Archival Application of Digital Forensics Methods for Authenticity, Description and Access Provision

Cal Lee

School of Information and Library Science

University of North Carolina, Chapel Hill

International Council on Archives Congress

August 20-24, 2012

Brisbane, Australia

BitCurator

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

What is an archivist to do with things like this?

Source: Simson Garfinkel

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

Same Fundamental Archival Principles 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]

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.

But digital is different…

“No computation without representation”

Smith, Brian Cantwell. "Limits of Correctness in Computers." In *Computerization and Controversy: Value Conflicts and Social Choices*, edited by Rob Kling, 810-25. San Diego, CA: Academic Press, 1996. 815.

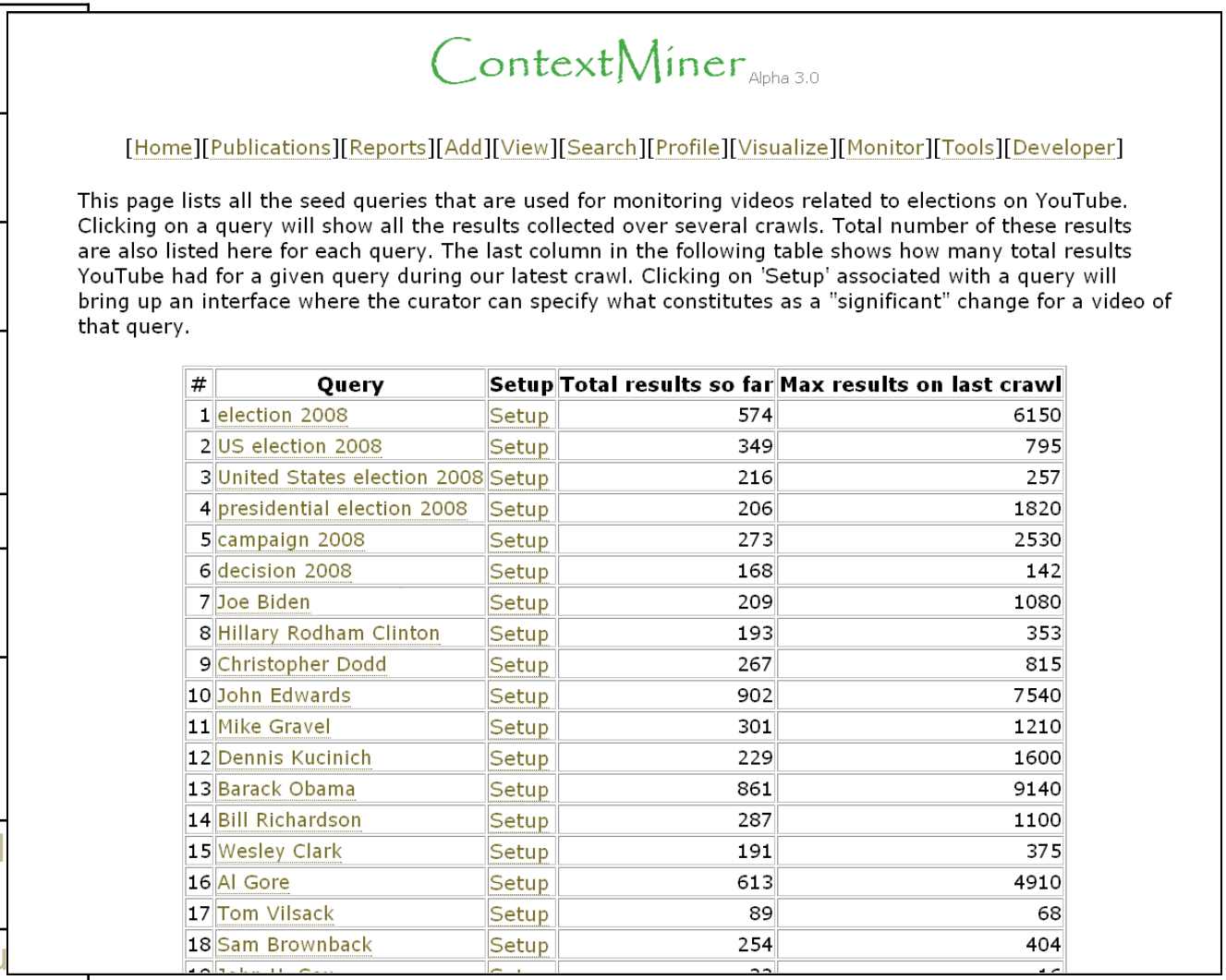
Digital Resources - Levels of Representation

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

Interaction Examples

Level

Aggregation of objects
Object or package
In-application rendering
File through filesystem
File as “raw” bitstream
Sub-file data structure
Bitstream through I/O equipment
Raw signal stream through I equipment
Bitstream on physical mediu



ContextMiner Alpha 3.0

[Home][Publications][Reports][Add][View][Search][Profile][Visualize][Monitor][Tools][Developer]

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

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

Interaction Examples

Level

- Aggregation of objects
- **Object or package**
- In-application rendering
- File through filesystem
- File as “raw” bitstream
- Sub-file data structure
- Bitstream through I/O equipment
- Raw signal stream throu equipment
- Bitstream on physical m

ContextMiner Alpha 3.0

[Home][Publications][Reports][Add][View][Search][Profile][Visualize][Monitor][Tools][Developer]

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

Query: *Rudy Giuliani*

I Got A Crush On.... Giuliani

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

Comedy

Crawling since 2007-07-19

Color coding for % changes

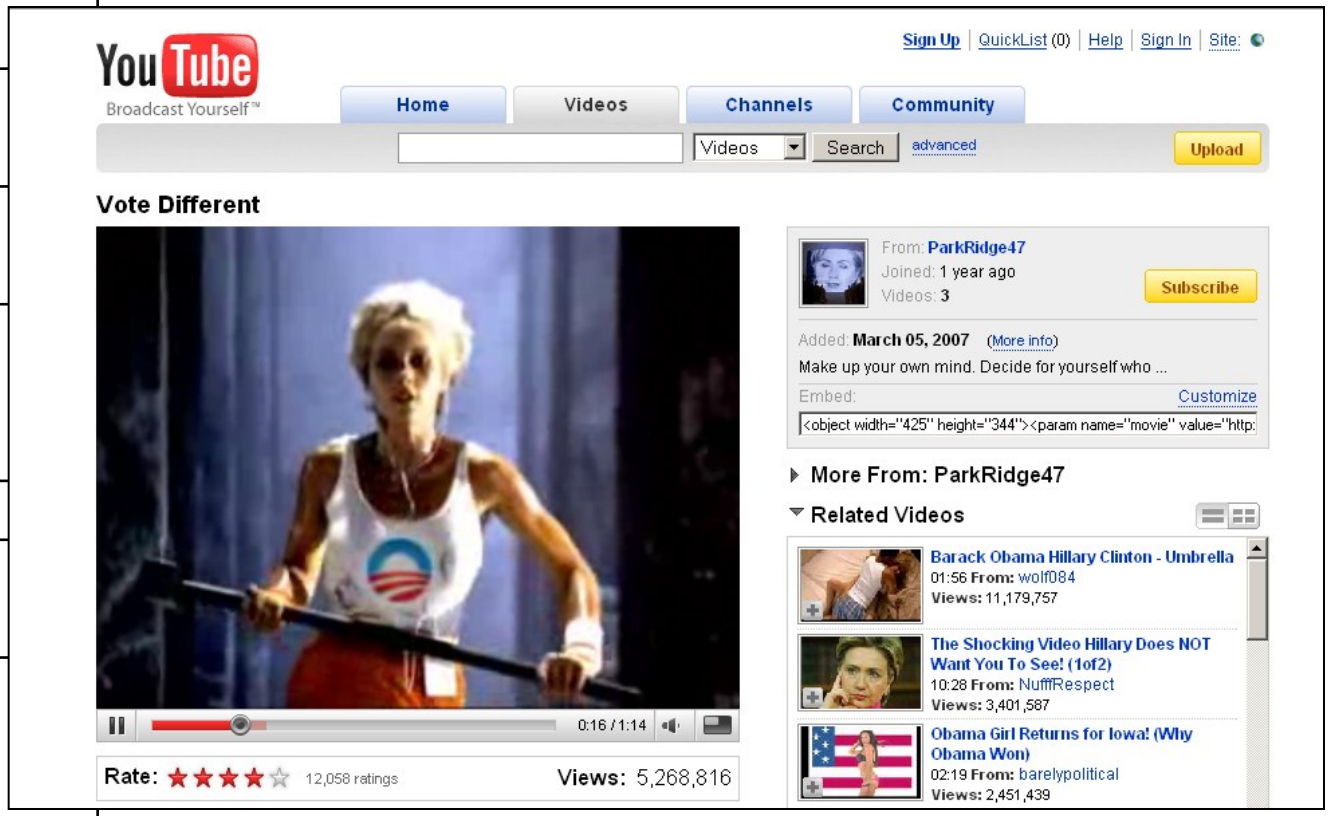
<	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	5.0	>

Crawl #	Crawl date	Rank	Views	Ratings	Avg Rating	Comments	Links	Favorited	Honors	Change
1	2007-07-31	5	27357	301	3.74	288	5	44	0	--
2	2007-08-01	5	27452	303	3.73	290	5	44	0	--
3	2007-08-02	5	27780	307	3.72	291	5	45	0	--
4	2007-08-03	5	28048	309	3.71	291	5	45	0	--
5	2007-08-04	2	28398	310	3.71	291	5	45	0	--
6	2007-08-05	2	28443	314	3.69	294	5	45	0	--
7	2007-08-06	3	28980	314	3.69	296	5	45	0	--
8	2007-08-07	3	29265	318	3.65	298	5	45	0	--
9	2007-08-08	3	29551	319	3.65	299	5	46	0	--
10	2007-08-09	3	30094	320	3.64	300	5	47	0	--
11	2007-08-10	3	30384	323	3.61	302	5	47	0	--
12	2007-08-10	5	30419	324	3.62	303	5	48	0	--
13	2007-08-11	3	30540	324	3.62	305	5	49	0	--
14	2007-08-12	3	30697	326	3.61	306	5	49	0	--
15	2007-08-13	3	30848	326	3.61	306	5	49	0	--
16	2007-08-14	3	31036	326	3.61	306	5	49	0	--
17	2007-08-15	2	31181	326	3.61	306	5	49	0	--
18	2007-08-16	2	31321	326	3.61	307	5	51	0	--
19	2007-08-17	2	31459	327	3.61	307	5	51	0	--
20	2007-08-18	2	31662	331	3.59	308	5	51	0	--
21	2007-08-19	2	31792	332	3.58	308	5	51	0	--
22	2007-08-20	2	31937	335	3.57	310	5	51	0	--
23	2007-08-21	2	32135	335	3.57	311	5	52	0	--

