Introduction to BitCurator: Using BitCurator to Support Digital Curation

Christopher (Cal) Lee

School of Information and Library Science

University of North Carolina at Chapel Hill

January 22, 2020

LYRASIS Webinar

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

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

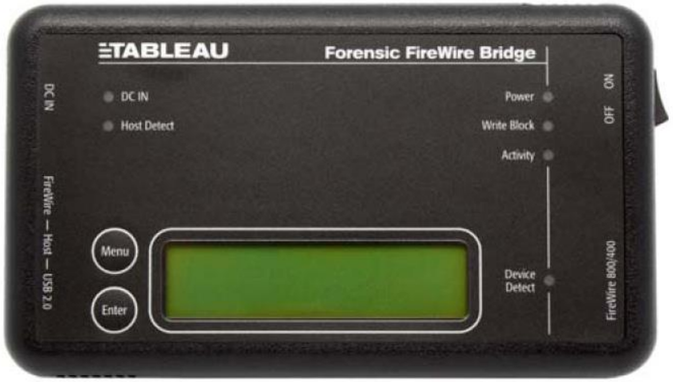
Chain of Custody

- “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 you might need some of this stuff:

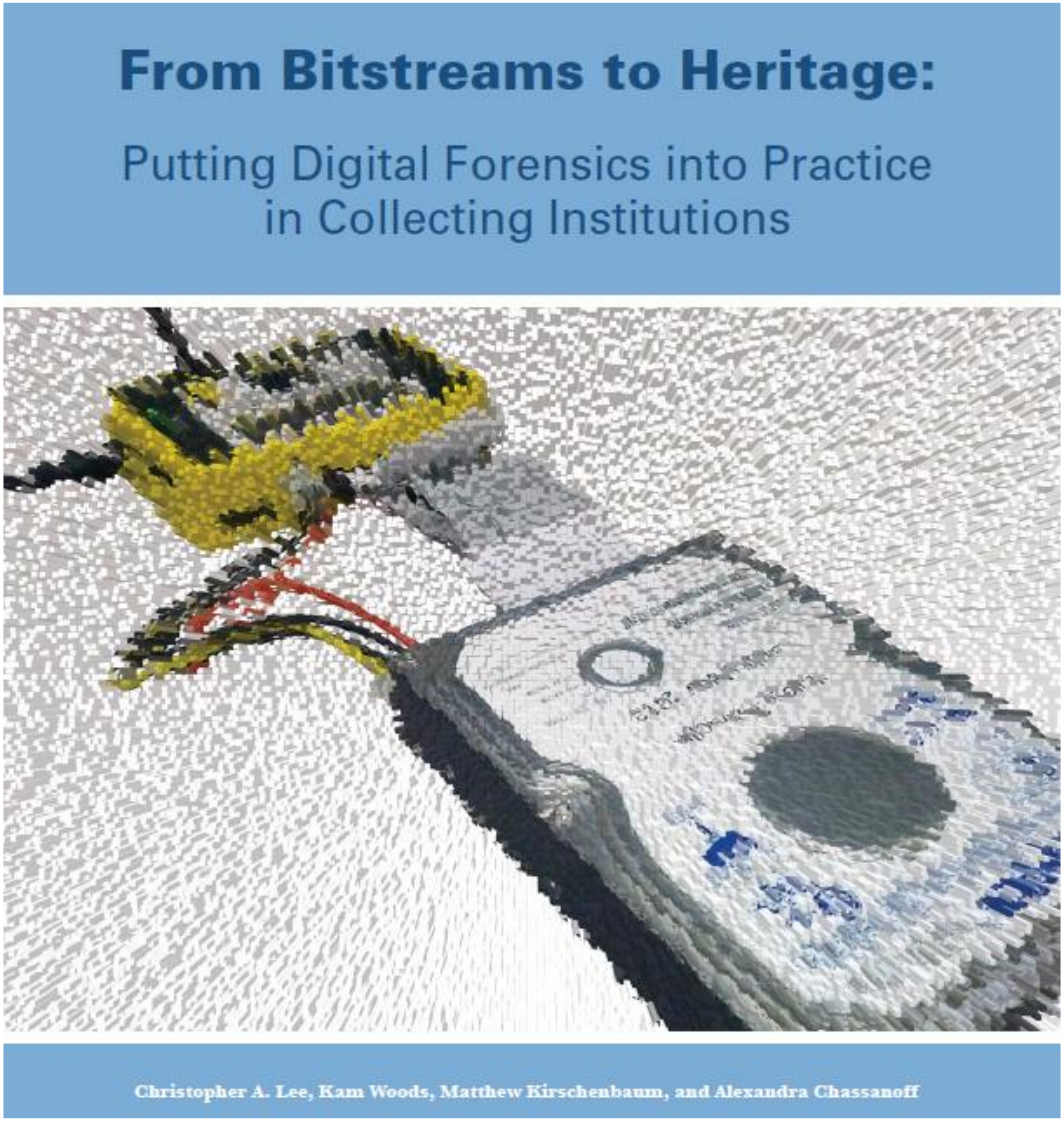


Motivation

- Archivists 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

Digital Forensics 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



From Bitstreams to Heritage:

