An Introduction to the BitCurator Environment

Cal Lee

School of Information and Library Science

University of North Carolina, Chapel Hill

American Libraries Association Annual Meeting

ALCTS PARS Preservation Metadata Interest Group

June 28, 2014

Las Vegas, NV

BitCurator

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Acquiring Born-Digital Materials: Same Goals as When Acquiring Analog Materials

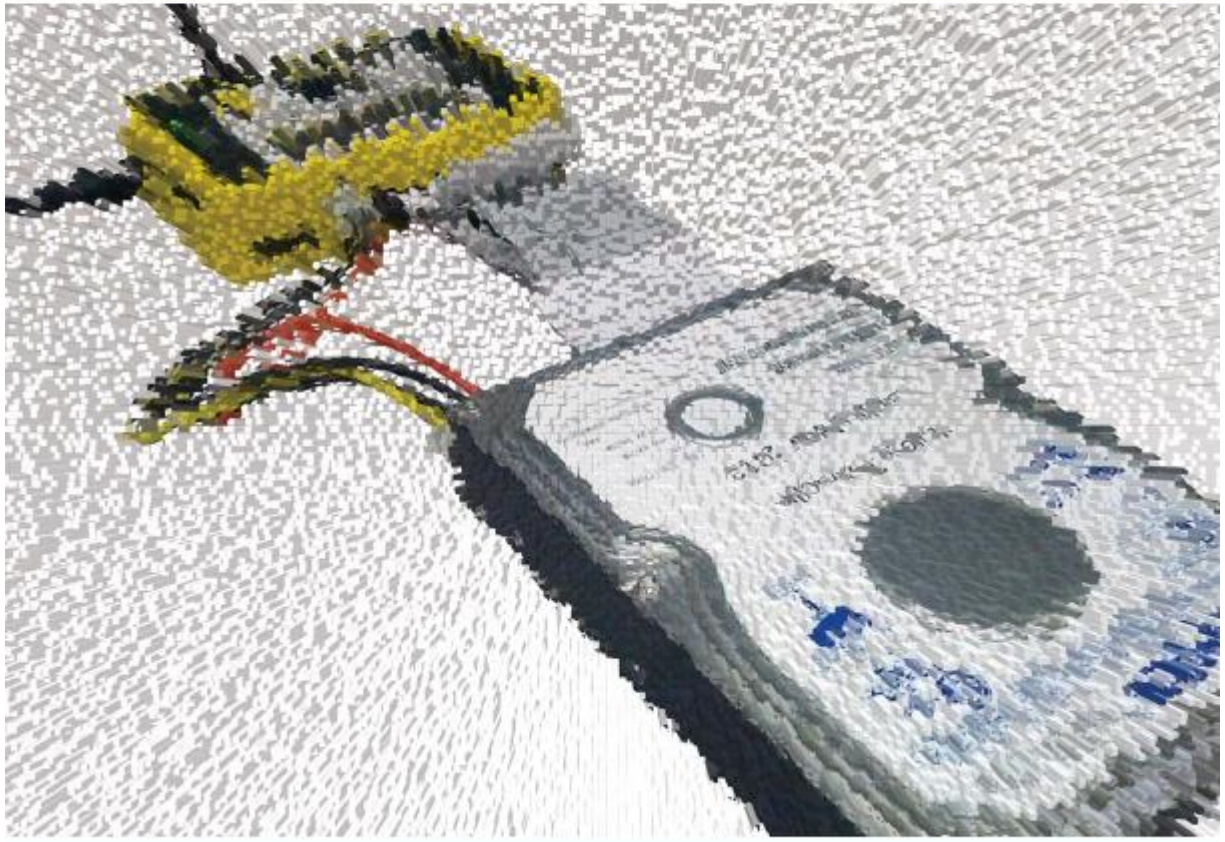
- Ensure integrity of materials
- Allow users to make sense of materials and understand their context
- Prevent inadvertent disclosure of sensitive data

Applying Digital Forensics to Library Materials

- Recognition of how data can be recovered when layers of technology fail or are no longer available
- Capturing information from places that are not immediately visible
- Ensuring that actions taken on files don't make irreversible changes to essential characteristics (e.g. MAC values)
- Attending to order of volatility – some types of data change more quickly and often than others
- Learning about available tools and techniques to deal with files
- Established practices for documenting acquisition and processing, so others will know what might have changed
- Overlap between technical knowledge required to do digital forensics and ad hoc acquisition of digital materials by libraries/archives

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

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)





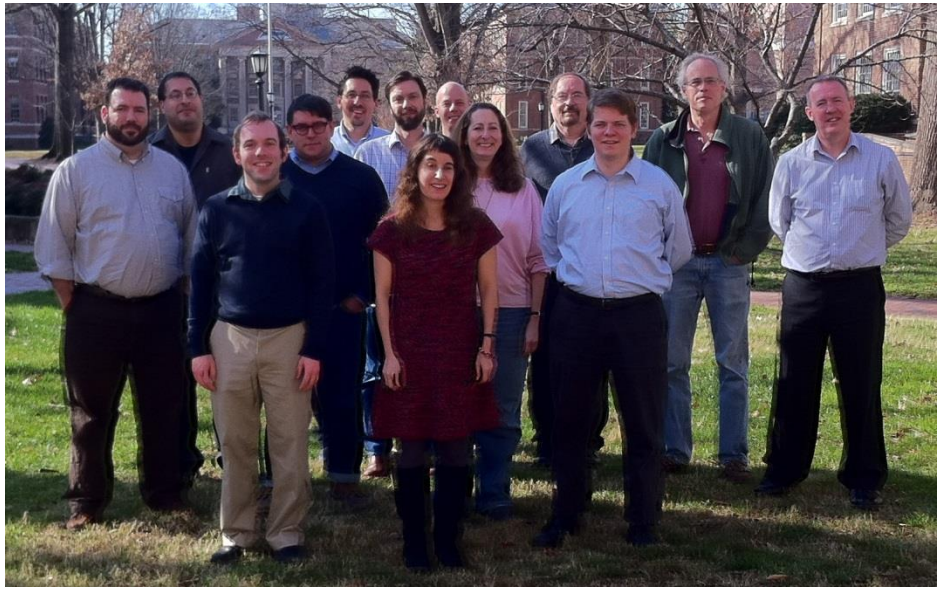

Two Groups of Advisors

Professional Experts Panel

- Bradley Daigle, University of Virginia Library
- Erika Farr, Emory University
- Jennie Levine Knies, University of Maryland
- Jeremy Leighton John, British Library
- Leslie Johnston, Library of Congress
- Naomi Nelson, Duke University
- Erin O'Meara, Gates Archive
- Michael Olson, Stanford University Libraries
- Gabriela Redwine, Harry Ransom Center, University of Texas
- Susan Thomas, Bodleian Library, University of Oxford

Development Advisory Group

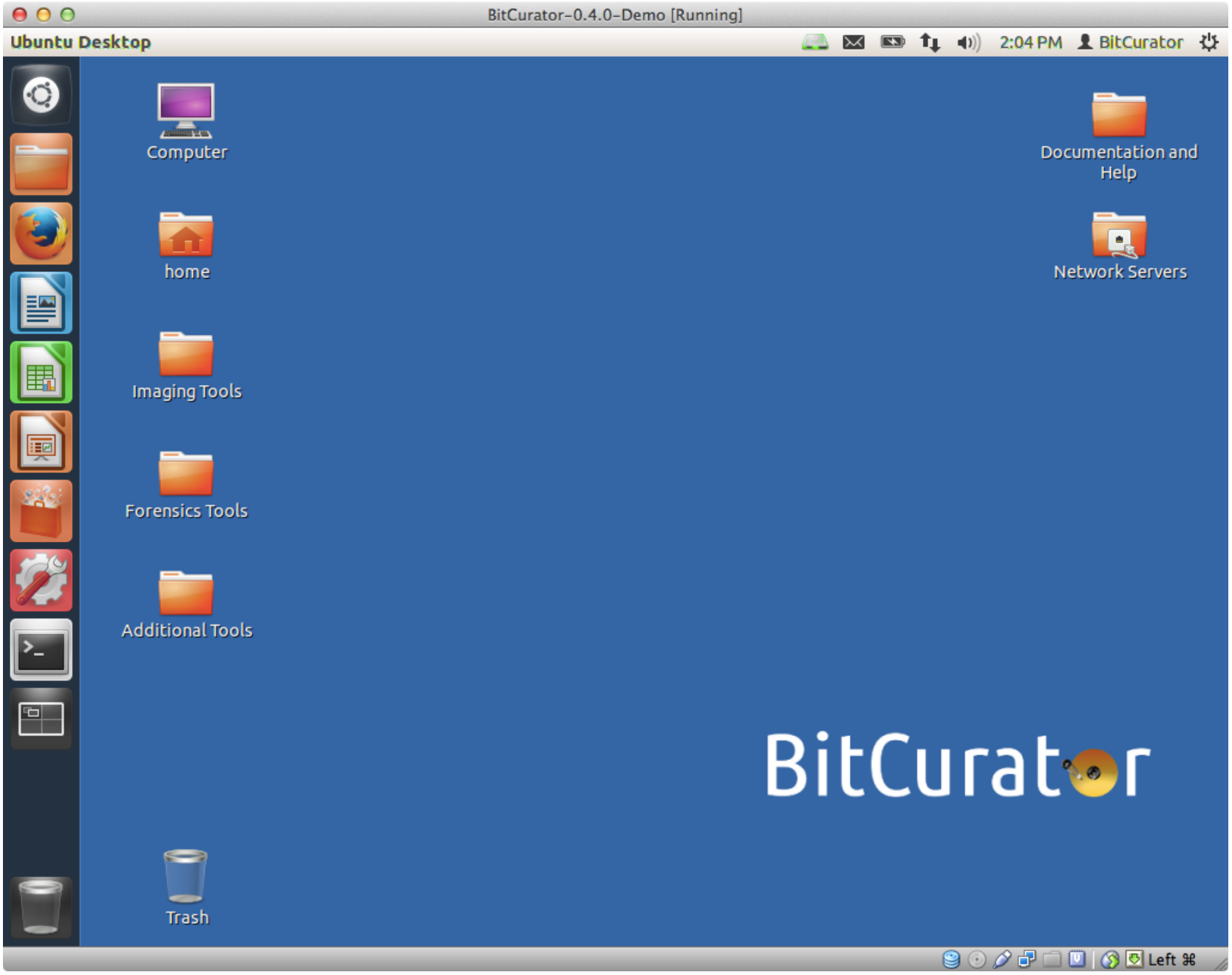
- Barbara Guttman, National Institute of Standards and Technology
- Jerome McDonough, University of Illinois
- Mark Matienzo, Yale University
- Courtney Mumma, Artefactual Systems
- David Pearson, National Library of Australia
- Doug Reside, New York Public Library
- Seth Shaw, University Archives, Duke University
- William Underwood, Georgia Tech