Interaction Examples

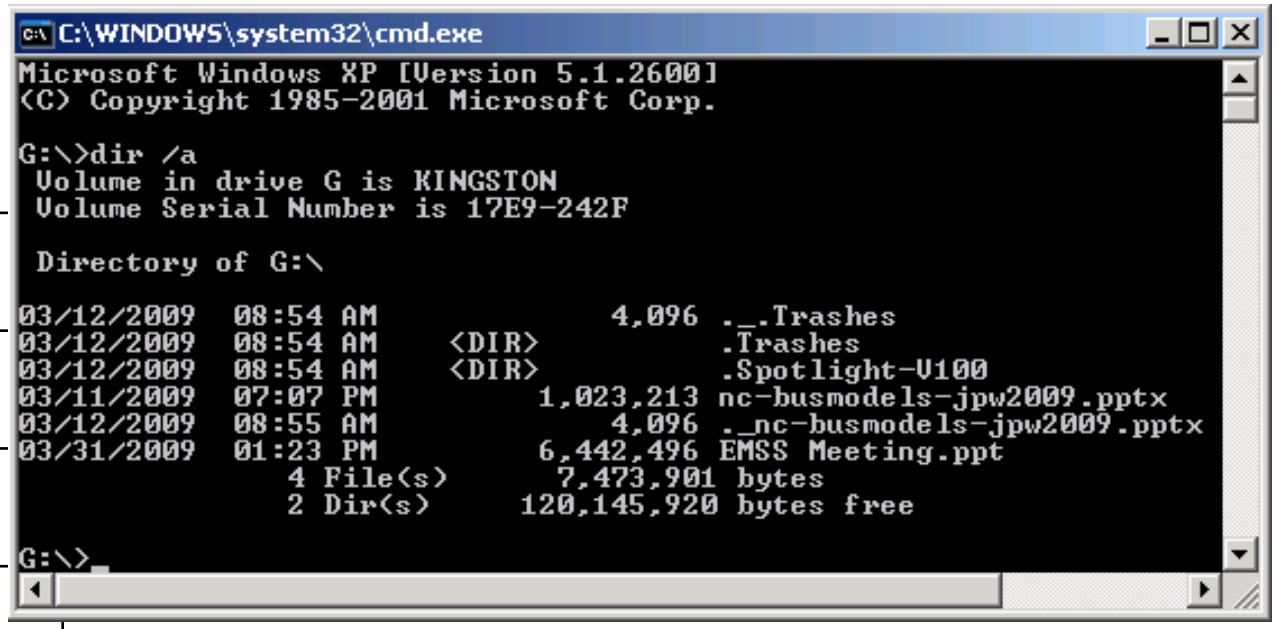
Level

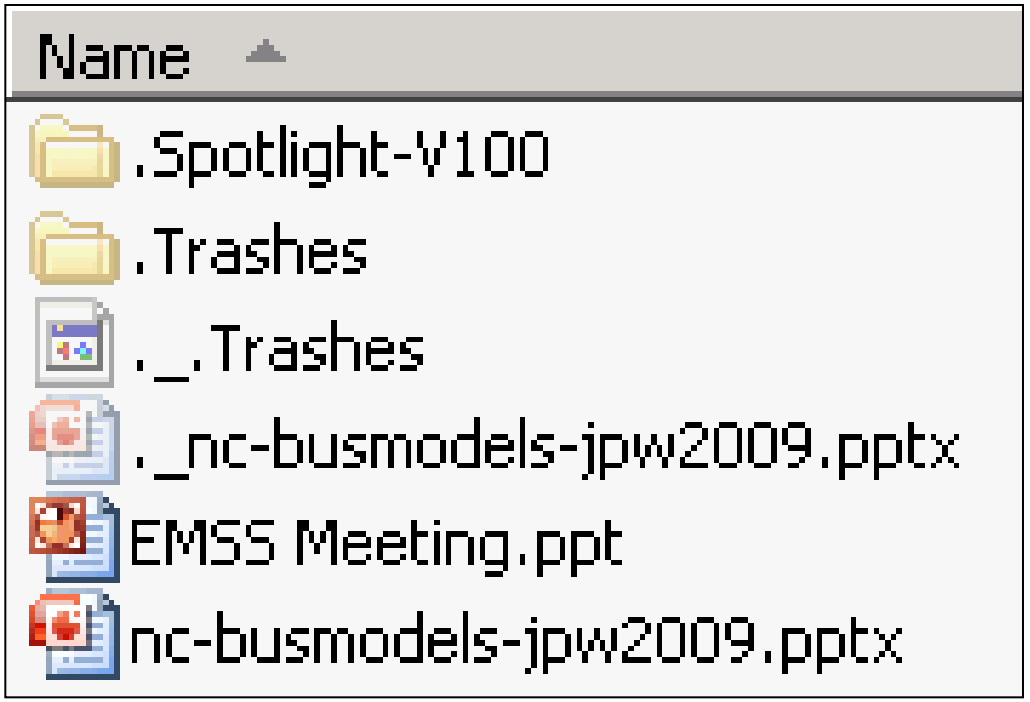
Aggregation of objects
Object or package
In-application rendering
File through filesystem
File as “raw” bitstream
Sub-file data structure
Bitstream through I/O equipment
Raw signal stream through I/O equipment
Bitstream on physical medium



Level

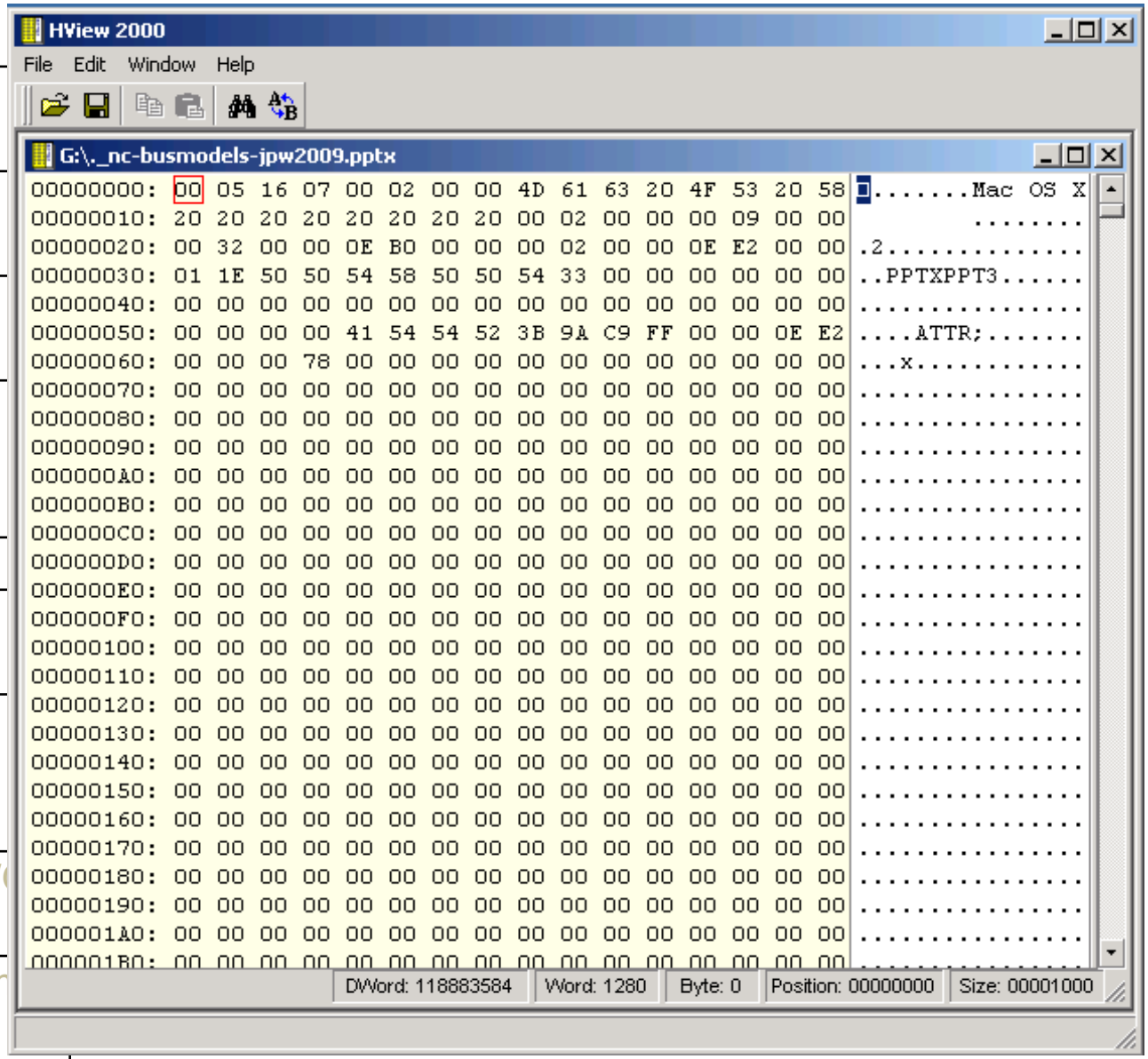
- Aggregation of objects
- Object or package
- In-application rendering
- **File through filesystem**
- File as “raw” bitstream
- Sub-file data structure
- Bitstream through I/O equipment
- Raw signal stream through I/O equipment
- Bitstream on physical medium





