....F.W<.+.
0000017A 10 E7 2A 00 00 78 99 B2 3C B9 24 39 45 81 5B 08 2D 71    ..*..x..<.$9E.[.-q
0000018C 07 57 57 2E 1E 28 CE 49 17 2B 14 36 61 02 61 9A 40 2E    .WW..(.I.+.6a.a.@.
0000019E C2 79 99 19 32 81 34 0F E0 F3 CC 00 00 A0 91 15 11 E0    .y..2.4..........
000001B0 83 E3 FD 78 CE 0F AE CE CE 36 8E B6 0E 5E 2D EA BE 06    .x...6...^-...
```

Signed 8 bit:	-119	Signed 32 bit:	1196314761	Hexadecimal:	89
Unsigned 8 bit:	137	Unsigned 32 bit:	1196314761	Octal:	211
Signed 16 bit:	20617	Float 32 bit:	5.281654e+04	Binary:	10001001
Unsigned 16 bit:	20617	Float 64 bit:	5.292398e-260	Stream Length:	8

.PNG........IHDR..

umentation and Help

twork Servers

11:42 PM   BitCurator

tCurator

8:41 AM BitCurator

Computer

Home
Desktop
Documents
Downloads
Music
Pictures
Videos
File System
Trash

Network

Browse Net...

Home Desktop Search

Additional Tools

Documentation and Help

Forensics

| Show AFF Info | Disk Image Info ▶ | Scripts ▶ |
| Show E01 Info | File Analysis ▶ | |

Open

Open With Other Application...

nps-2010-emails. E01

FileInfo

Find Files ▶

Open Browser as Root

Safe Mount

Open Scripts Folder

charlie-work-usb-09-12-11.E01

Cut

Copy

Make Link

Rename...

Copy to ▶

Move to ▶

Move to Trash

Revert to Previous Version...

Compress...

Send To...

Properties

29

"charlie-work-usb-2009-12-11.E01" selected (9.3 MB)

Computer

- Home
- Desktop
- Documents
- Downloads
- Music
- Pictures
- Videos
- File System
- Trash

Network

- Browse Net...
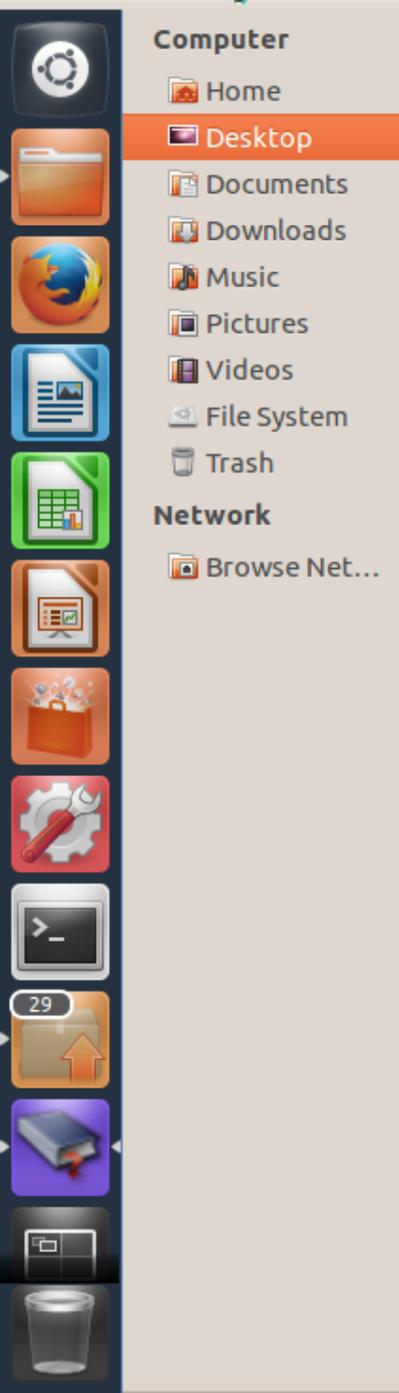
Home   Desktop   Search

EnCase Disk Image Info

ewfinfo 20130416

Acquiry information
        Acquisition date: Wed Jan 19 12:09:18 2011
        System date: Wed Jan 19 12:09:18 2011
        Operating system used: Linux
        Software version used: 20100226
        Password: N/A

EWF information
        File format: EnCase 6
        Sectors per chunk: 64
        Error granularity: 64
        Compression method: deflate
        Compression level: best compression
        Set identifier: 4eb6701d-6cf0-2f4a-a0c6-0cb5d5e20959

Media information
        Media type: fixed disk
        Is physical: yes
        Bytes per sector: 512
        Number of sectors: 2068480
        Media size: 1010 MiB (1059061760 bytes)

Digest hash information
        MD5:   9c0de6c8532d7a66ddcf01861dfb6535

charlie-work-usb-2009-12-11.E01

Cancel   OK

"charlie-work-usb-2009-12-11.E01" selected (9.3 MB)

Quick Start Guide
Most recent version always available at:
http://wiki.bitcurator.net/

BitCurator

Quick Start Guide v0.9.16

Last updated: August 3, 2014

UNC
SCHOOL OF INFORMATION
AND LIBRARY SCIENCE

MITH MARYLAND INSTITUTE FOR
TECHNOLOGY IN THE HUMANITIES