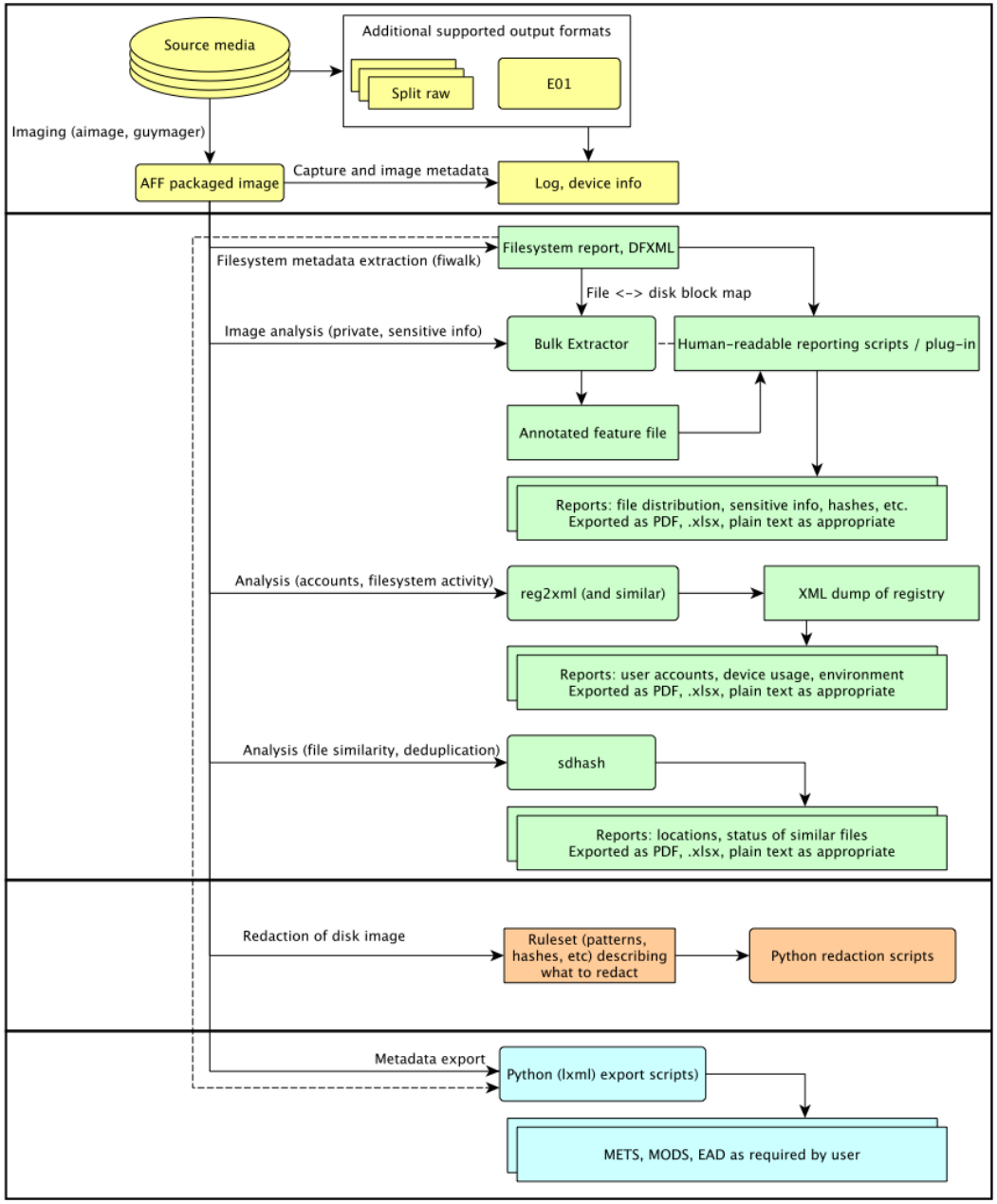
BitCurator Environment*

- Bundles, integrates and extends functionality (primarily data capture and reporting) 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-0.4.0-Demo [Running]
Ubuntu Desktop
2:04 PM
BitCurator
Computer
home
Imaging Tools
Forensics Tools
Additional Tools
Trash
Documentation and Help
Network Servers
BitCurator
Left ⌘

BitCurator-Supported Workflow Elements



Acquisition

Reporting

Redaction

Metadata export

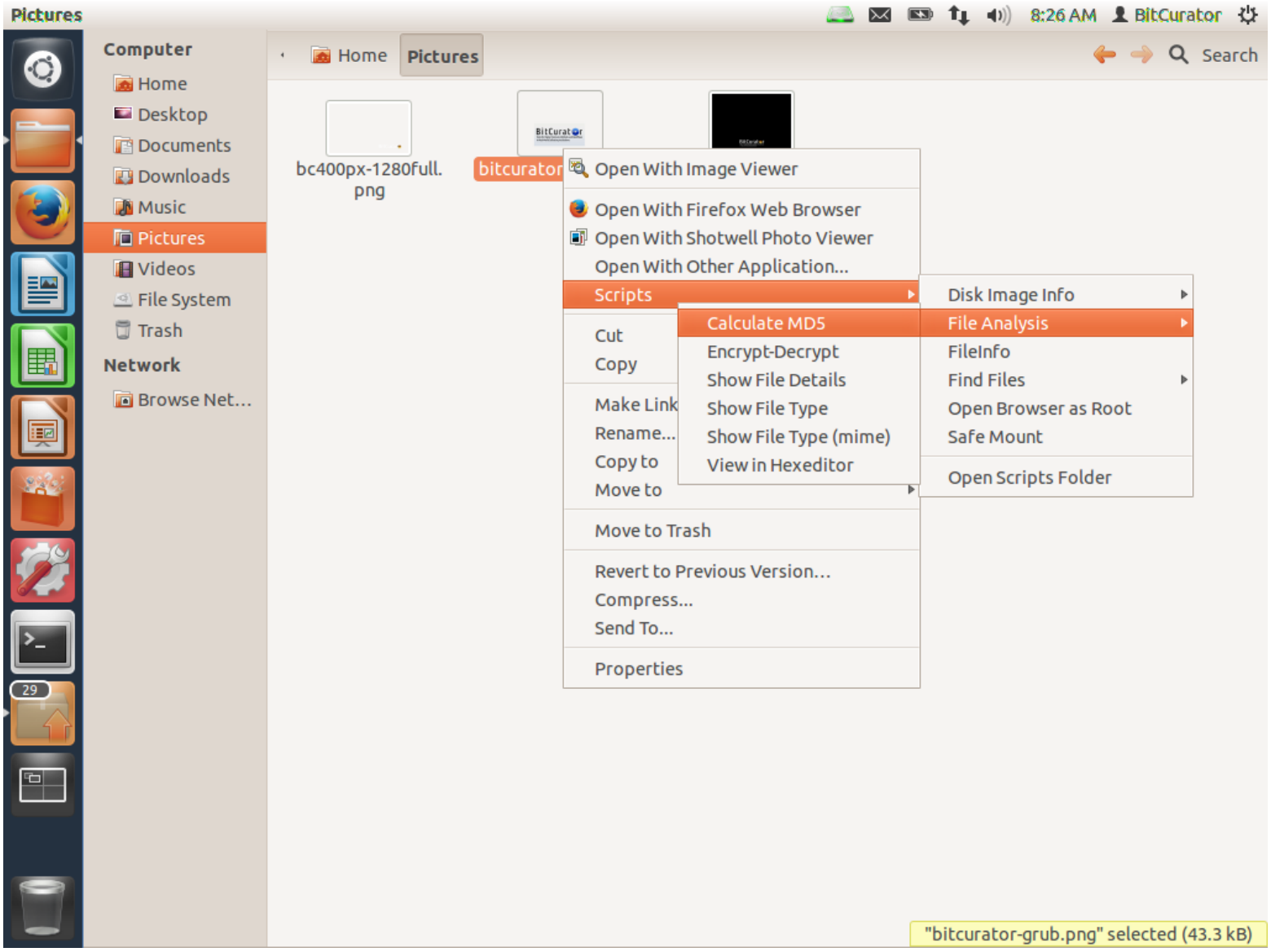
- **Acquisition**
- **Reporting**
- **Redaction**
- **Metadata Export**