Putting Digital Forensics into Practice in Collecting Institutions

Christopher A. Lee, Kam Woods, Matthew Kirschenbaum, and Alexandra Chassanoff

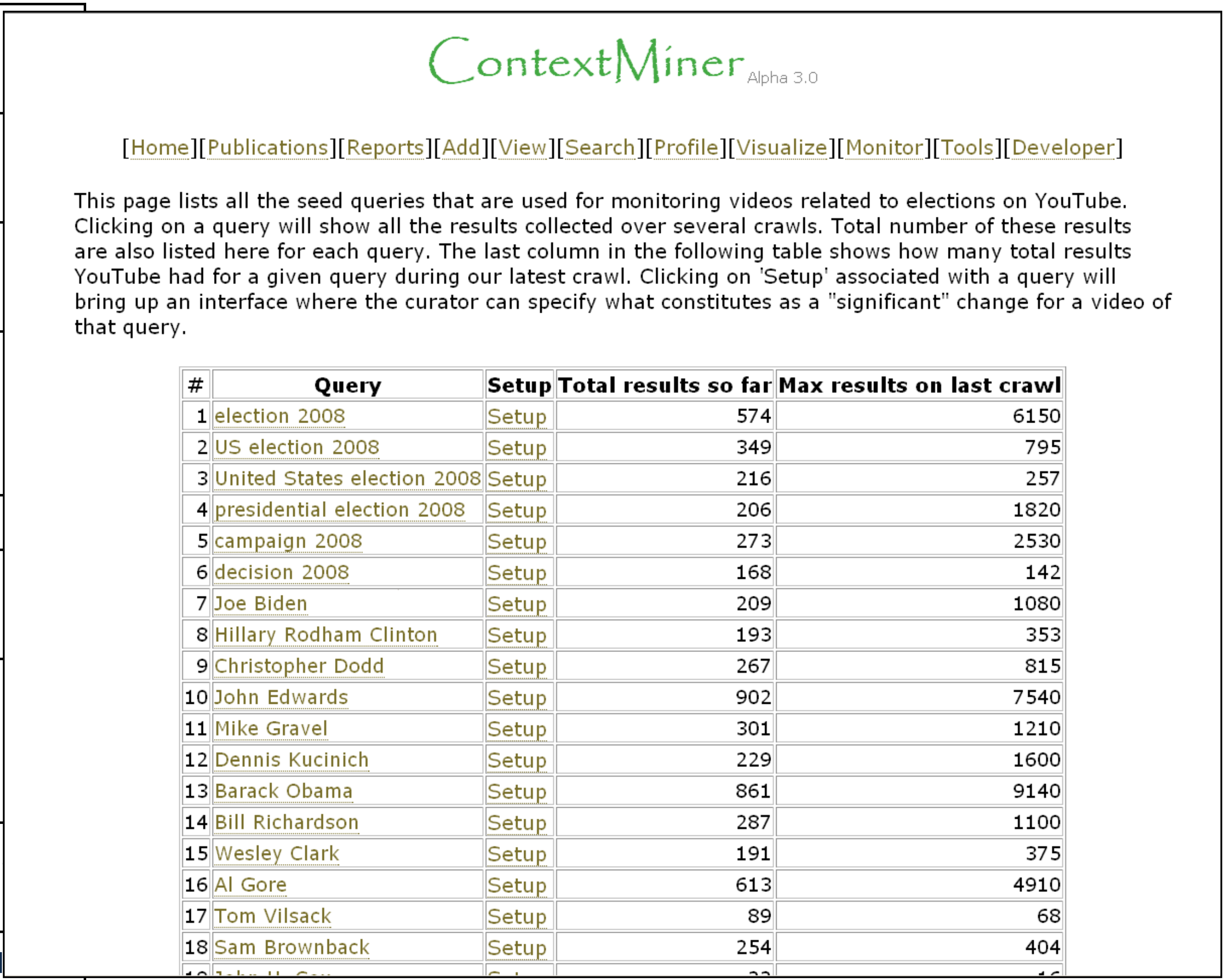
http://www.bitcurator.net/docs/bitstreams-to-heritage.pdf

Digital 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

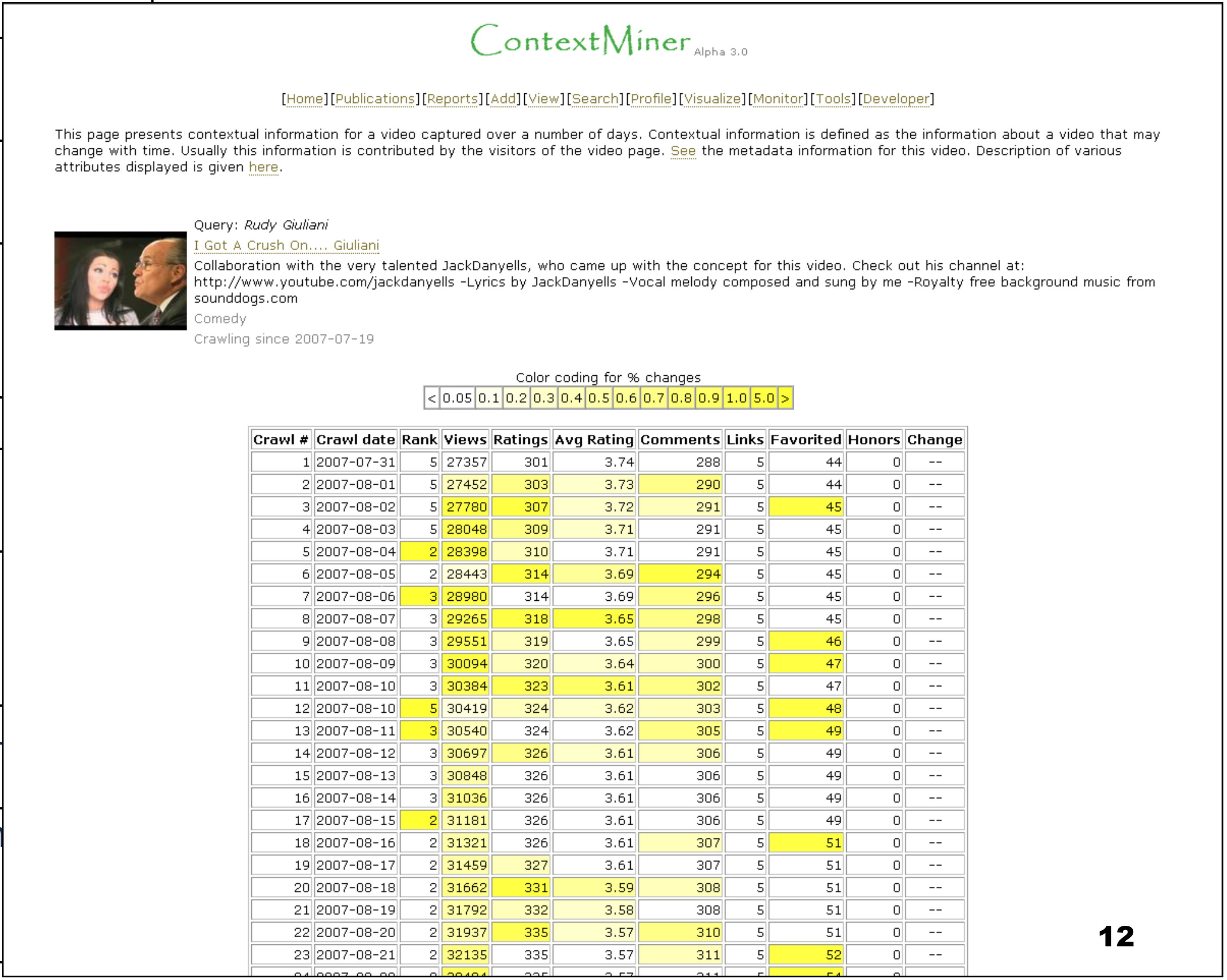
[Home][Publications][Reports][Add][View][Search][Profile][Visualize][Monitor][Tools][Developer]