Other Functionality to Meet Identified LAM Needs:

Function	Tool(s)
Identify duplicate files	**FSLint**
Characterize files	**FITS**
Examine, copy and extract information from old Mac disks	**HFSExplorer**
Package files for storage and/or transfer	**BagIt (Java) library**
Scan for viruses	ClamTK
Read contents of Microsoft Outlook PST files	readpst
Examine embedded header information in images	pyExifToolGUI
Generate images of problematic disks or particular disk types	dd, dcfldd, ddrescue, cdrdao (in addition to Guymager)
Identify files that are partially similar but not identical	sdhash, ssdeep

Considerable Uptake

- 118 members of the BitCurator Users mailing list

- 640 individuals who have participated in BitCurator-related events (workshops, tutorials, hackathons)

- Numerous publications and reports by library/archives students and professionals about their testing and use of the software

- 934 Twitter followers (whatever that means)

Open Source Development Strategy

- Rapid development with numerous iterations based on several channels of user feedback

- Code released under GPL, v3 (perhaps moving to Apache) – available through GitHub

- Existing code incorporated is generally GPL or public domain (government products)

- Packaging elements of the code to be integrated into other environments (e.g. Archivematica)

- Significant community engagement, including work of Community Lead (Porter Olsen)

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

- Institutions as members, with two categories of membership: Charter and General

- Software and documentation will continue to be free and open source, but membership provides further benefits (e.g. support, training, development priority)

DIMAC (Disk Image Access for the Web)

- Developed by Sunitha Misra and Kam Woods

- To dynamically navigate and download contents of a disk image, without having to download or mount the full image

- See: https://github.com/kamwoods/dimac

- Demo at:

http://www.youtube.com/watch?v=BwiWFqxYzQ8

See: Sunitha Misra, Christopher A. Lee, and Kam Woods, "A Web Service for File-Level Access to Disk Images," *Code4Lib Journal*, http://journal.code4lib.org/articles/9773.

APPLYING FORENSICS TO PRESERVING THE PAST: CURRENT ACTIVITIES AND FUTURE POSSIBILITIES

- Organizers: Cal Lee, Jeremy Leighton John, Susan Thomas
- To be held at Digital Libraries 2014, London, September, 8-12, 2014
- One-day event, split across an afternoon and following morning (Sept 11-12)
- Short papers and talks, group discussion and formation of steps for further action

Thank You!

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

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

Twitter: @bitcurator