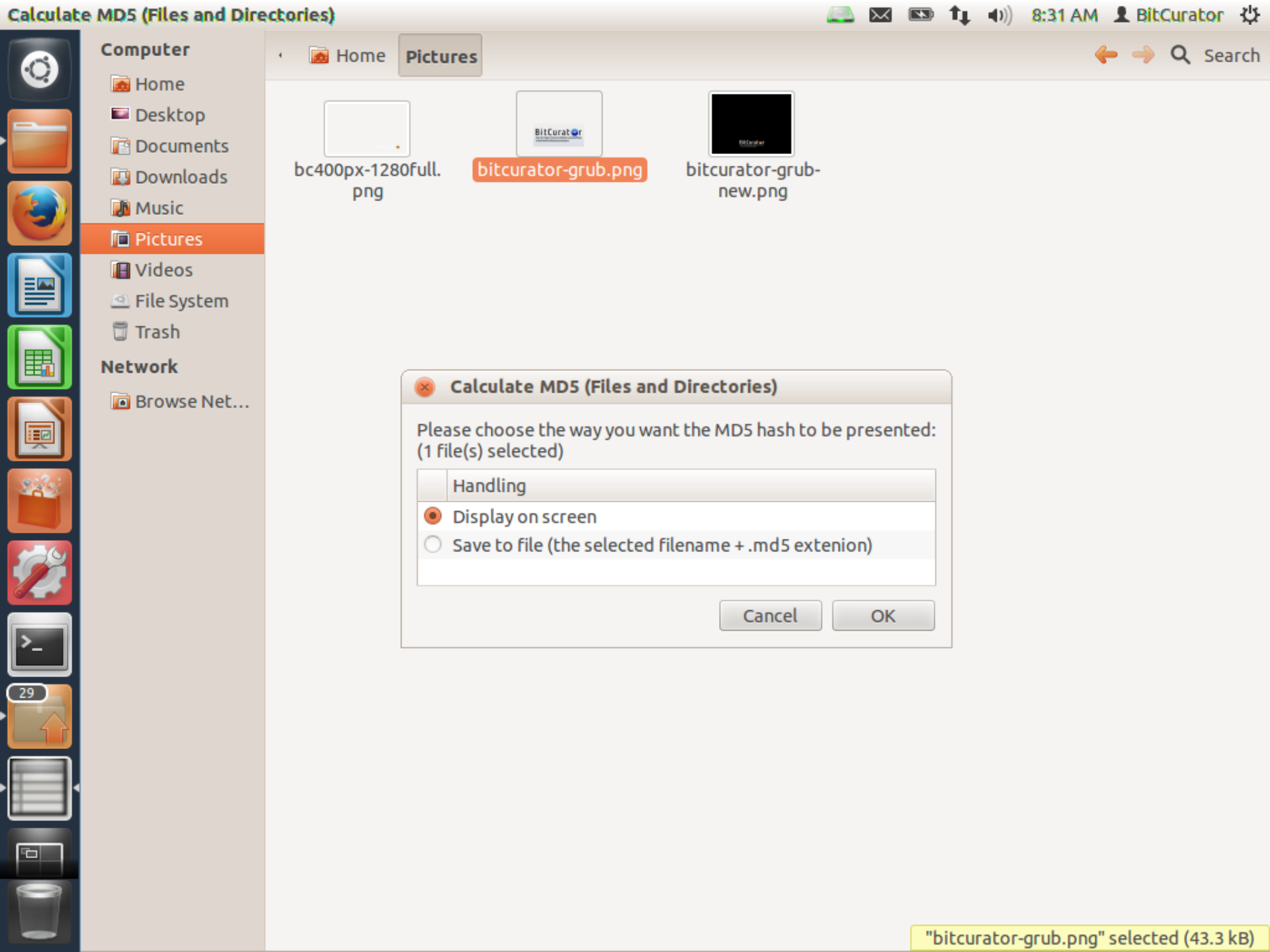
See: http://bitcurator.net

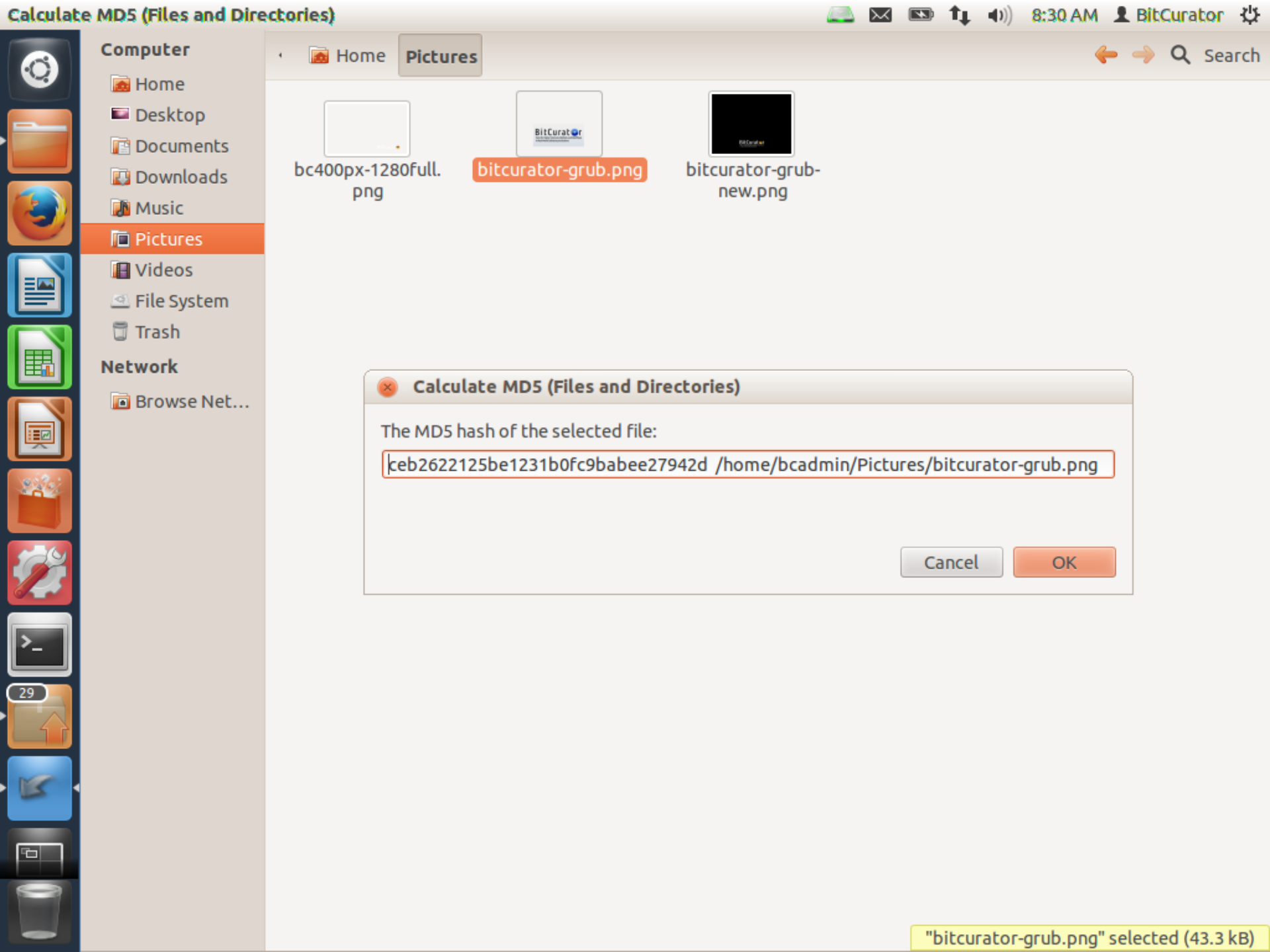
Cryptographic Hashes (aka Checksums) – Compact Representations of Bitstreams

- A given bitstream, fed into an algorithm, will generate a short string of characters that is **extremely** unlikely to be generated by a different bistream 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




Calculate MD5 (Files and Directories)
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
"bitcurator-grub.png" selected (43.3 kB)


Calculate MD5 (Files and Directories)
8:30 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)
The MD5 hash of the selected file:
ceb2622125be1231b0fc9babee27942d /home/bcadmin/Pictures/bitcurator-grub.png
Cancel
OK
"bitcurator-grub.png" selected (43.3 kB)

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)