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

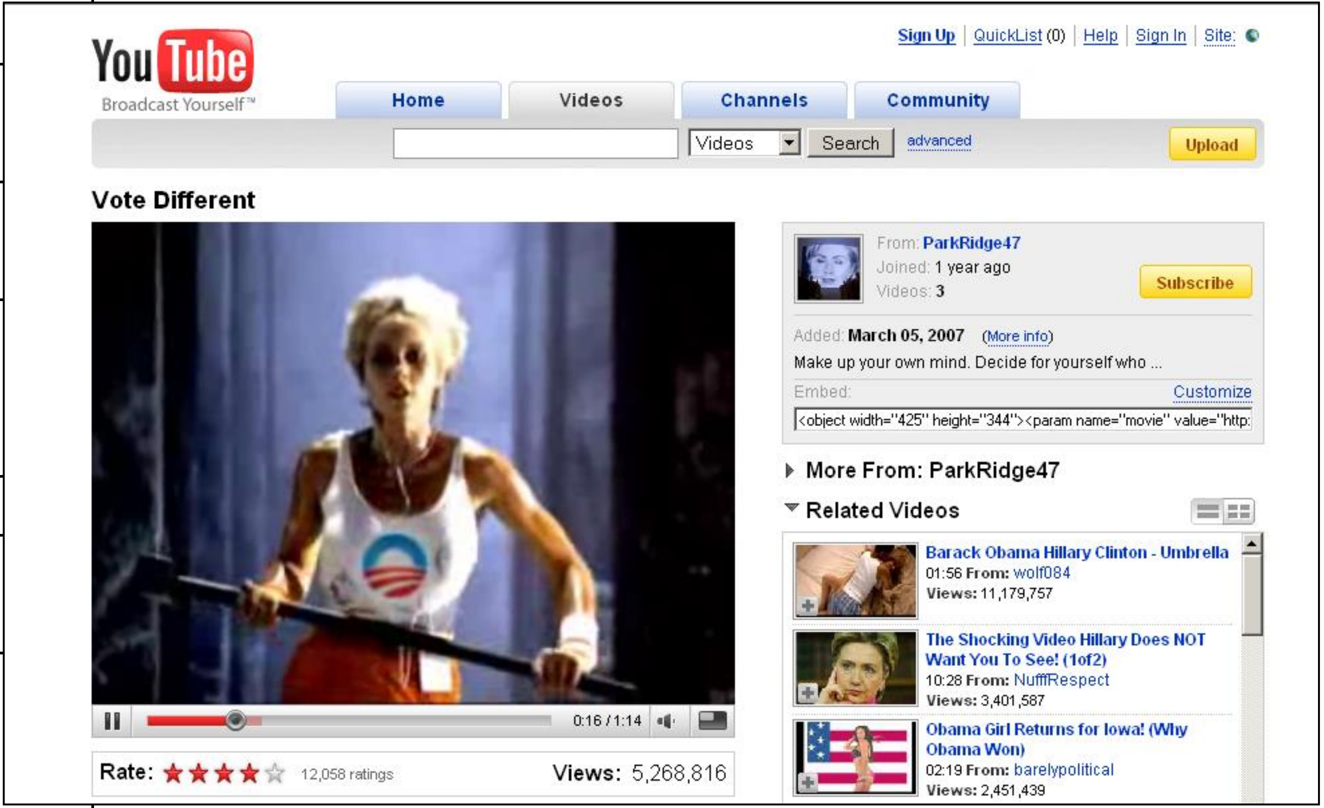
Color coding for % changes

<	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

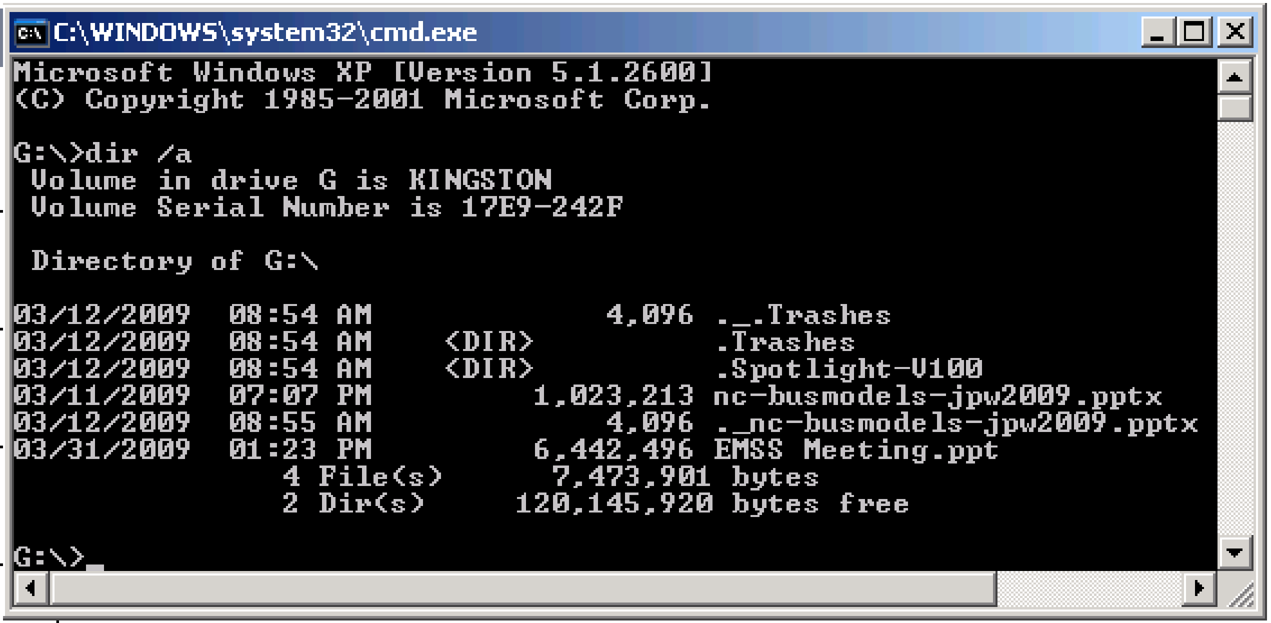
Level
Aggregation of objects
Object or package
In-application rendering
File through filesystem
File as “raw” bitstream
Sub-file data structure
Bitstream through I/O equipment
Raw signal stream through I/O equipment
Bitstream on physical medium