Interaction Examples

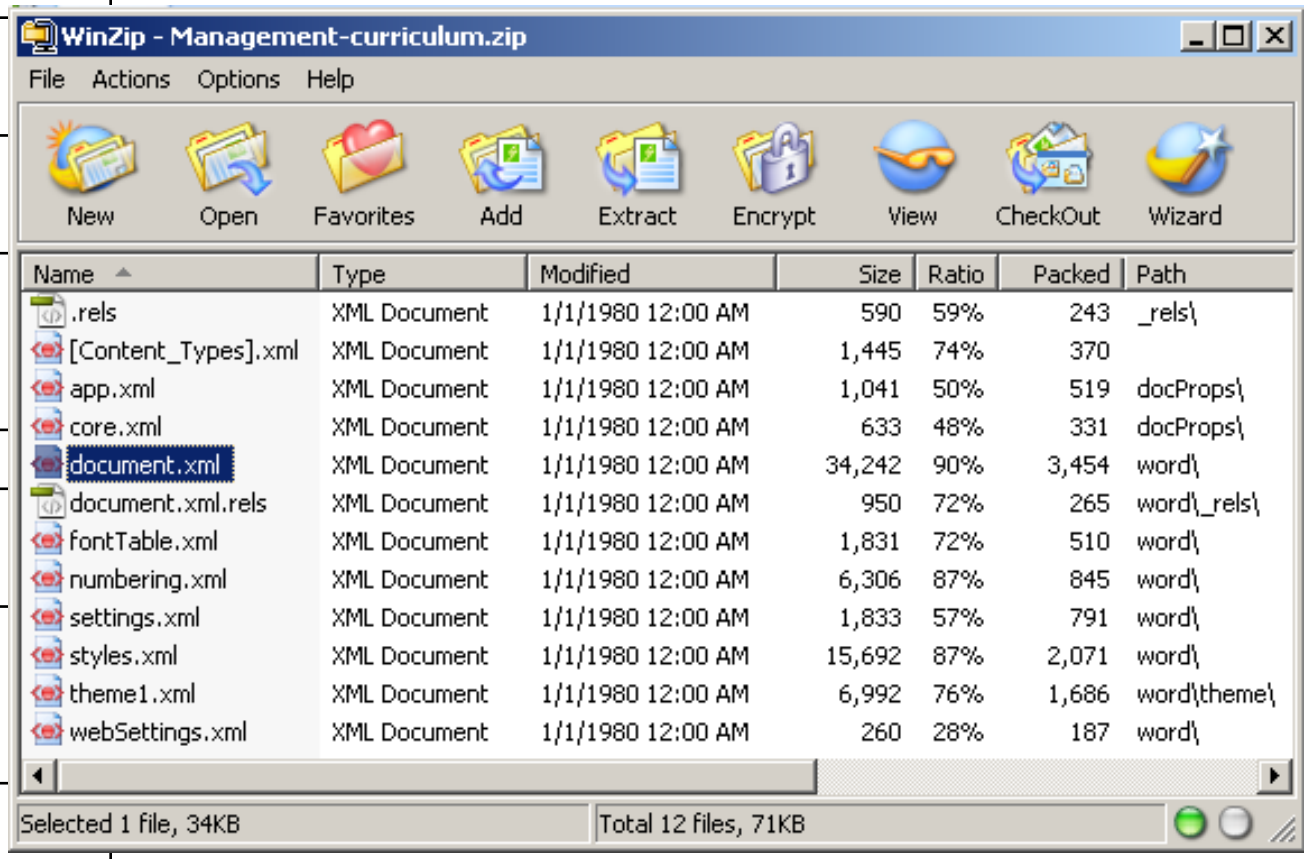
Level
Aggregation of objects
Object or package
In-application rendering
File through filesystem
File as “raw” bitstream
Sub-file data structure
Bitstream through I/O equipment
Raw signal stream through I/ equipment
Bitstream on physical mediur



Interaction Examples

Level

- Aggregation of objects
- Object or package
- In-application rendering
- File through filesystem
- File as “raw” bitstream
- **Sub-file data structure**
- Bitstream through I/O equipment
- Raw signal stream through equipment
- Bitstream on physical medium



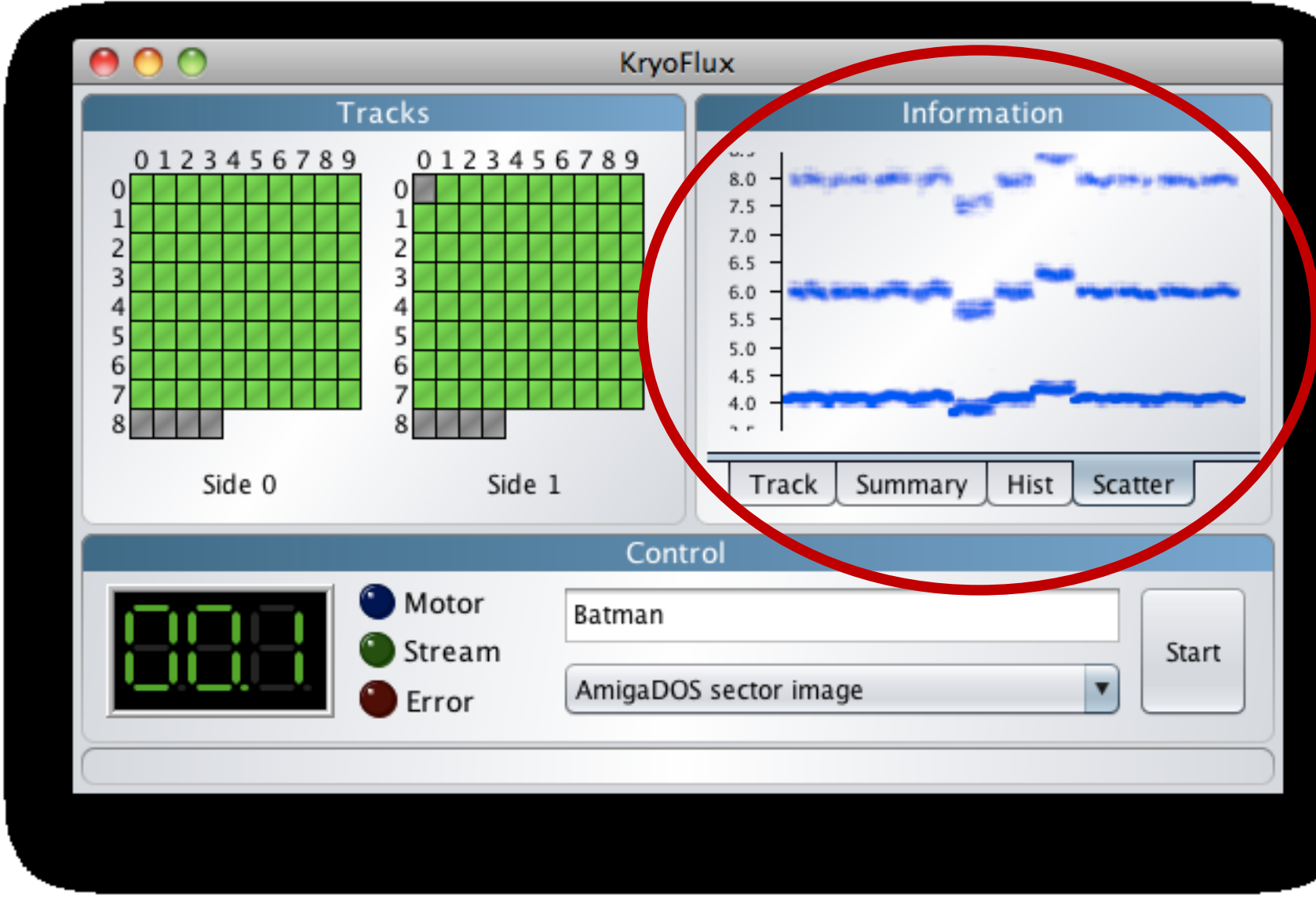
Interaction Examples

Level

- Aggregation of objects
- Object or package
- In-application rendering
- File through filesystem
- File as “raw” bitstream
- Sub-file data structure
- **Bitstream through I/O equipment**
- Raw signal stream through I/O equipment
- Bitstream on physical medium