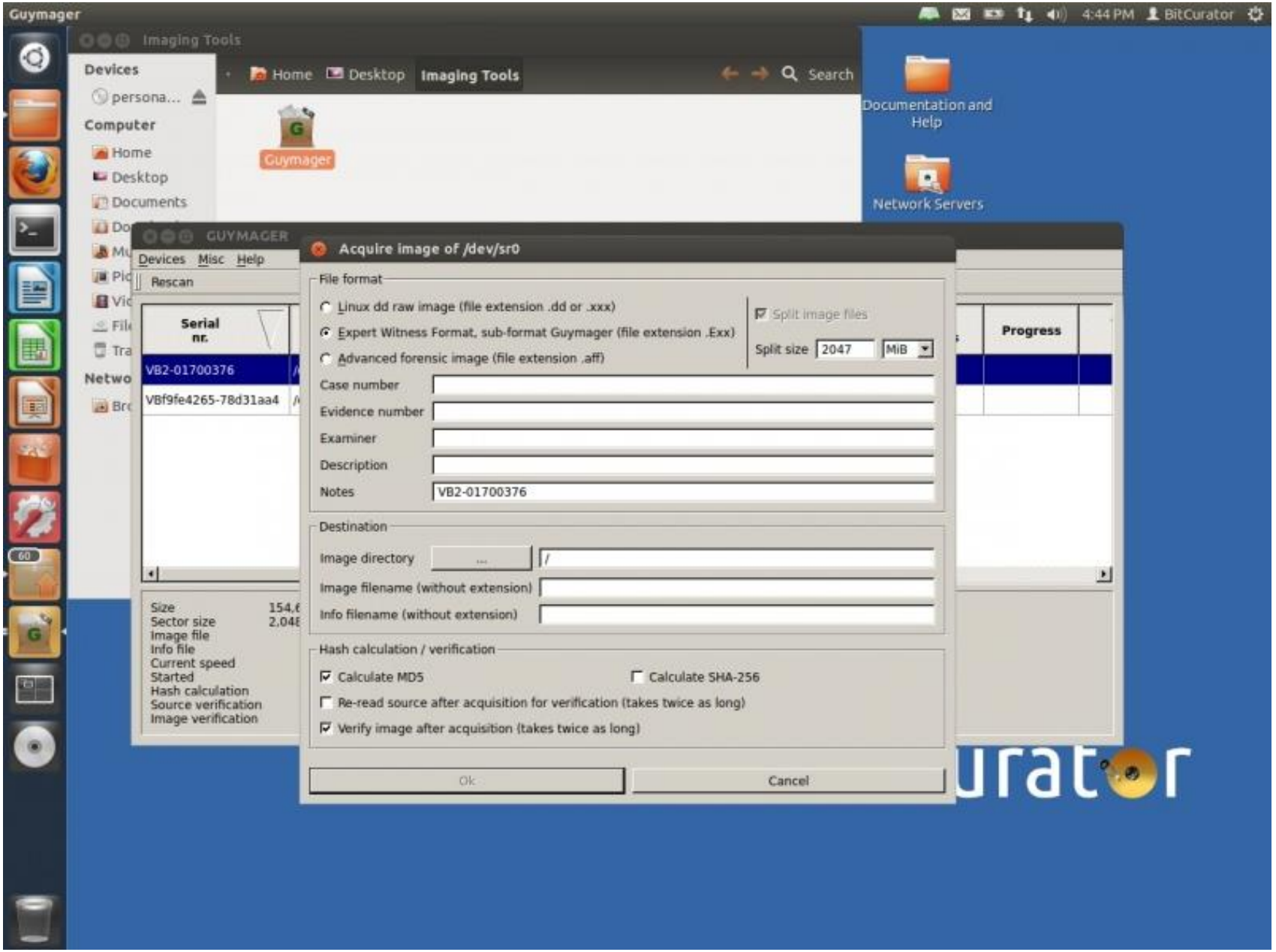
Strategies for avoiding accidental manipulation of volatile data

- Use write-blocking equipment when first reading from a medium (hardware, if possible)
- Make bit-level image
- Create checksums before and after file transfers and transformations
- Pay special attention to irreversible changes

Getting below the File System – Low-Level Copying

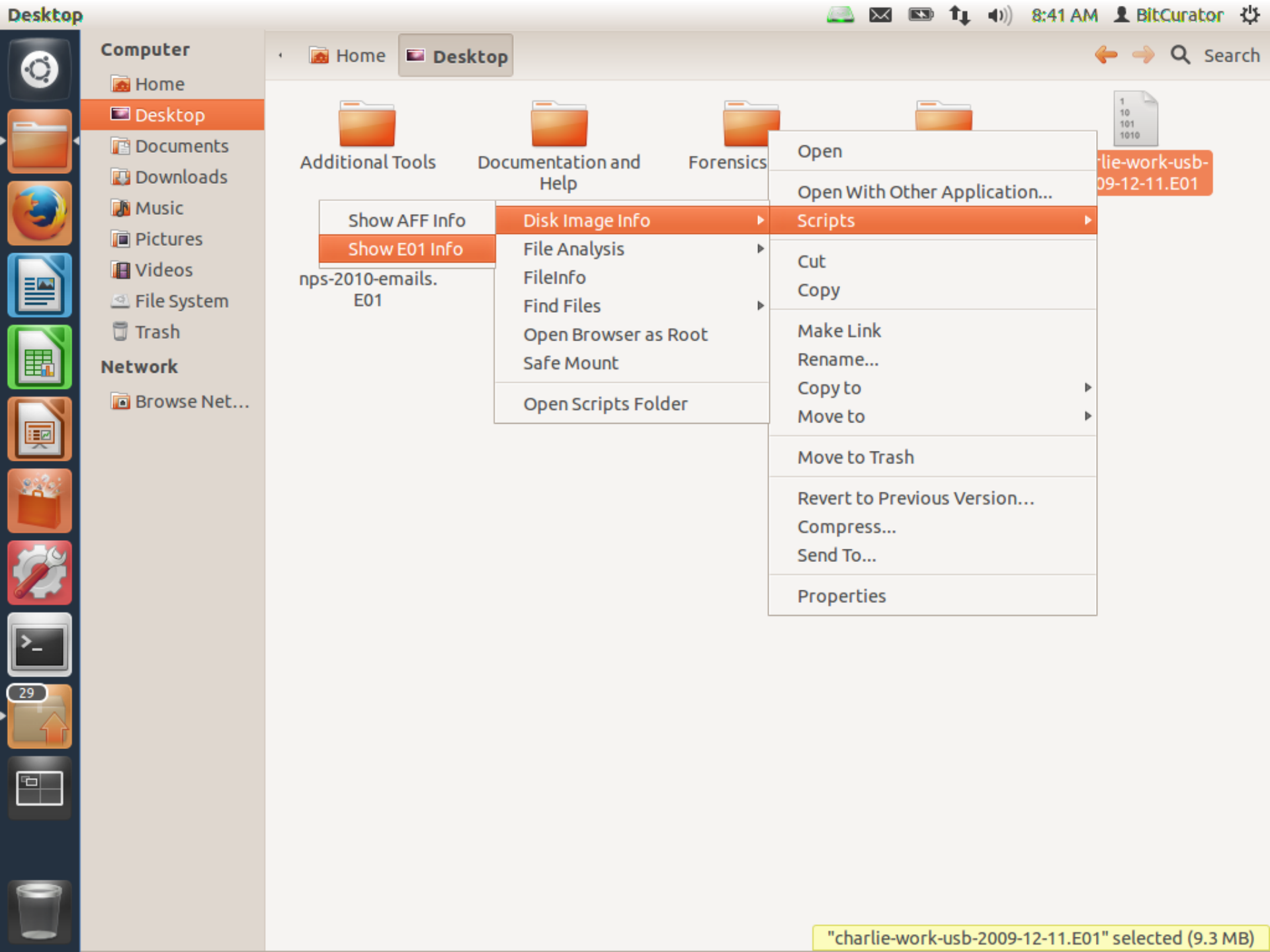
- Getting an "image" of a storage medium involves working at a level below the file system
  - Can get at file attributes and deleted files not visible through higher-level copy operations
- Most commonly used tool is dd (or variant) - UNIX program for low-level copying and conversion of data from a storage device
- More specialized tools for creating forensic images include:
  - FTK Imager
  - Guymager
  - Imaging utilities in commercial applications (including EnCase and FTK)

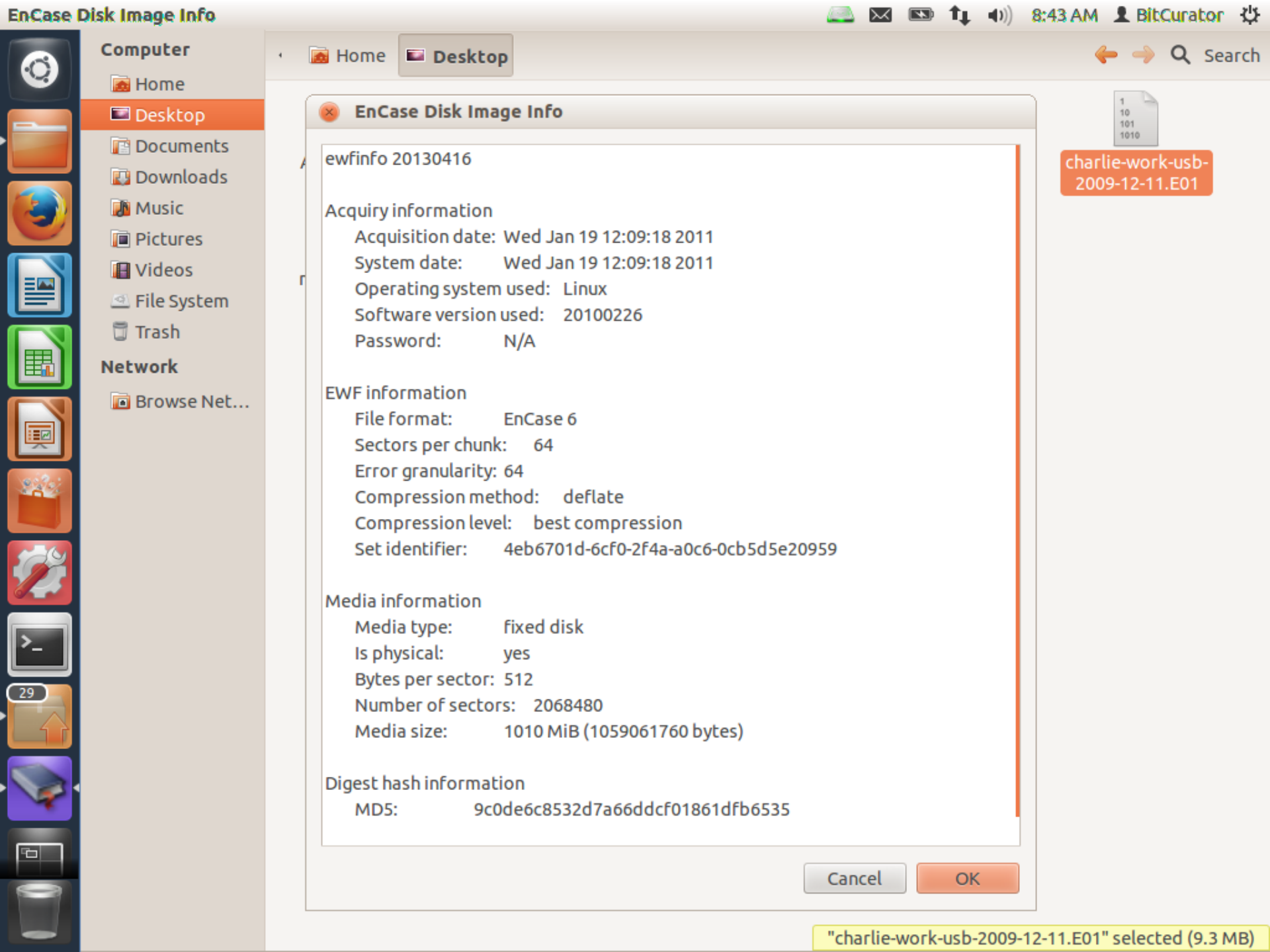
Creating a Disk Image in Guymager



Why Create Disk Images?