Make up your own mind. Decide for yourself who ...

<object width="425" height="344"><param name="movie" value="http:

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



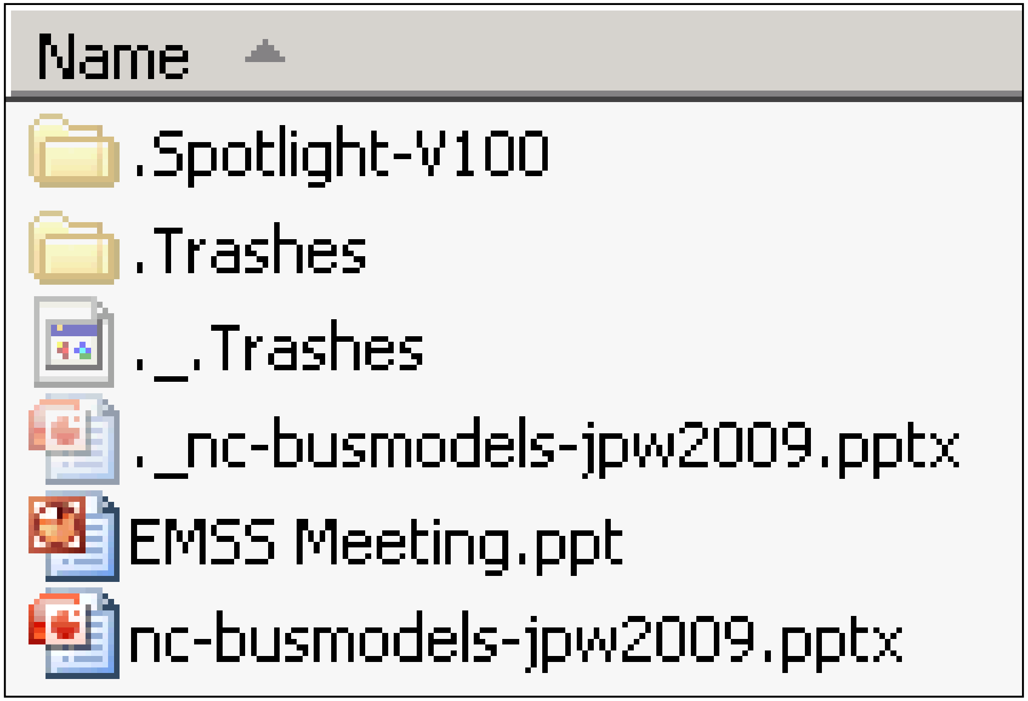
```
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.

G:\>dir /a
 Volume in drive G is KINGSTON
 Volume Serial Number is 17E9-242F

 Directory of G:\

03/12/2009  08:54 AM             4,096 ._.Trashes
03/12/2009  08:54 AM    <DIR>          .Trashes
03/12/2009  08:54 AM    <DIR>          .Spotlight-V100
03/11/2009  07:07 PM         1,023,213 nc-busmodels-jpw2009.pptx
03/12/2009  08:55 AM             4,096 ._nc-busmodels-jpw2009.pptx
03/31/2009  01:23 PM         6,442,496 EMSS Meeting.ppt
               4 File(s)      7,473,901 bytes
               2 Dir(s)     120,145,920 bytes free

G:\>_
```