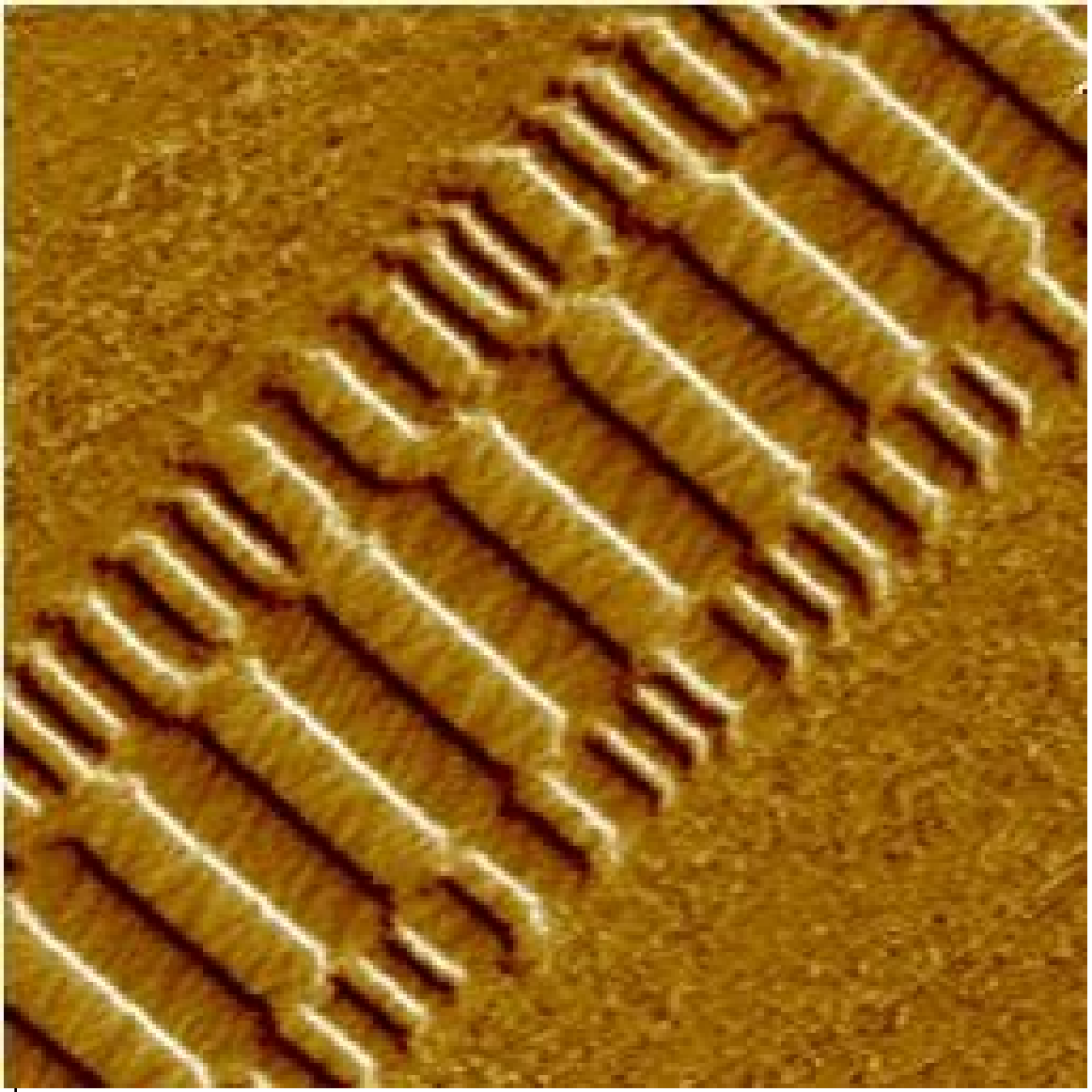
```
Segment                    arg    length    data
=======                ========   =======   ====
. . .
afflib_version                0         7   "3.3.3"
aff_file_type                 0         3   AFF
acquisition_commandline       0        36   aimage /dev/sda
                                             /mnt/charlie-002
acquisition_device            0         8   /dev/sda
sectorsize                 1024         0
pagesize                16777216         0
devicesectors                 2         8   = 9999864 (64-bit value)
acquisition_macaddr           0        18   00:0b:db:4f:6b:10.
acquisition_dmesg             0     27298   [    0.000000]
                                             Initializing cgro
image_gid                     0        16   7256 F895 DE4F E304
                                             233E 21C0 2347 CCC5
acquisition_date              0        20   2009-11-12 19:12:18.
md5                           0        16   0609 2DFE AA4F B183
                                             946F 95D8 AD84 519E
acquisition_seconds        1570         0   = 00:26:10 (hh:mm:ss)
imagesize                     2         8   = 10239860736 (64-bit value)
```

Level
Aggregat
Object or
In-applica
File throu
File as “r
Sub-file c
Bitstream equipmer
Raw sig I/O equipment
Bitstream on physical medium



Interaction Examples

Level
Aggregation of objects
Object or package
In-application rendering
File through filesystem
File as “raw” bitstream
Sub-file data structure
Bitstream through I/O equipment
Raw signal stream through I/O equipment
Bitstream on physical medium



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

Digital Forensics Principles Can Help Archivists to Fulfill their 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

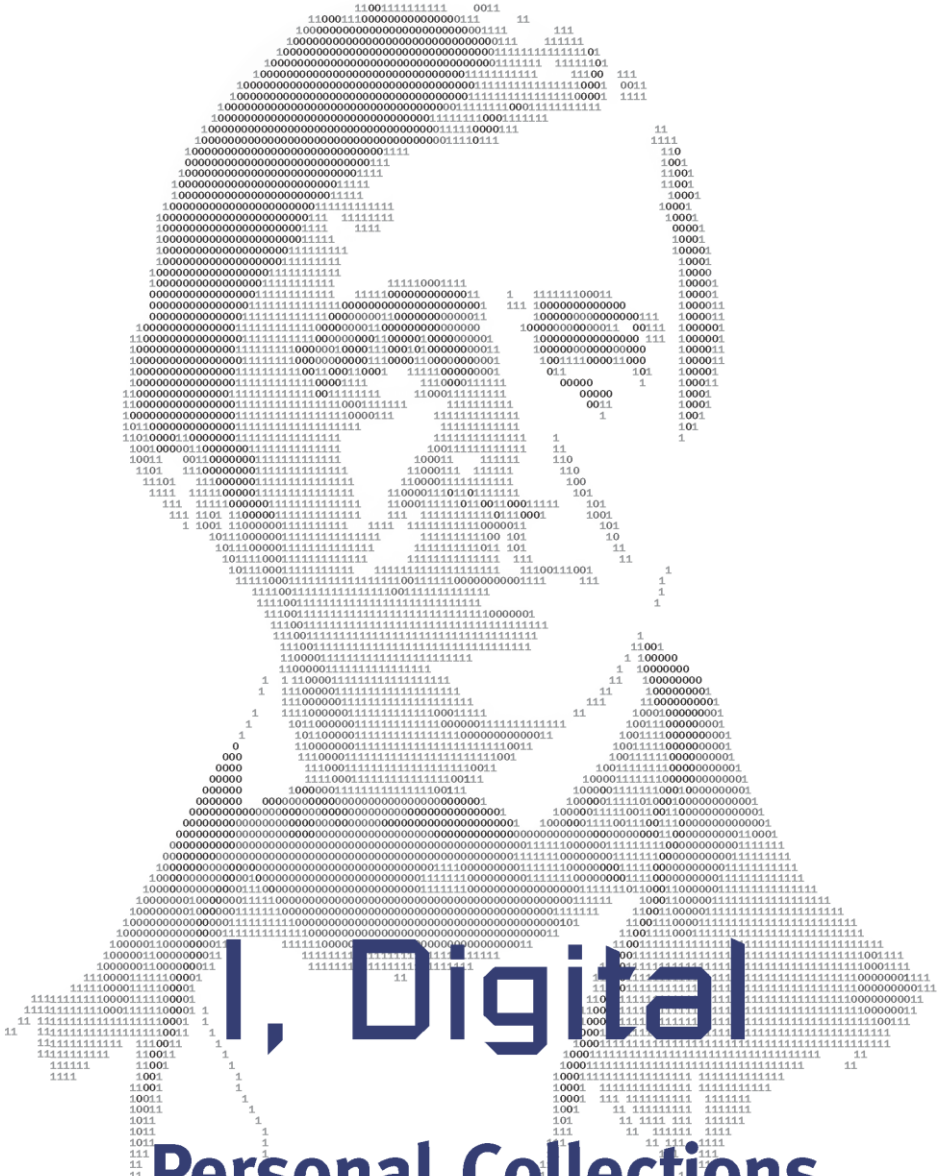
Previous Work*

- Some library/archives literature on recovering data from media
- Report by Ross and Gow (1999) on relevance of advances in data recovery and digital forensics to collecting institutions
- More recently, stream of literature related to use of forensic tools and methods for acquiring and managing digital collections
- Related projects:
  - Computer Forensics and Born-Digital Content in Cultural Heritage Collections
  - Born Digital Collections: An Inter-Institutional Model for Stewardship (AIMS)
  - Digital Records Forensics project

*See citations in full paper:
http://www.ica2012.com/files/data/Full%20papers%20upload/ica12Final00290.pdf

Renewed Energy and Attention around Personal Archives

- Arguably under-represented in first few decades of literature about electronic records
- Technical challenges of unruly acquisitions from individuals are channeling new energy into personal archives issues



I, Digital

Personal Collections in the Digital Era

Edited by Christopher A. Lee

What is Digital Forensics (aka Forensic Computing)?

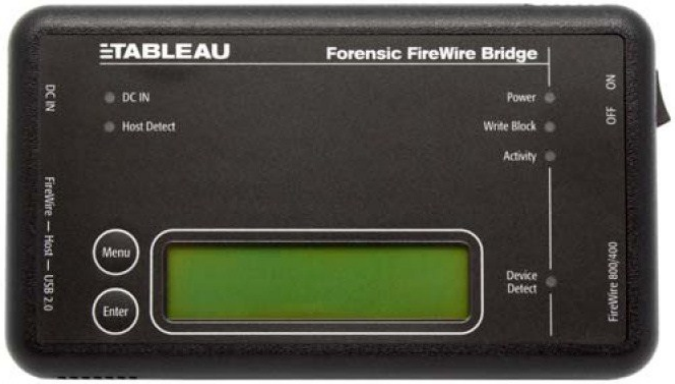
- "The process of identifying, preserving, analysing and presenting digital evidence in a manner that is legally acceptable."[1]
- "Involves multiple methods of
  - **Discovering digital data (computer system, mobiles)**
  - **Recovering deleted, encrypted, or damaged file information**
  - Monitoring live activity
  - Detecting violations of corporate policy"[2]

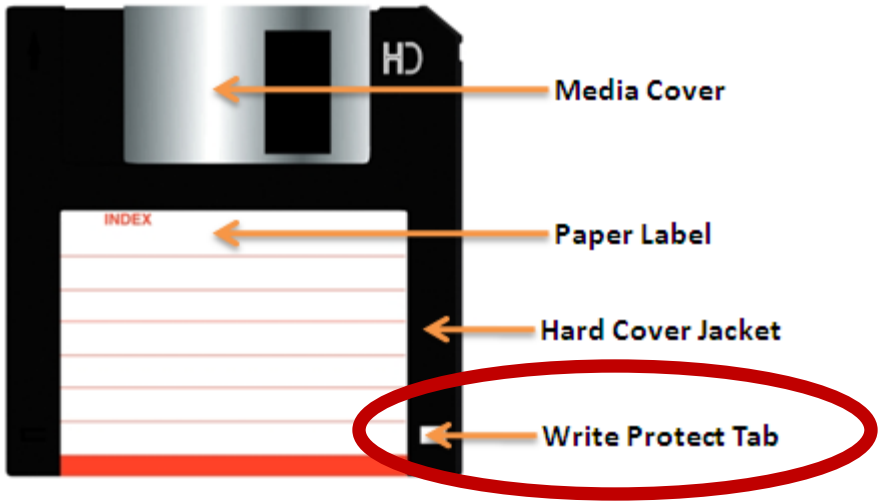
1. McKemmish, R. "What is Forensic Computing?" *Trends and Issues in Crime and Criminal Justice* 118 (1999).
2. Brad Glisson, Introduction to Computer Forensics & E-discovery, University of Glasgow, Week 1 Lecture, September 2008. (emphasis mine)