- **Simplify and compartmentalize** processing tasks – don't need to solve all technical challenges at the same time
- Make sure full set of bits is safe – e.g. have the disk but not depend on fragile physical medium
- Surprises about how things were structured within the file system
- You could inadvertently change something in the act of examining or dealing with the files
- Proof of file integrity and chain of custody
- Corrupted files and viruses - to determine what subset of the bitstream can be recovered
- Recovery of traces of online activity
- Avoid irreversible transformations
- Changes in preservation strategy over time

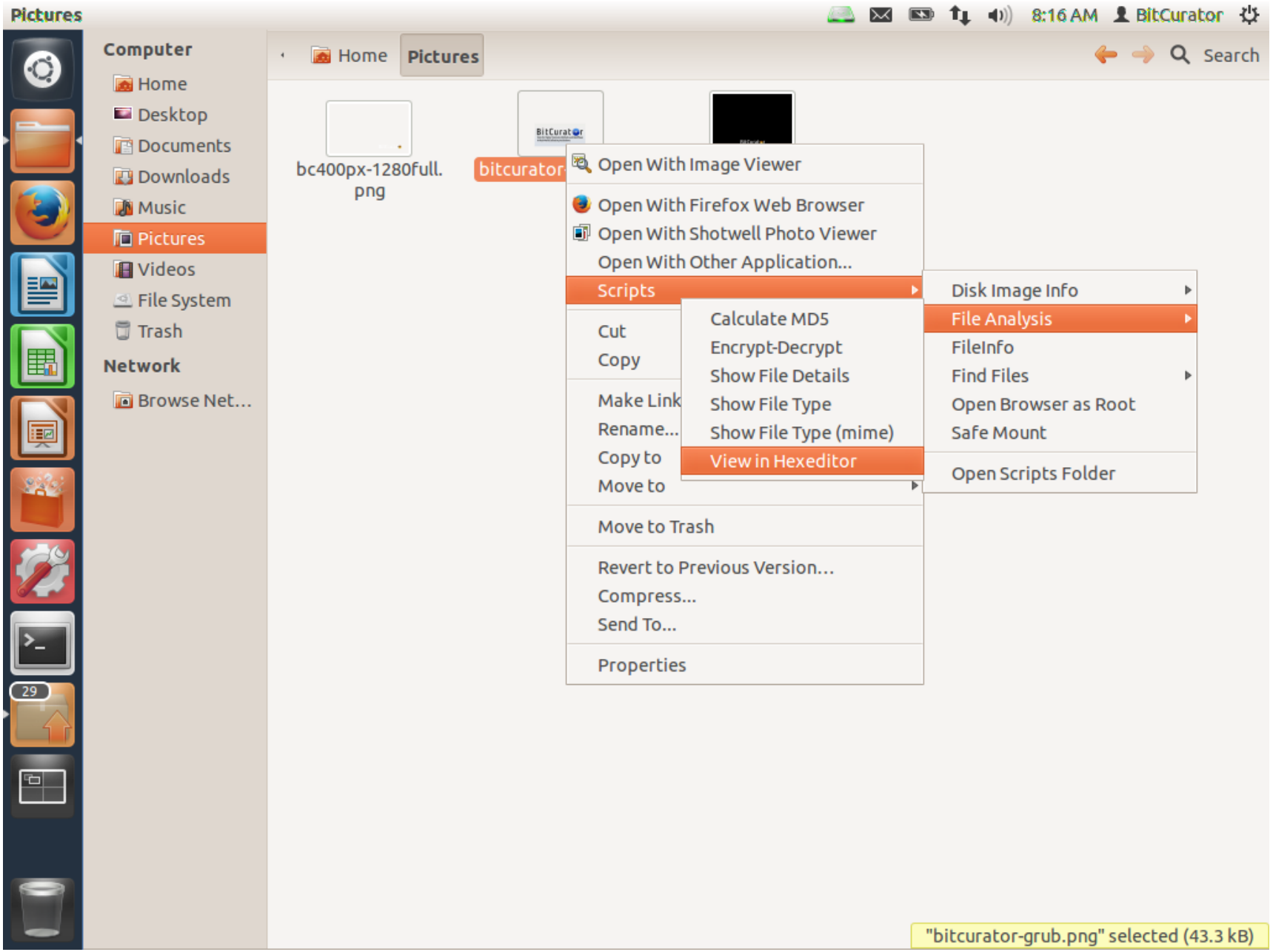

Desktop
8:41 AM
BitCurator
Computer
Home
Desktop
Documents
Downloads
Music
Pictures
Videos
File System
Trash
Network
Browse Net...
Home
Desktop
Search
Additional Tools
Documentation and Help
Forensics
nps-2010-emails.E01
Show AFF Info
Show E01 Info
Disk Image Info
File Analysis
FileInfo
Find Files
Open Browser as Root
Safe Mount
Open Scripts Folder
Open
Open With Other Application...
Scripts
Cut
Copy
Make Link
Rename...
Copy to
Move to
Move to Trash
Revert to Previous Version...
Compress...
Send To...
Properties
"charlie-work-usb-2009-12-11.E01" selected (9.3 MB)

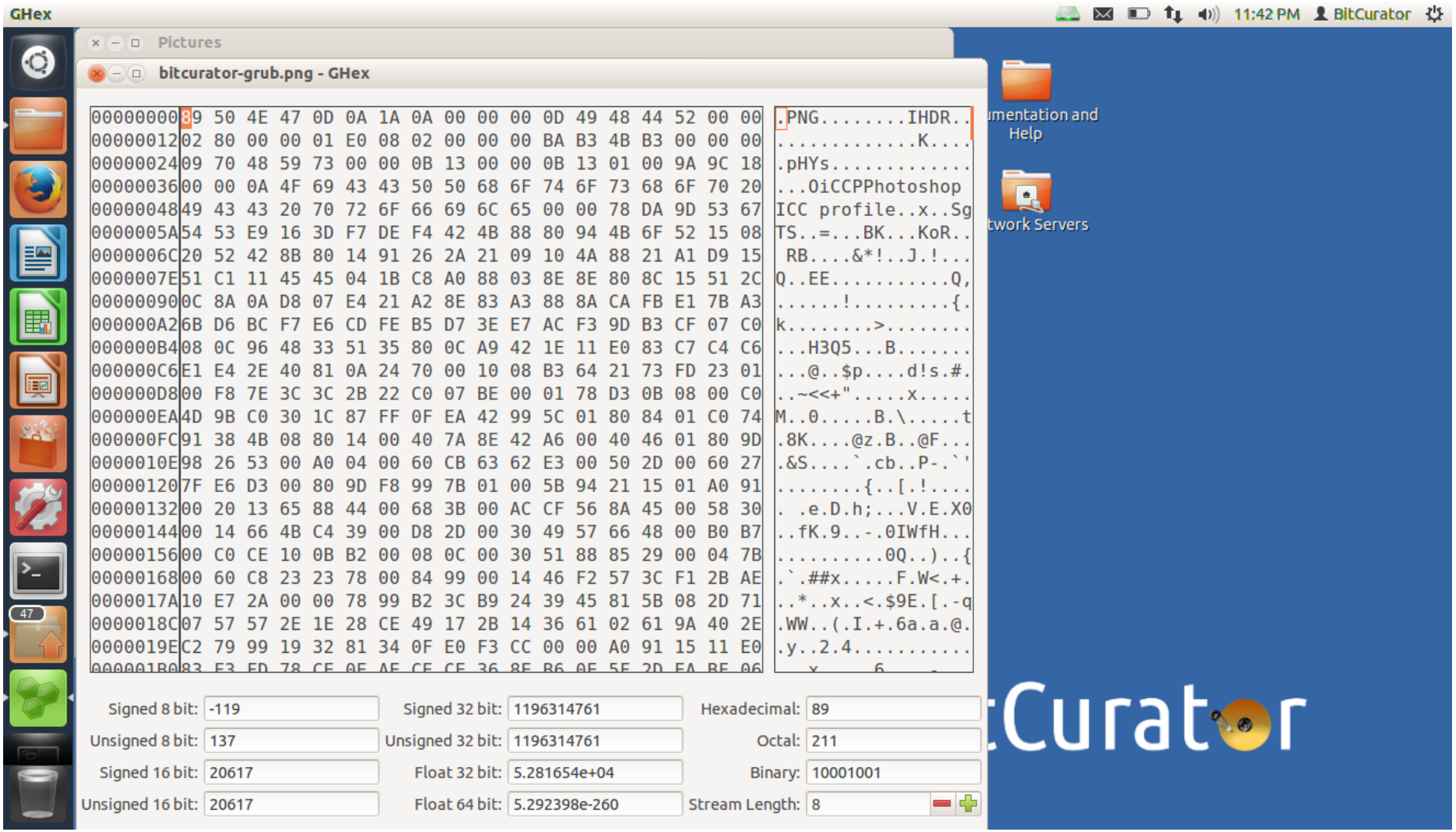

EnCase Disk Image Info
8:43 AM
BitCurator
Computer
Home
Desktop
Documents
Downloads
Music
Pictures
Videos
File System
Trash
Network
Browse Net...
Home
Desktop
Search
EnCase Disk Image Info
ewfinfo 20130416
Acquiry information
Acquisition date: Wed Jan 19 12:09:18 2011
System date: Wed Jan 19 12:09:18 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
Cancel
OK
charlie-work-usb-2009-12-11.E01
"charlie-work-usb-2009-12-11.E01" selected (9.3 MB)