Name

.Spotlight-V100

.Trashes

._.Trashes

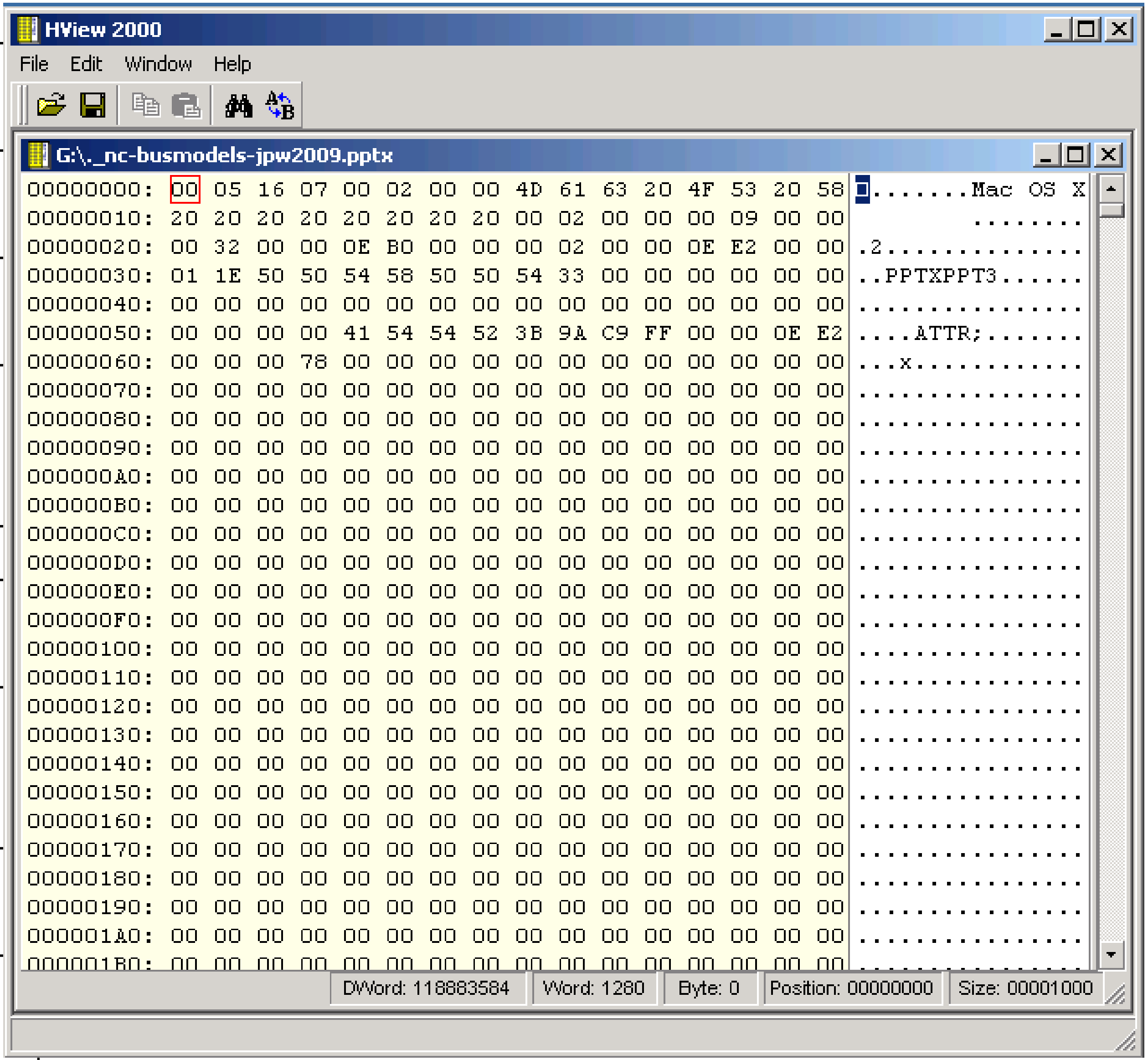
._nc-busmodels-jpw2009.pptx

EMSS Meeting.ppt

nc-busmodels-jpw2009.pptx

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



G:\._nc-busmodels-jpw2009.pptx

```
00000000: 00 05 16 07 00 02 00 00 4D 61 63 20 4F 53 20 58 ........Mac OS X
00000010: 20 20 20 20 20 20 20 20 00 02 00 00 00 09 00 00         ........
00000020: 00 32 00 00 0E B0 00 00 00 02 00 00 0E E2 00 00 .2..............
00000030: 01 1E 50 50 54 58 50 50 54 33 00 00 00 00 00 00 ..PPTXPPT3......
00000040: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
00000050: 00 00 00 00 41 54 54 52 3B 9A C9 FF 00 00 0E E2 ....ATTR;.......
00000060: 00 00 00 78 00 00 00 00 00 00 00 00 00 00 00 00 ...x............
00000070: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
00000080: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
00000090: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
000000A0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
000000B0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
000000C0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
000000D0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
000000E0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
000000F0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
00000100: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
00000110: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
00000120: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
00000130: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
00000140: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
00000150: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
00000160: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
00000170: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
00000180: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
00000190: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
000001A0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
000001B0: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 ................
```