Guidelines for Evidence Collection & Archiving (RFC 3227) – Main Lessons

- “Such collection represents a considerable efforts on the part of the System Administrator.”
- “Keep detailed notes.”
- “Minimise changes to the data as you are collecting it.”
- “Do collection first and analysis later.”
- “Proceed from the volatile to the less volatile.”
- Computer evidence should be: admissible, authentic, complete, reliable, believable

Write Blocking – One-Way Streets for Data





www.techmint.info/2009/09/security-write-protecting-floppy-disks.html

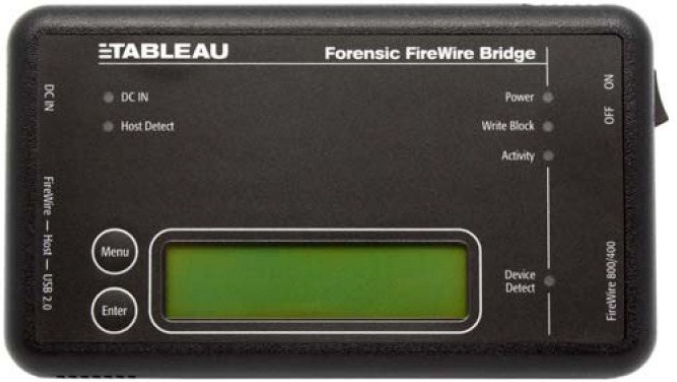
File System – An Essential Layer for Metadata

- 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)

Getting below the File System – Low-Level Copying

- Getting an “image” of a storage medium involves working at a level below the file system
- Image is a copy of all of the storage sectors from the drive, rather than just copying the files
- Can get at file attributes and deleted files not visible through higher-level copy operations

Digital Forensics Tools – Hardware and Software



AFFLIB

Open Source Computer Forensics Software

Digital forensics tools are designed primarily to be used in places like this:

El Paso County Sheriff's Office (Colorado)

http://shr.elpasoco.com/Law+Enforcement+Bureau/Investigations+Division/Computer+Crime+Lab.htm

But they’re also be used in places like this:

Stanford University Libraries and Academic Information Resources (SULAIR)

British Library, London

BitCurator Project

- Funded by Andrew W. Mellon Foundation - October 1, 2011 – September 30, 2013
- Partners: SILS and Maryland Institute for Technology in the Humanities (MITH)
- Core Team:
  - Cal Lee, PI
  - Matt Kirschenbaum, Co-PI
  - Kam Woods, Technical Led
  - Alex Chassonoff, Project Manager (UNC), Sunitha Misra, GA (UNC), Porter Olsen, GA (MITH)

Professional Experts Panel	Development Advisory Group
• Bradley Daigle, University of Virginia Library • Erika Farr, Emory University • Jeremy Leighton John, British Library • Leslie Johnston, Library of Congress • Courtney Mumma, Artefactual Systems • 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	• Geoffrey Brown, Indiana University • Barbara Guttman, National Institute of Standards and Technology • Jerome McDonough, University of Illinois • Mark Matienzo, Yale University • David Pearson, National Library of Australia • Doug Reside, New York Public Library • Seth Shaw, University Archives, Duke University • William Underwood, Georgia Tech • Peter Van Garderen, Artefactual Systems

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

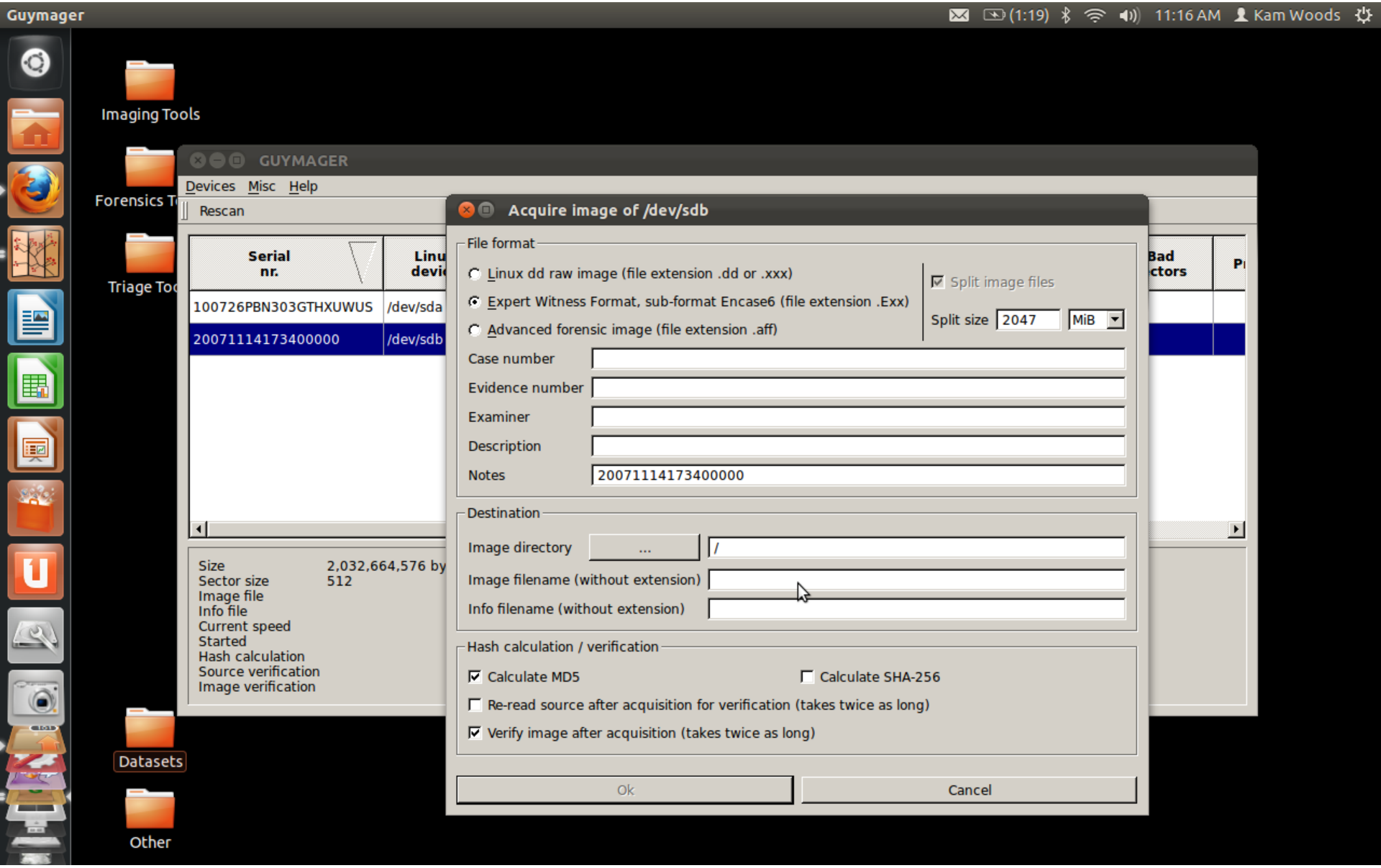
BitCurator Environment*

- Bundles, integrates and extends functionality of open source software: 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

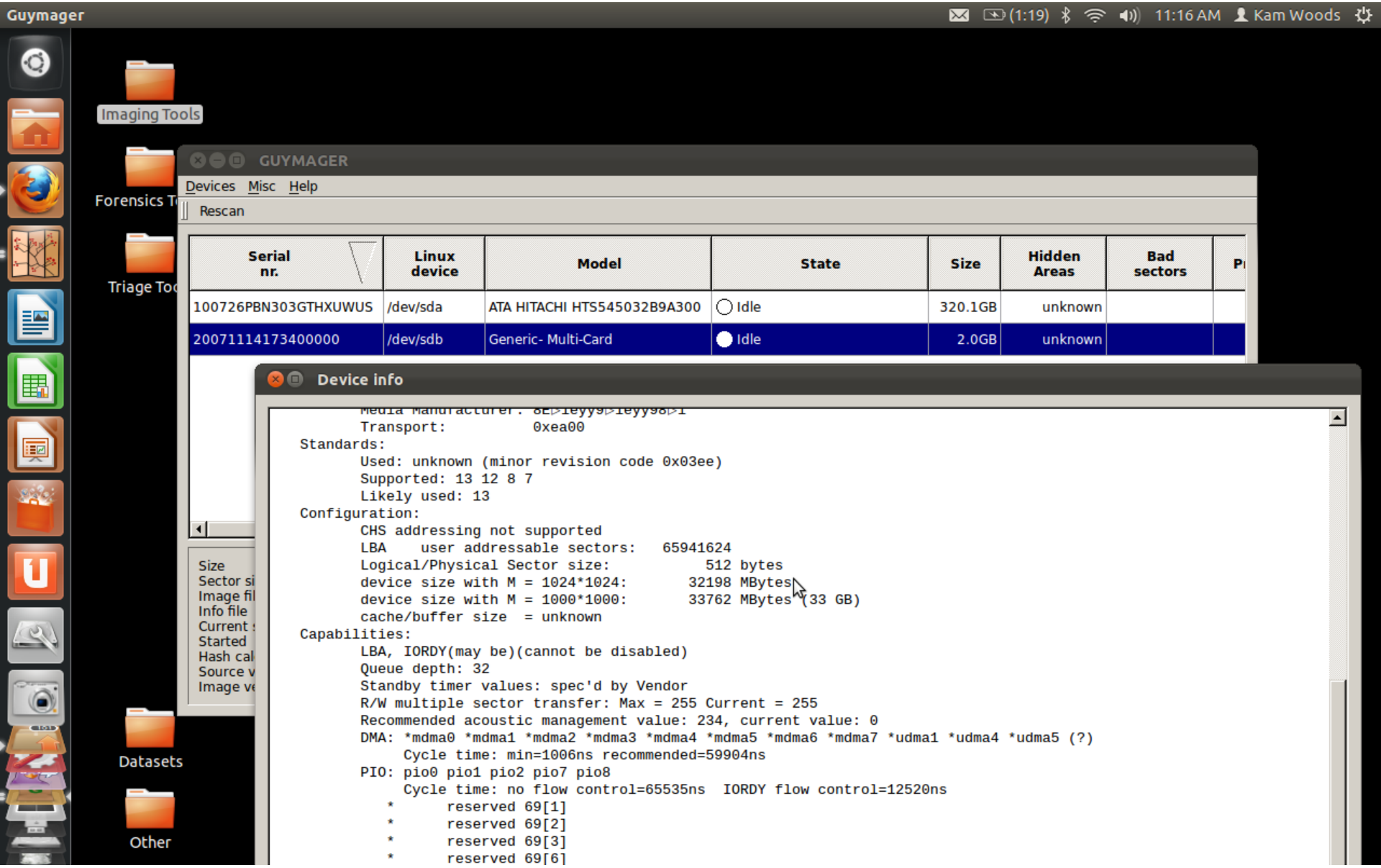
*To read about and download the environment, see: http://wiki.bitcurator.net/