Hex Dump

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

In the BitCurator environment:

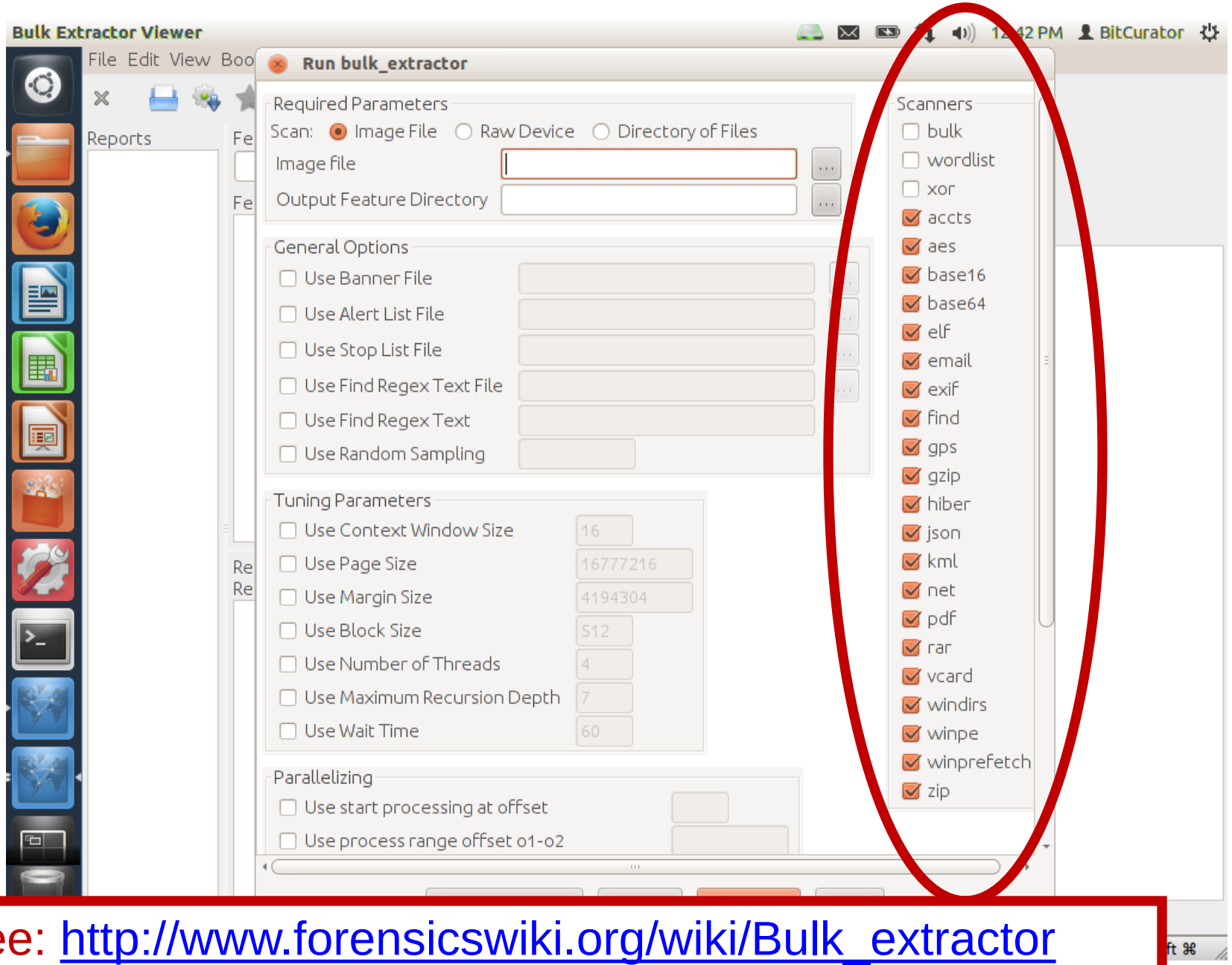



GHex
11:42 PM
BitCurator
Pictures
bitcurator-grub.png - GHex
00000000 89 50 4E 47 0D 0A 1A 0A 00 00 00 0D 49 48 44 52 00 00 .PNG........IHDR..
00000012 02 80 00 00 01 E0 08 02 00 00 00 BA B3 4B B3 00 00 00 ....K....
00000024 09 70 48 59 73 00 00 0B 13 00 00 0B 13 01 00 9A 9C 18 .pHYs.............
00000036 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.........
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
umentation and Help
twork Servers
Curator