DWord: 118883584 | Word: 1280 | Byte: 0 | Position: 00000000 | Size: 00001000

Interaction Examples

Level

Aggregation of objects

Object or package

In-application rendering

File through filesystem

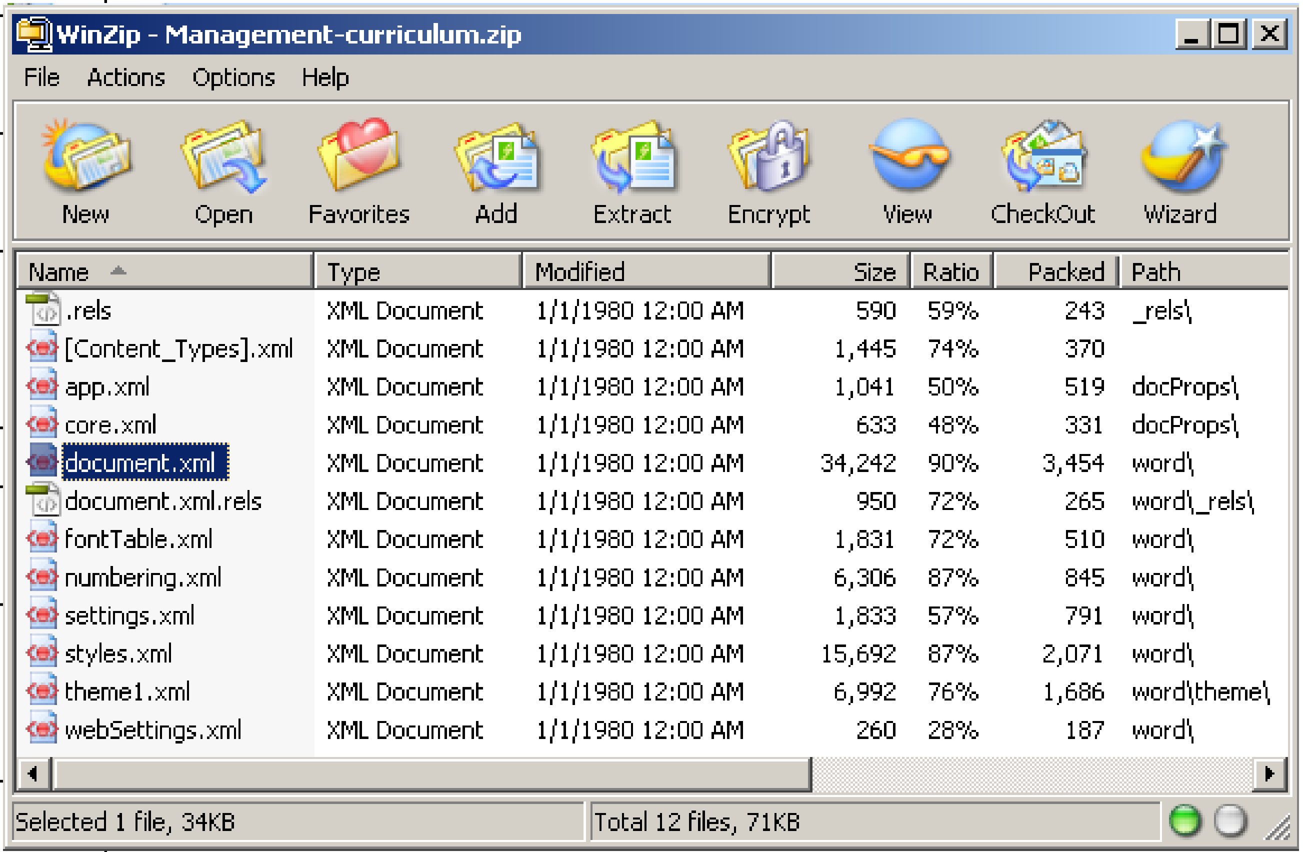
File as “raw” bitstream

Sub-file data structure

Bitstream through I/O equipment

Raw signal stream through equipment

Bitstream on physical medium



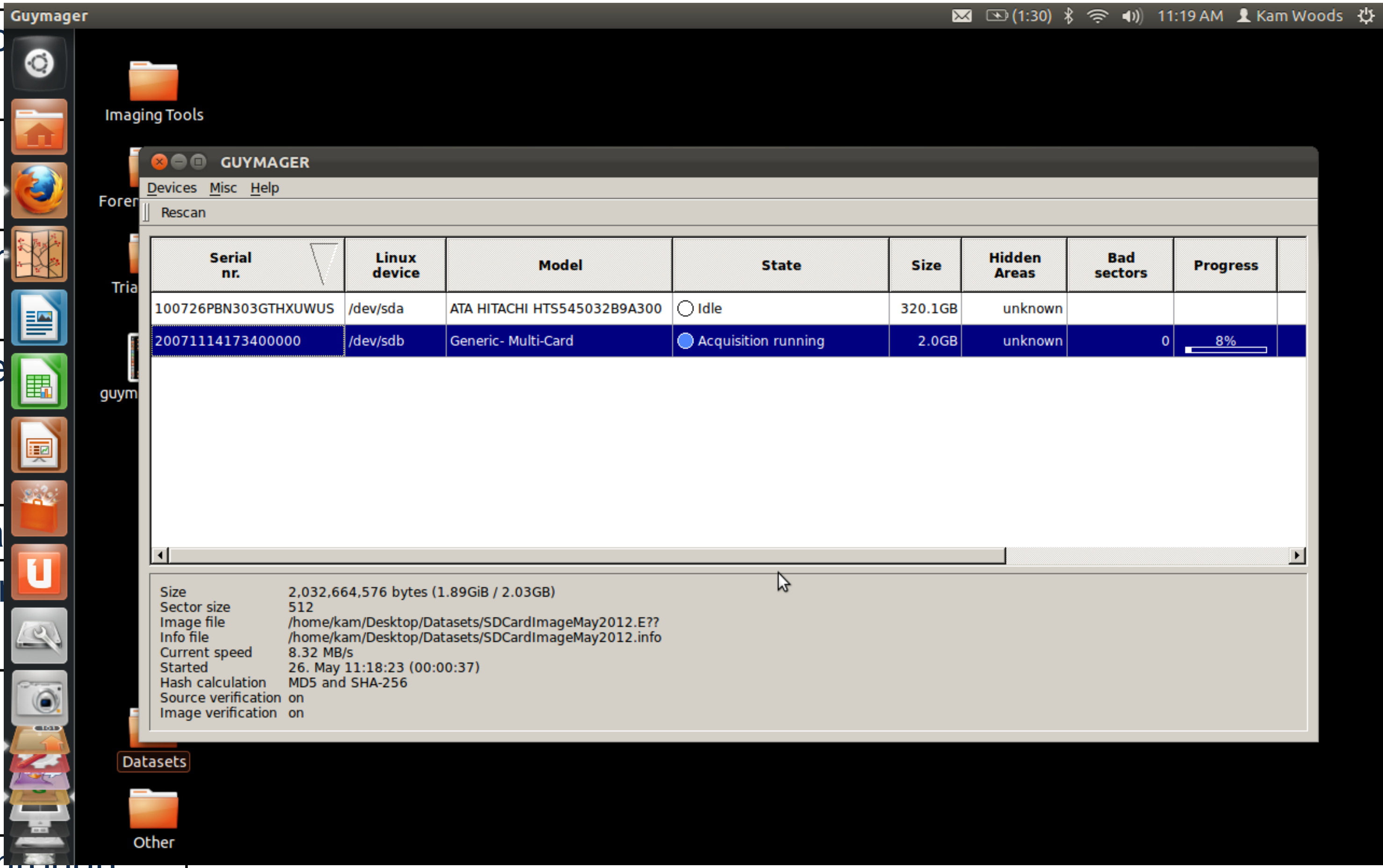
Name	Type	Modified	Size	Ratio	Packed	Path
.rels	XML Document	1/1/1980 12:00 AM	590	59%	243	_rels\
[Content_Types].xml	XML Document	1/1/1980 12:00 AM	1,445	74%	370	
app.xml	XML Document	1/1/1980 12:00 AM	1,041	50%	519	docProps\
core.xml	XML Document	1/1/1980 12:00 AM	633	48%	331	docProps\
document.xml	XML Document	1/1/1980 12:00 AM	34,242	90%	3,454	word\
document.xml.rels	XML Document	1/1/1980 12:00 AM	950	72%	265	word_rels\
fontTable.xml	XML Document	1/1/1980 12:00 AM	1,831	72%	510	word\
numbering.xml	XML Document	1/1/1980 12:00 AM	6,306	87%	845	word\
settings.xml	XML Document	1/1/1980 12:00 AM	1,833	57%	791	word\
styles.xml	XML Document	1/1/1980 12:00 AM	15,692	87%	2,071	word\
theme1.xml	XML Document	1/1/1980 12:00 AM	6,992	76%	1,686	word\theme\
webSettings.xml	XML Document	1/1/1980 12:00 AM	260	28%	187	word\