Update Manager
(1:54)
10:59 AM
Kam Woods
Imaging Tools
Forensics Tools
Triage Tools
BitCurator
Datasets
Other

Main Acquisition Interface for Guymager



Guymager Showing Technical Metadata about an SD Card (Right Click)



Exporting Filesystem Metadata - Output from fiwalk (XML)*

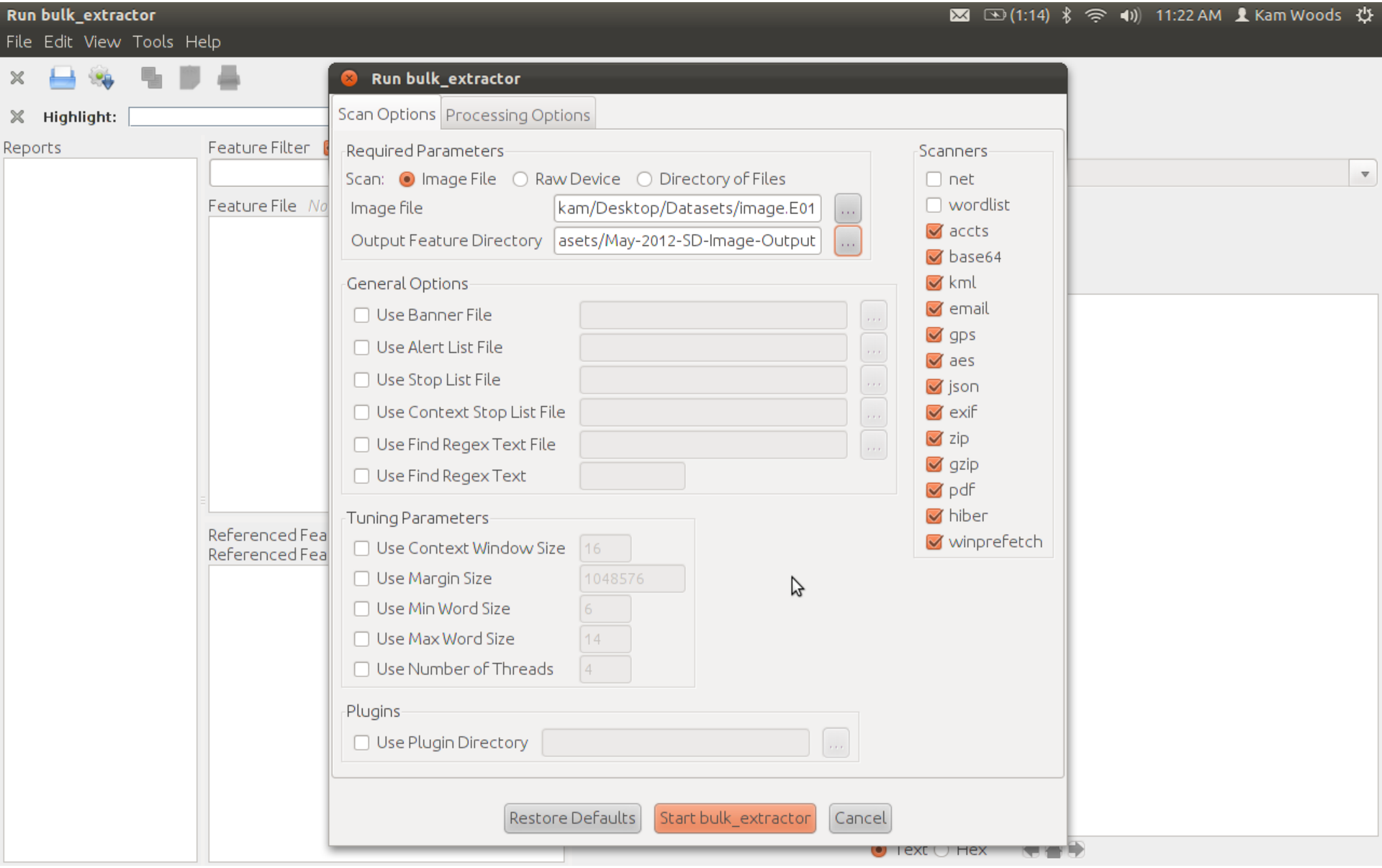
```
<fileobject>
      <filename>Documents and Settings/All Users/Documents/
                My Pictures/Sample Pictures/Blue hills.jpg
      </filename>
      ...
      <filesize>28521</filesize>
      <alloc>1</alloc>
      <used>1</used>
      <inode>6245</inode>
      ...
      <uid>0</uid>
      <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>
```

*fiwalk created by Simson Garfinkel

Identifying “Features” of Interest in Disk Images

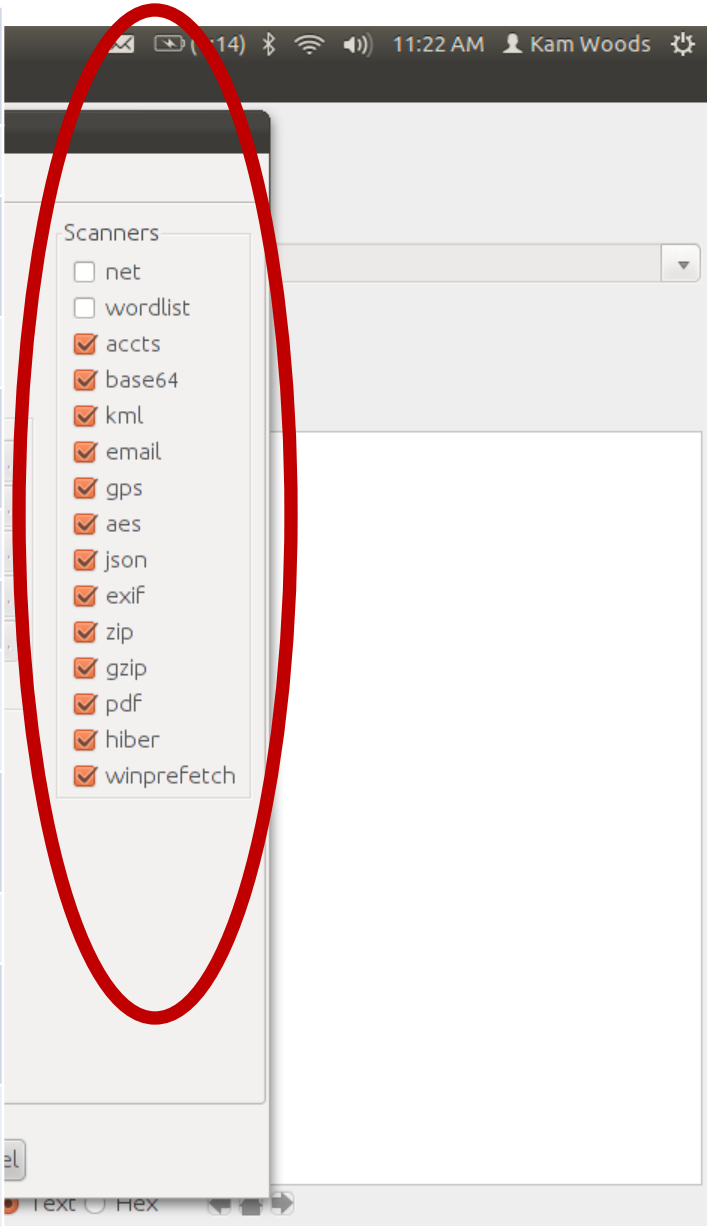
Bulk Extractor
(Created by Simson Garfinkel)