Identifying “Features” of Interest in Disk Images

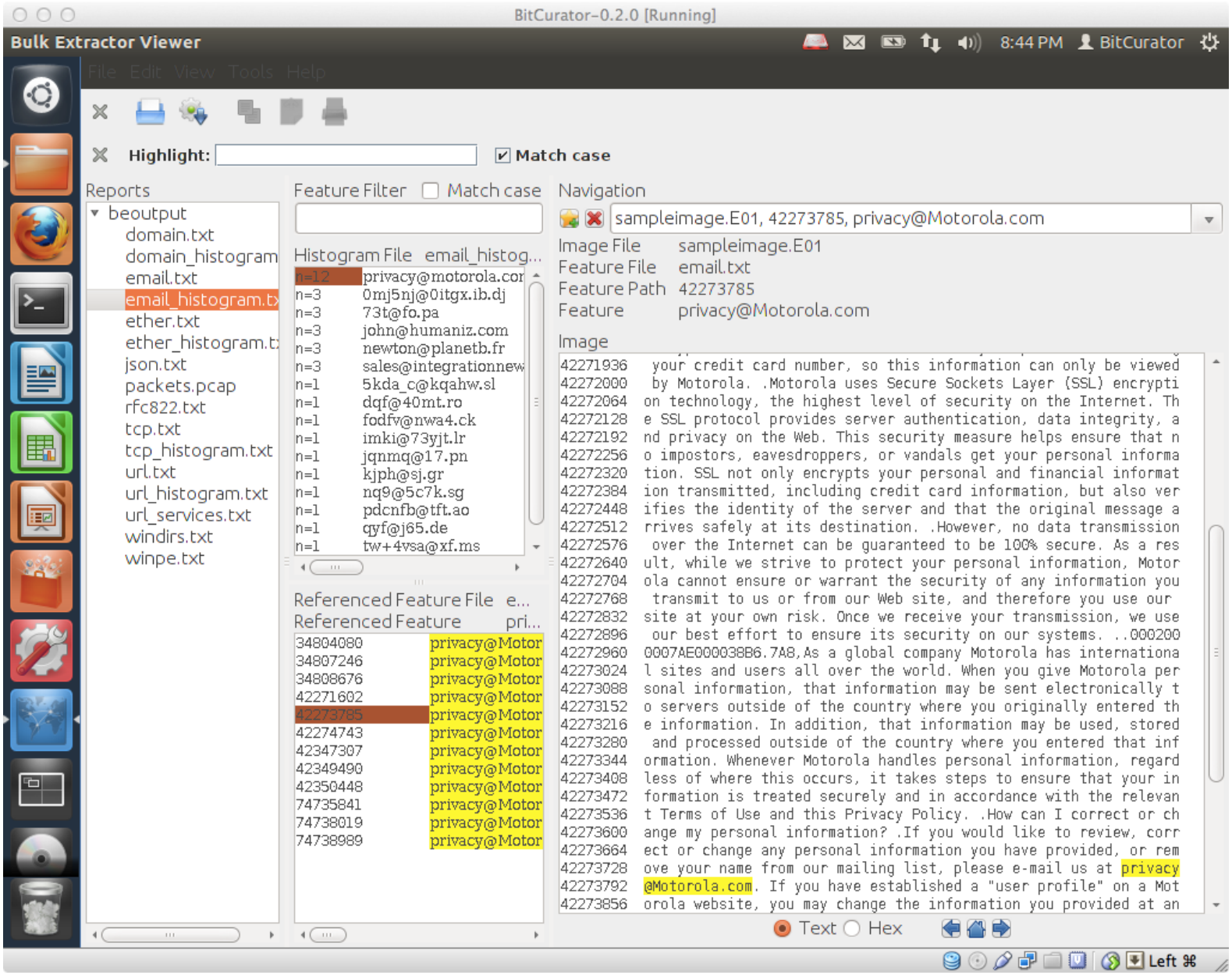
Bulk Extractor

Bulk Extractor Scanning Options

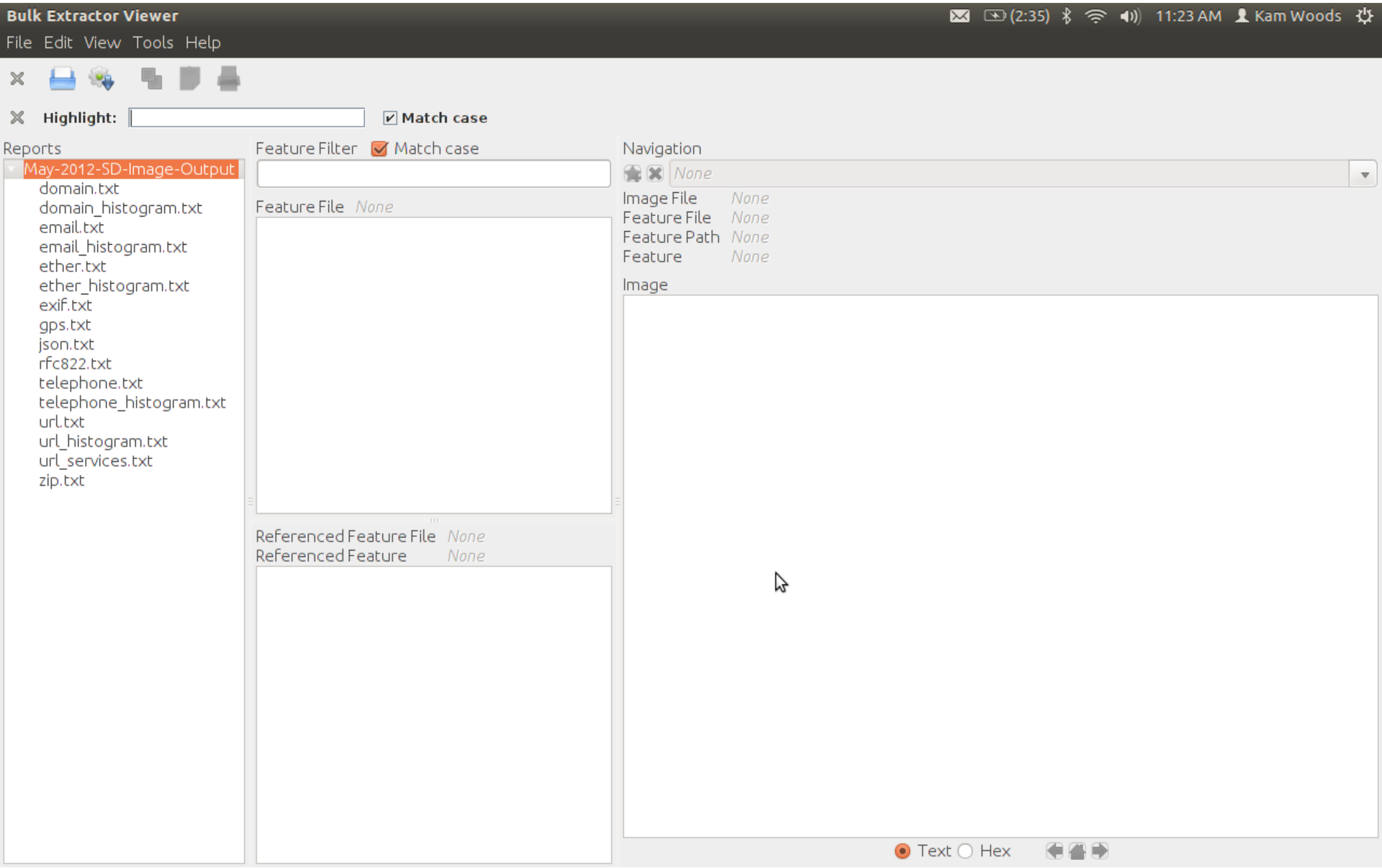


See: http://www.forensicswiki.org/wiki/Bulk_extractor

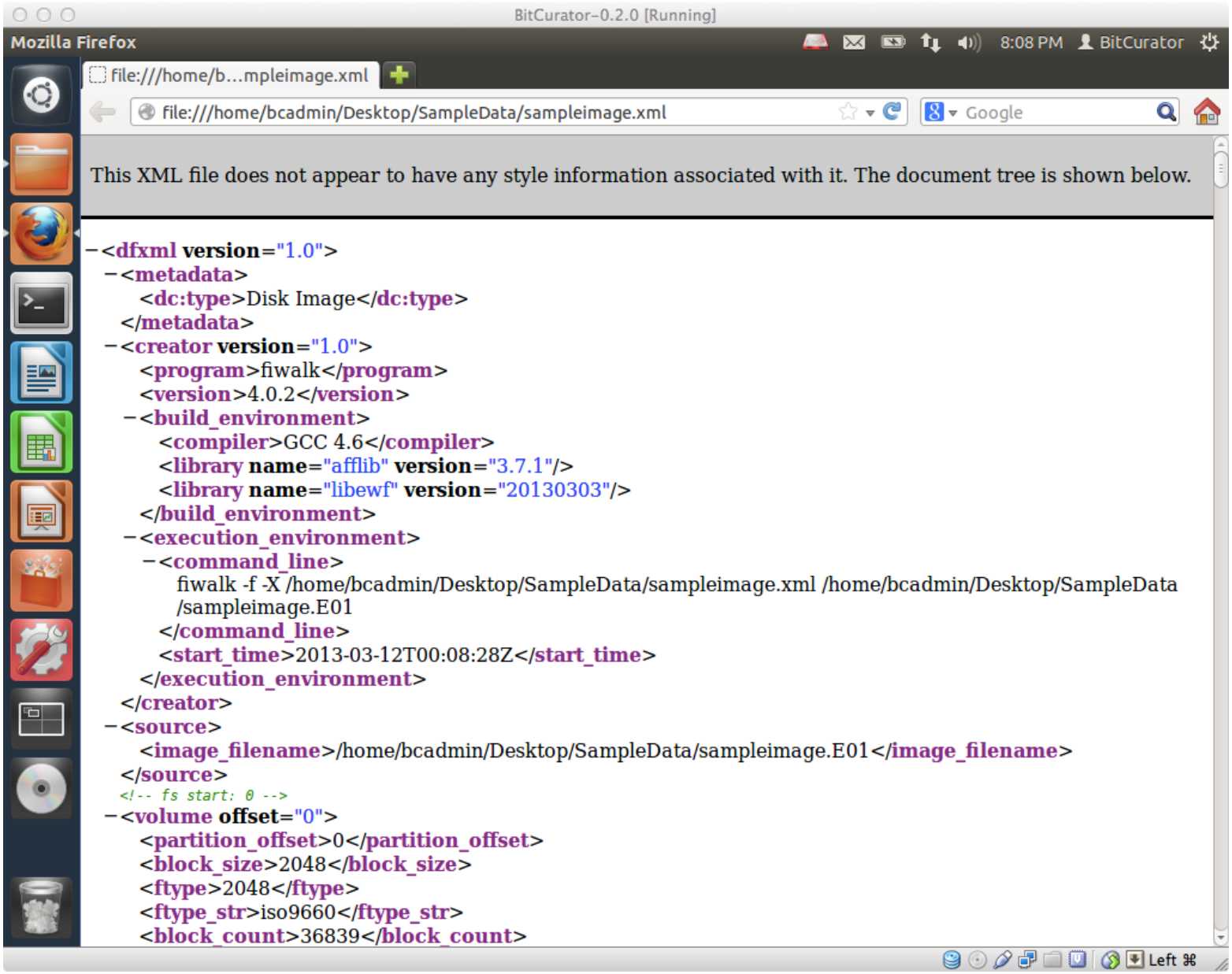
Histogram of Email Addresses (Specific Instances in Context on Right)