Selected 1 file, 34KB

Total 12 files, 71KB

Interaction Examples

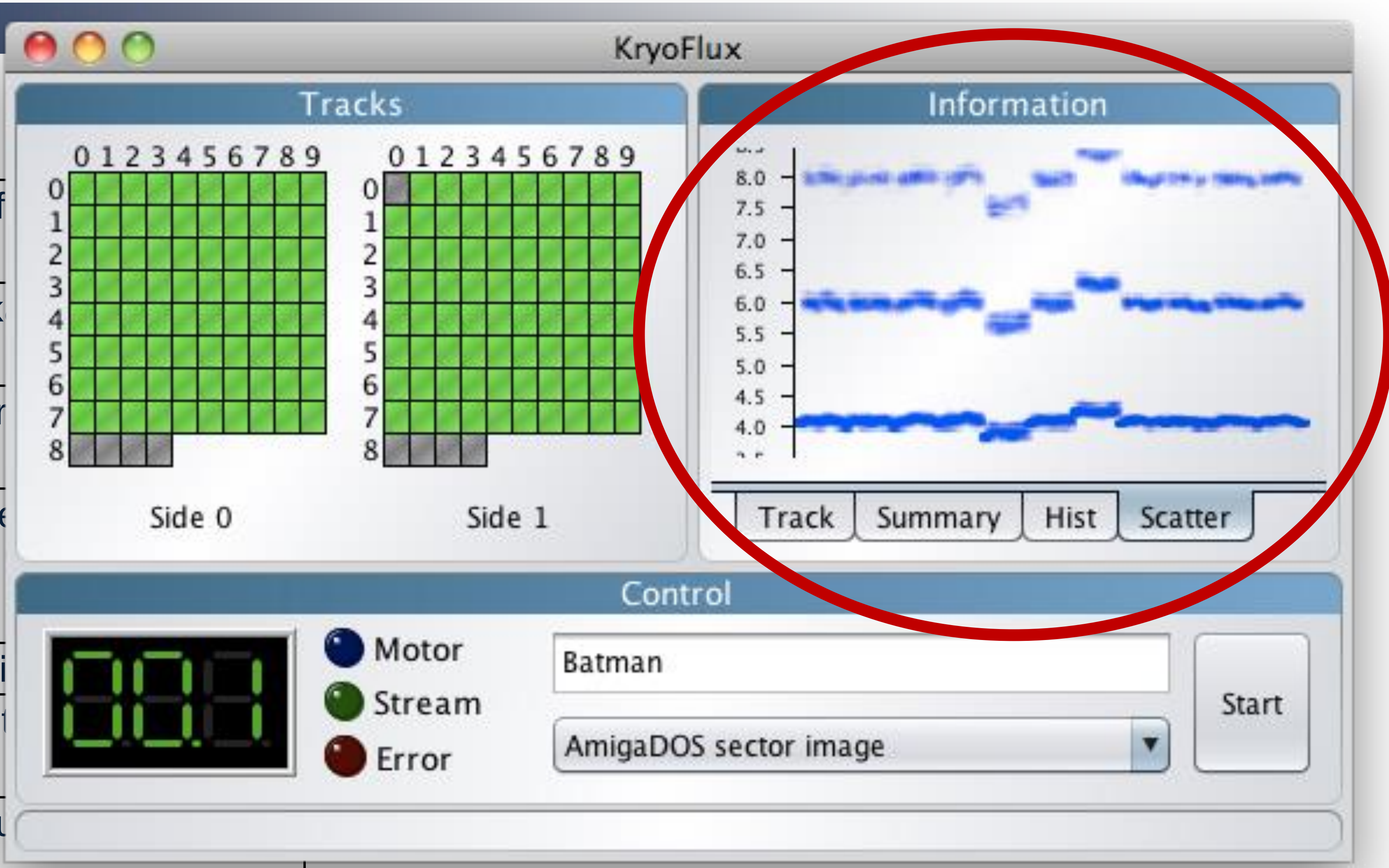
Level
Aggregation of objec
Object or package
In-application render
File through filesyste
File as “raw” bitstrea
Sub-file data structu
Bitstream through I/O equipment
Raw signal stream through I/O equipment
Bitstream on physical medium



Serial nr.	Linux device	Model	State	Size	Hidden Areas	Bad sectors	Progress
100726PBN303GTHXUWUS	/dev/sda	ATA HITACHI HTS545032B9A300	○ Idle	320.1GB	unknown		
20071114173400000	/dev/sdb	Generic- Multi-Card	● Acquisition running	2.0GB	unknown	0	8%

```
Size                 2,032,664,576 bytes (1.89GiB / 2.03GB)
Sector size          512
Image file           /home/kam/Desktop/Datasets/SDCardImageMay2012.E??
Info file            /home/kam/Desktop/Datasets/SDCardImageMay2012.info
Current speed        8.32 MB/s
Started              26. May 11:18:23 (00:00:37)
Hash calculation     MD5 and SHA-256
Source verification  on
Image verification   on
```

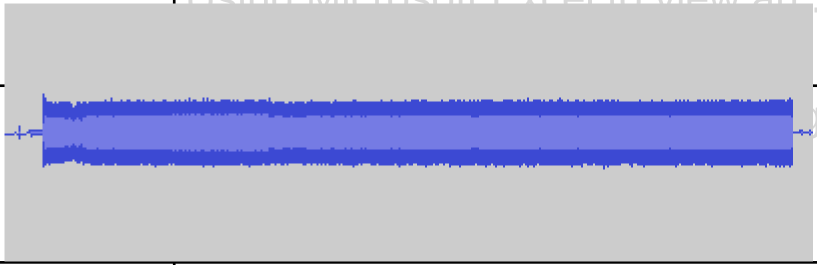
Level
Aggregation of
Object or pack
In-application r
File through file
File as “raw” bi
Sub-file data st
Bitstream throu equipment
Raw signal stream through I/O equipment
Bitstream on physical medium