Bulk Extractor Scanning Options

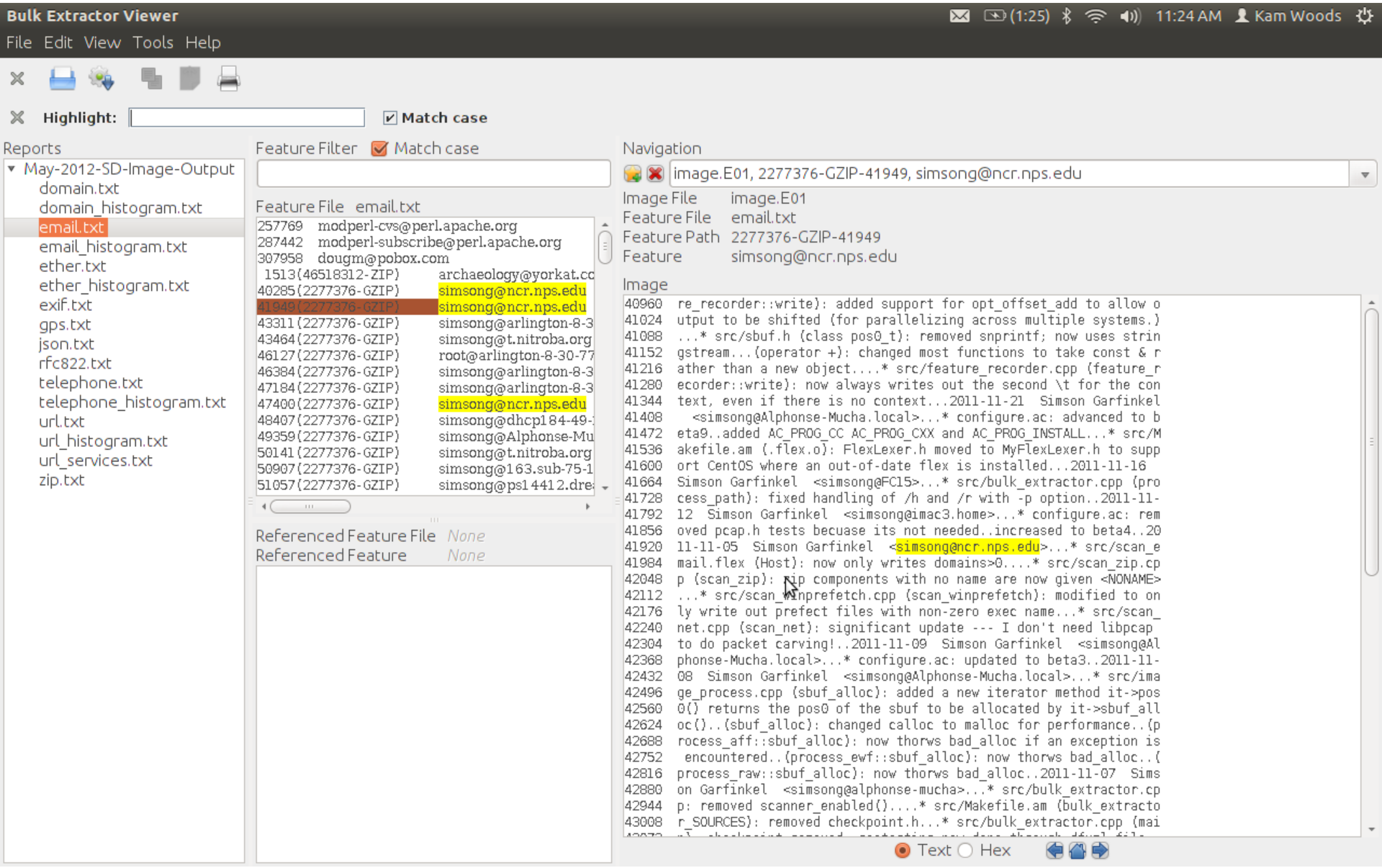


Options

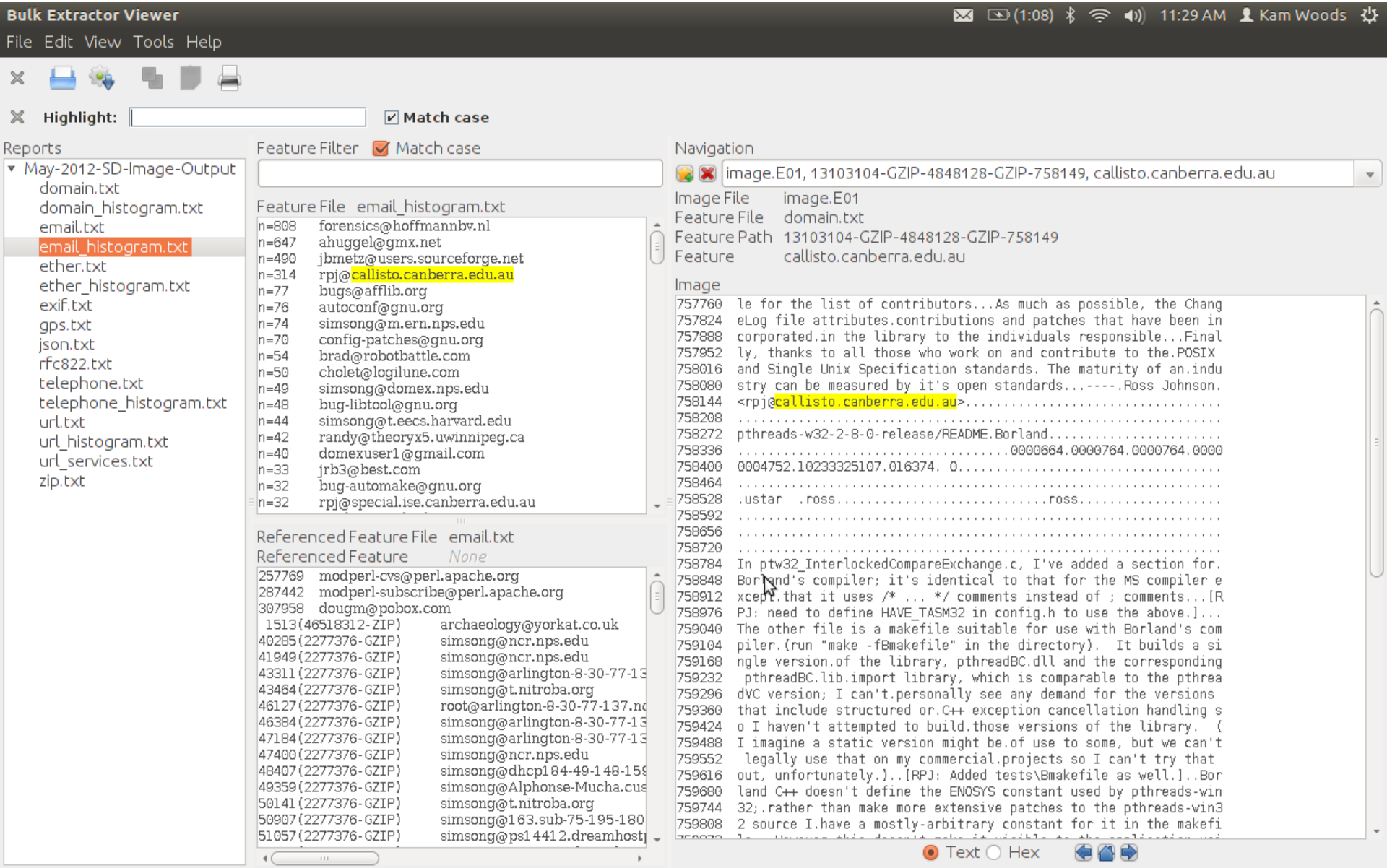
Scanner	Description
scan-accts	Looks for phone numbers, credit card numbers, etc
scan_base64	Decodes BASE64 text
scan_kml	Detects KML (Keyhole Markup Language) files – used to identify geographic locations
scan_gps	Detects XML from Garmin GPS devices
scan_aes	Detects in-memory AES (Advanced Encryption Standard) keys from the key schedules
scan_json	Detects JavaScript Object Notation files
scan_exif	Detects EXIF structures from JPEG files
scan_zip	Detects and decompresses ZIP files and zlib streams
scan_gzip	Detects and decompresses GZIP files and gzip streams
scan_pdf	Extracts text from some kinds of PDF files
scan_hiber	Detects and decompresses Windows hibernation file fragments
scan_winprefetch	Detects and extracts fields from windows prefetch fields from Windows prefetch files and file fragments



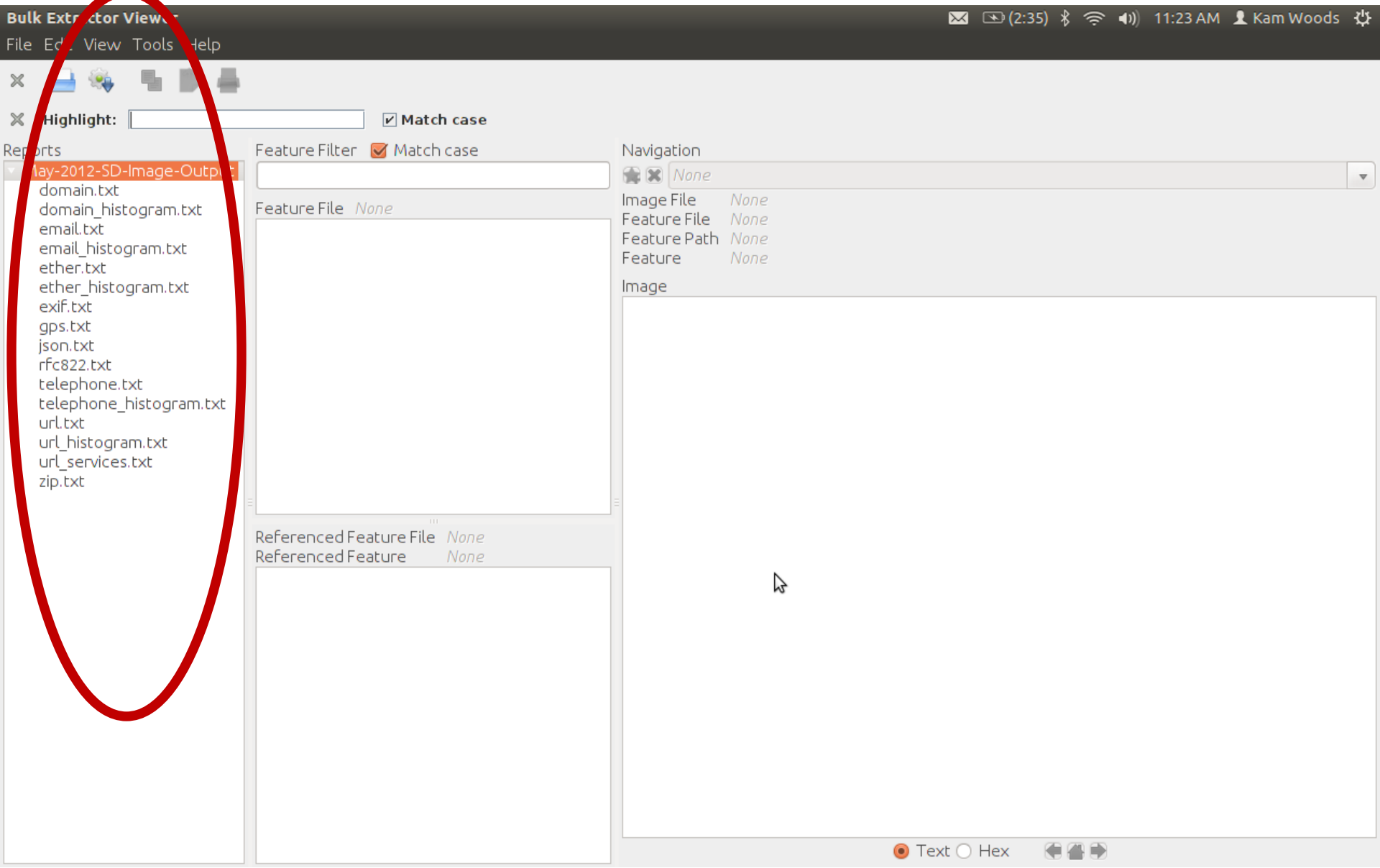
Email Addresses Identified - With Repeats Highlighted