Bulk Extractor Reports



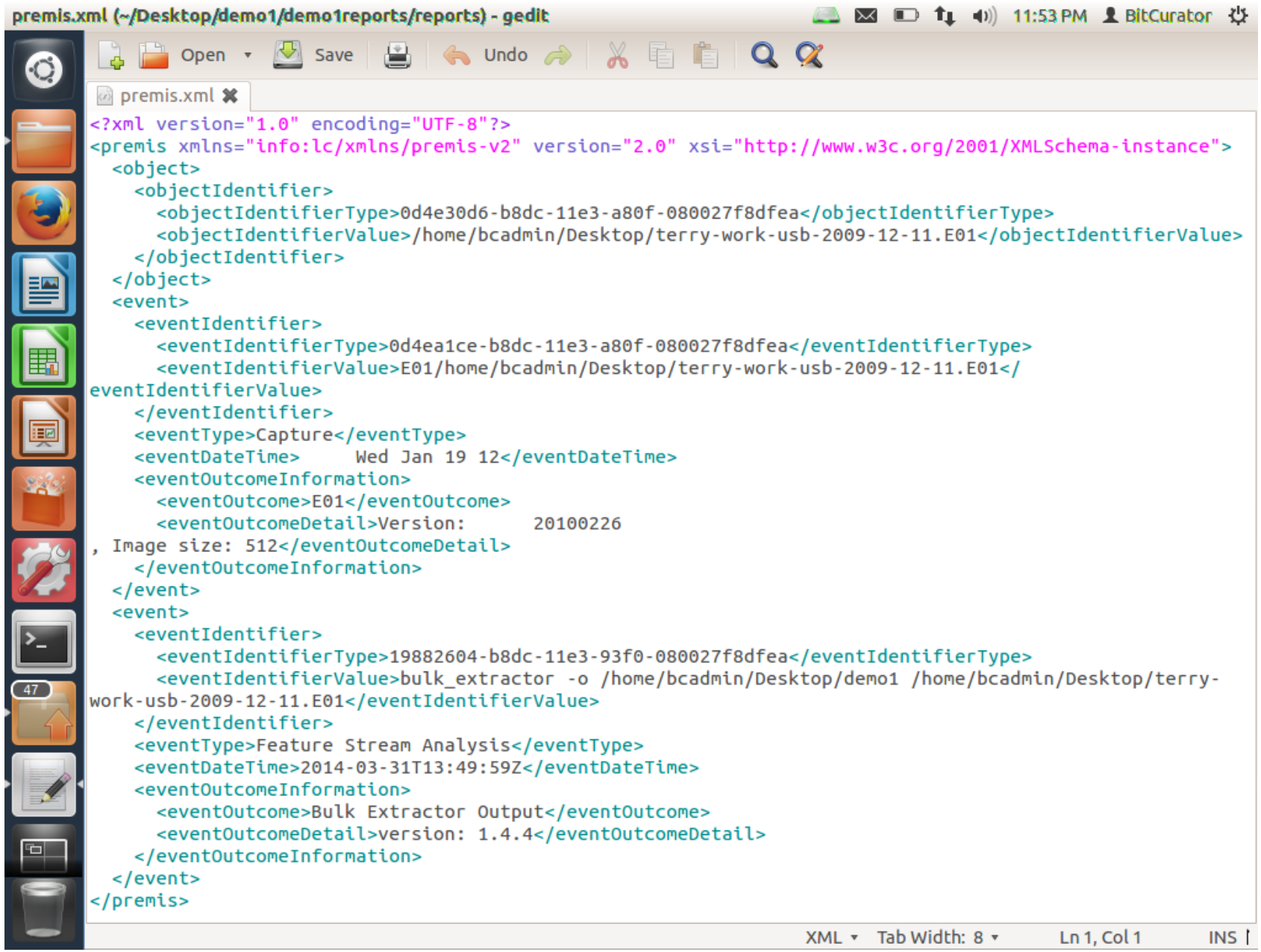
Metadata about a Captured Disk



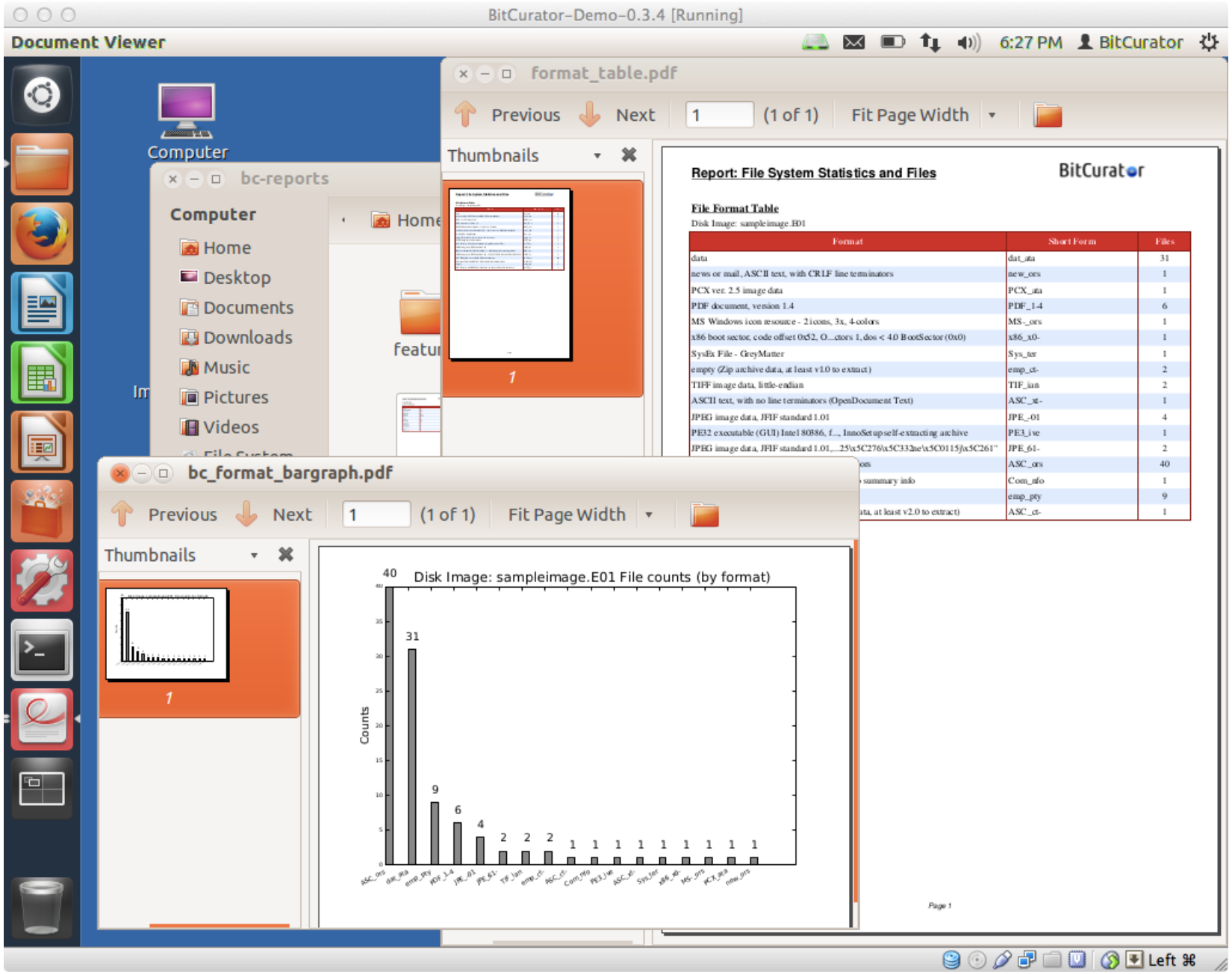
Filesystem Metadata about a Specific File - Output from fiwalk

```
<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>
```

PREMIS Metadata Generated from Running BitCurator Tools



Various Specialized BitCurator Reports



Nautilus Scripts

- In addition to the specialized forensics tools in the BitCurator environment, there are a variety of scripts that can be run using the GNOME file manager called Nautilus (Linux analog to Windows Explorer or Mac OS X Finder)
- Can be used in the BitCurator environment or your own Linux environment
- You've already seen several of these (calculating MD5s, showing in hex view, showing .E01 disk image internal metadata

Other Functionality:

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

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

BitCurator

Quick Start Guide v0.9.12

Last updated: June 9, 2014

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

MITH MARYLAND INSTITUTE FOR TECHNOLOGY IN THE HUMANITIES

Open Source Software Strategy

- Code released under GPL, v3 (perhaps moving to Apache License) – available through GitHub
- Existing code incorporate is generally GPL or public domain (government products)
- Packaging elements of the code to be integrated into other environments (e.g. Archivematica)
- Regular contact with individuals and organizations responsible for other development efforts

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
- The most important member benefit is assurance that the BitCurator software will persist in future years

http://www.bitcurator.net/bitcurator-consortium/

Other Membership Benefits

General Members:

- Access to a BitCurator Consortium help desk
- Prioritization in future enhancement requests
- Dedicated educational offerings
- Voting rights
- Eligibility to serve on the BitCurator Consortium Executive Council and BitCurator Consortium Committees
- Service opportunities
- Community engagement and networking
- Professional development and training
- Subscription to a dedicated BitCurator Consortium member electronic mailing list
- Special rates for BitCurator Consortium events, including the annual BitCurator User Forum

Charter Members - all to the left, and:

- Opportunity to participate in and shape the initial BitCurator Consortium Executive Council and BitCurator Consortium Committees, including exclusive eligibility for election or appointment to the Executive Council (General Members can serve on committees but will not be eligible for election to the Executive Council before 2015).
- Participation in the development of the initial BitCurator Consortium user, technical and services roadmaps.
- Recognition through the placement of your institution name, logo and link on the BitCurator Consortium web site.
- Use of the “BitCurator Consortium Charter Member” icon

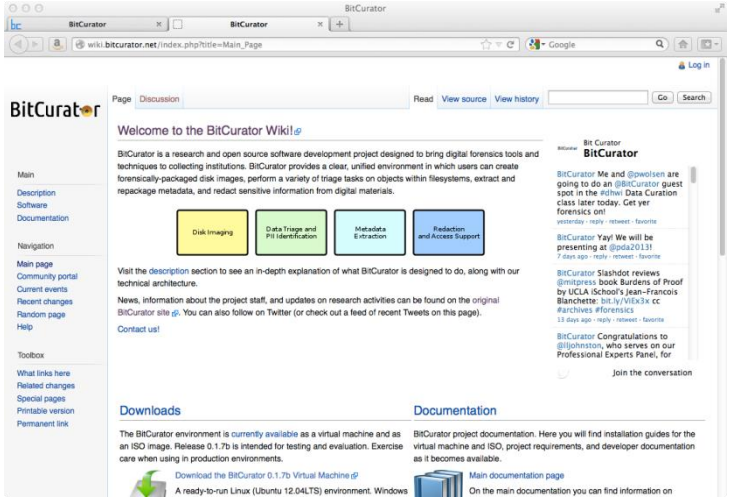
Becoming a Charter Member

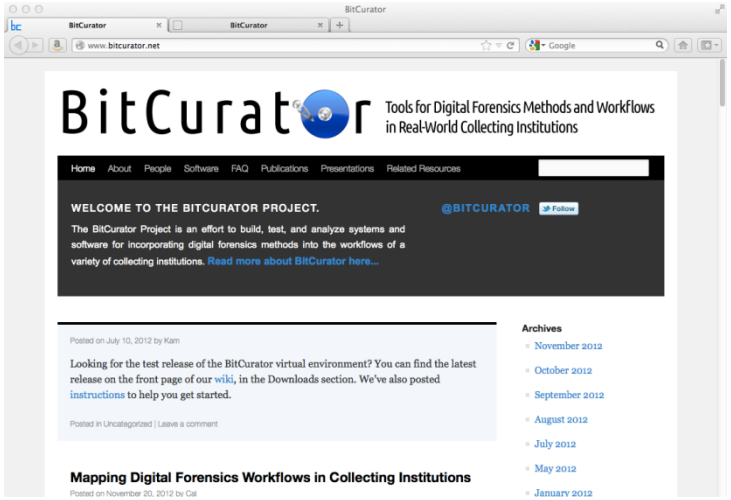
- Charter Membership drive June-December 2014
- Charter Members will play an early, active role in the shaping of the BitCurator Consortium's governance, ongoing development, and overall sustainability.
- Charter Membership is a one-time membership option, available only through December 31, 2014.

Membership Dues

- Dues for Charter Members in the first year: $5000 (US). After the first year of membership, dues will be the same as those of General Members.
- General Member dues: $2000 (US) per institution per year, for a three-year period with annual billing opportunities.

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