Histogram of Email Addresses*



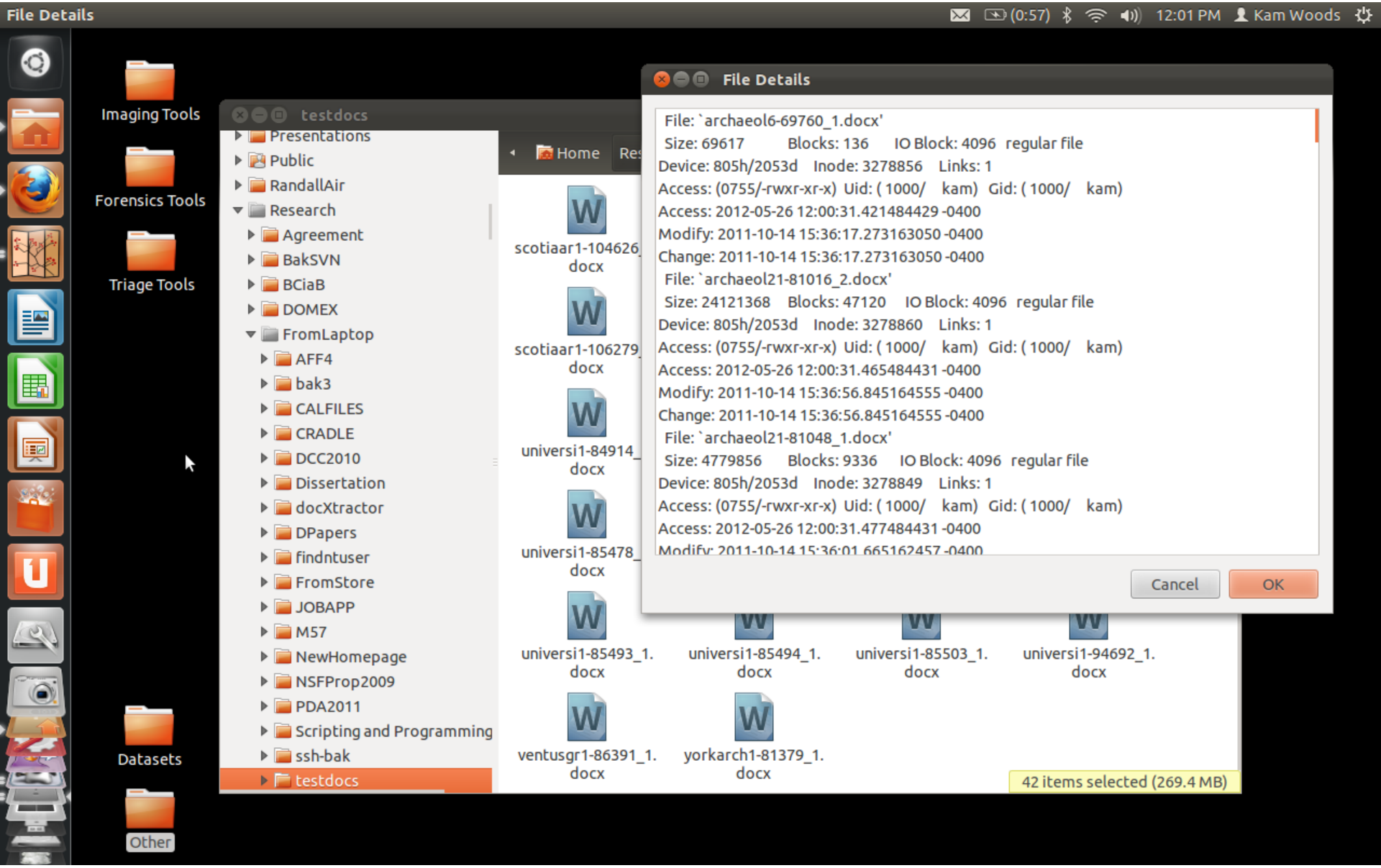
Bulk Extractor Report Options



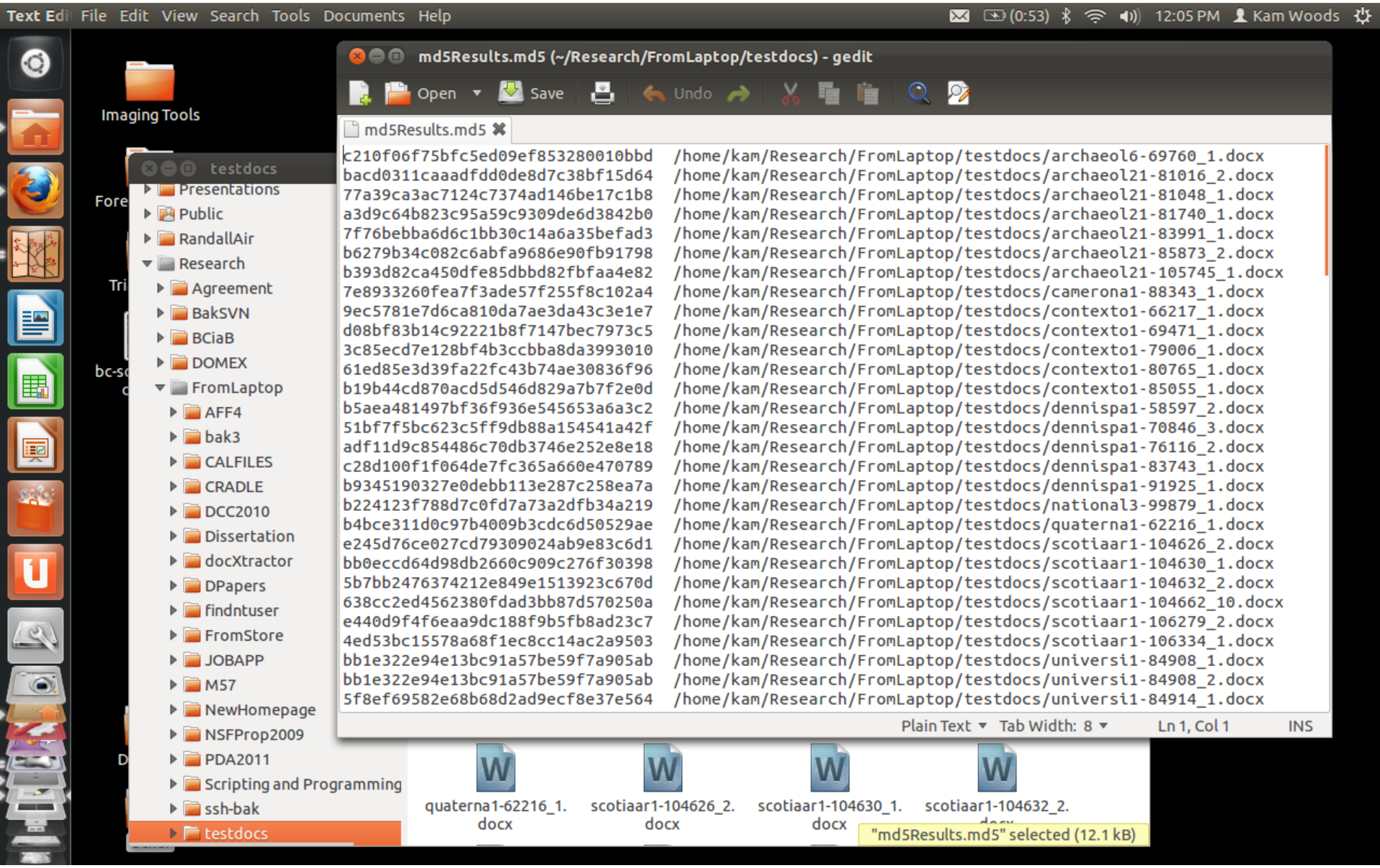
Nautilus Scripts

- 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

File Details for Word Documents in a Directory (Nautilus Script)



MD5 Hashes of Files (Nautilus Script)



Conclusions: Implied Changes with the Archival Profession

- Professional vocabulary evolving to include terms such as disk image, hex viewer, cryptographic hash, and filesystem
- Gaining access to new professional communities and guidance
- Use of tools designed to treat data at a low level – as raw bitstreams off media – rather than at the file level
- Potential to shift “center of gravity” about electronic records from design of institutional recordkeeping systems toward acquisition and management of records from a more diverse and unpredictable set of sources

Thank you!

http://bitcurator.net

http://wiki.bitcurator.net

Twitter: @bitcurator