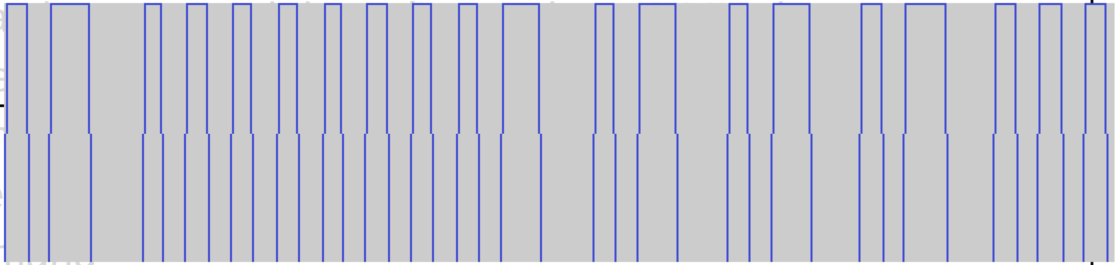


Batman

AmigaDOS sector image

Interaction Examples



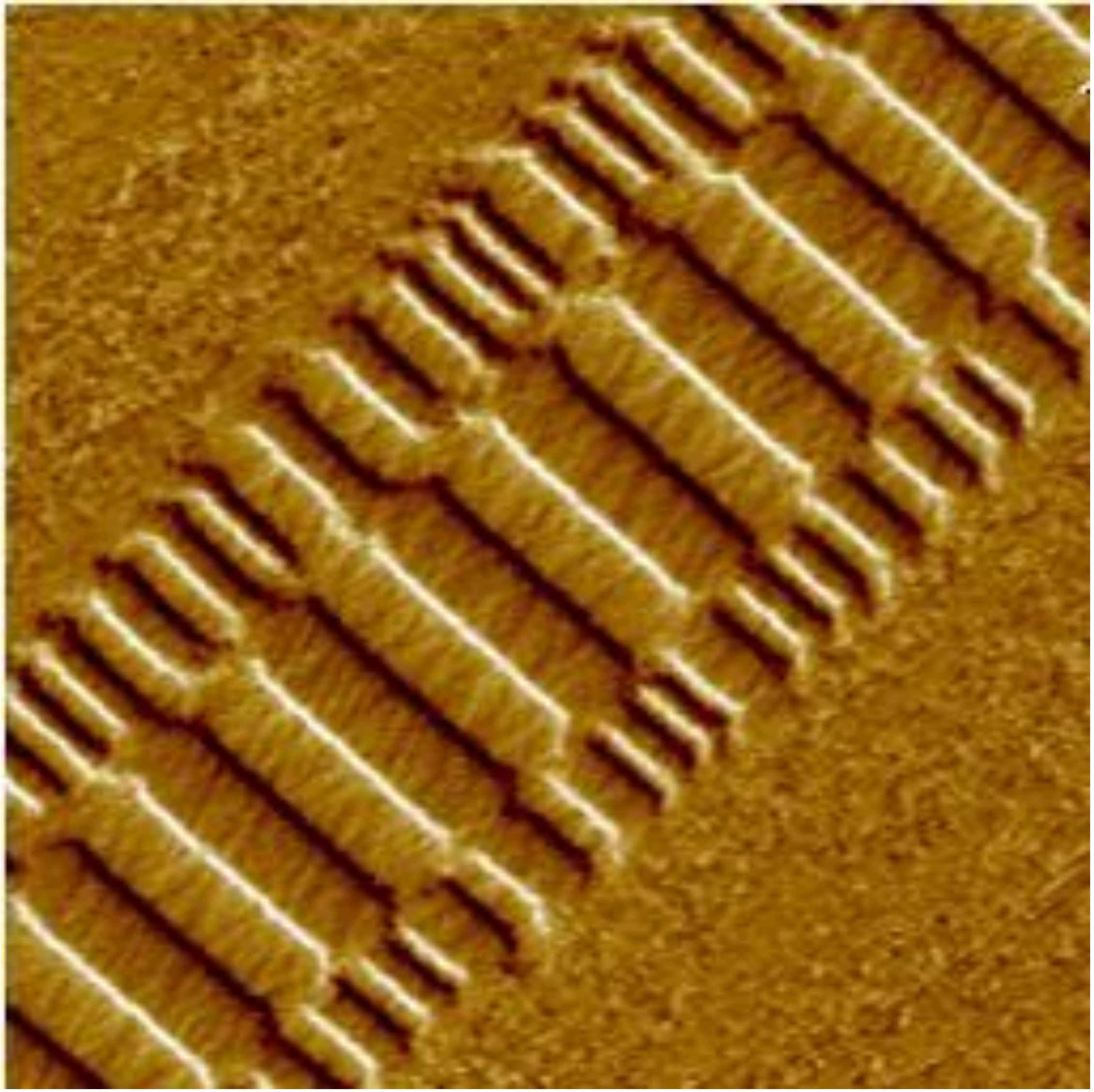


	Examples
	Browsing the contents of an archival collection using a finding aid
	Viewing a web page that contains several files, including HTML, a style sheet and several images
In-application rendering	Using Microsoft Excel to view an .xls file, watching an online
File through filesystem	Windows Explorer, typing to show the contents of a
File as “raw” bitstream	Opening an individual file in a hex editor
Sub-file data structure	Extra value
Bitstream through I/O equipment	Conn gene command
Raw signal stream through I/O equipment	Connecting a floppy drive to a host computer and then generating a magnetic flux transition image of the disk
Bitstream on physical medium	Using a high-power microscope and camera to take a picture of the patterns of magnetic charges on the surface of a hard drive or pits and lands on an optical disk

http://www.pagetable.com/?p=32

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

BitCurator

- Funded by Andrew W. Mellon Foundation
  - Phase 1: October 1, 2011 – September 30, 2013
  - Phase 2 – October 1, 2013 – September 30, 2014
- Partners: School of Information and Library Science (SILS) at UNC and Maryland Institute for Technology in the Humanities (MITH)

BitCurator Goals

- Develop a system for collecting professionals that incorporates the functionality of open-source digital forensics tools
- Address two fundamental needs not usually addressed by the digital forensics industry:
  - Incorporation into the workflow of archives/library ingest and collection management environments
  - Provision of public access to the data

BitCurator Environment*

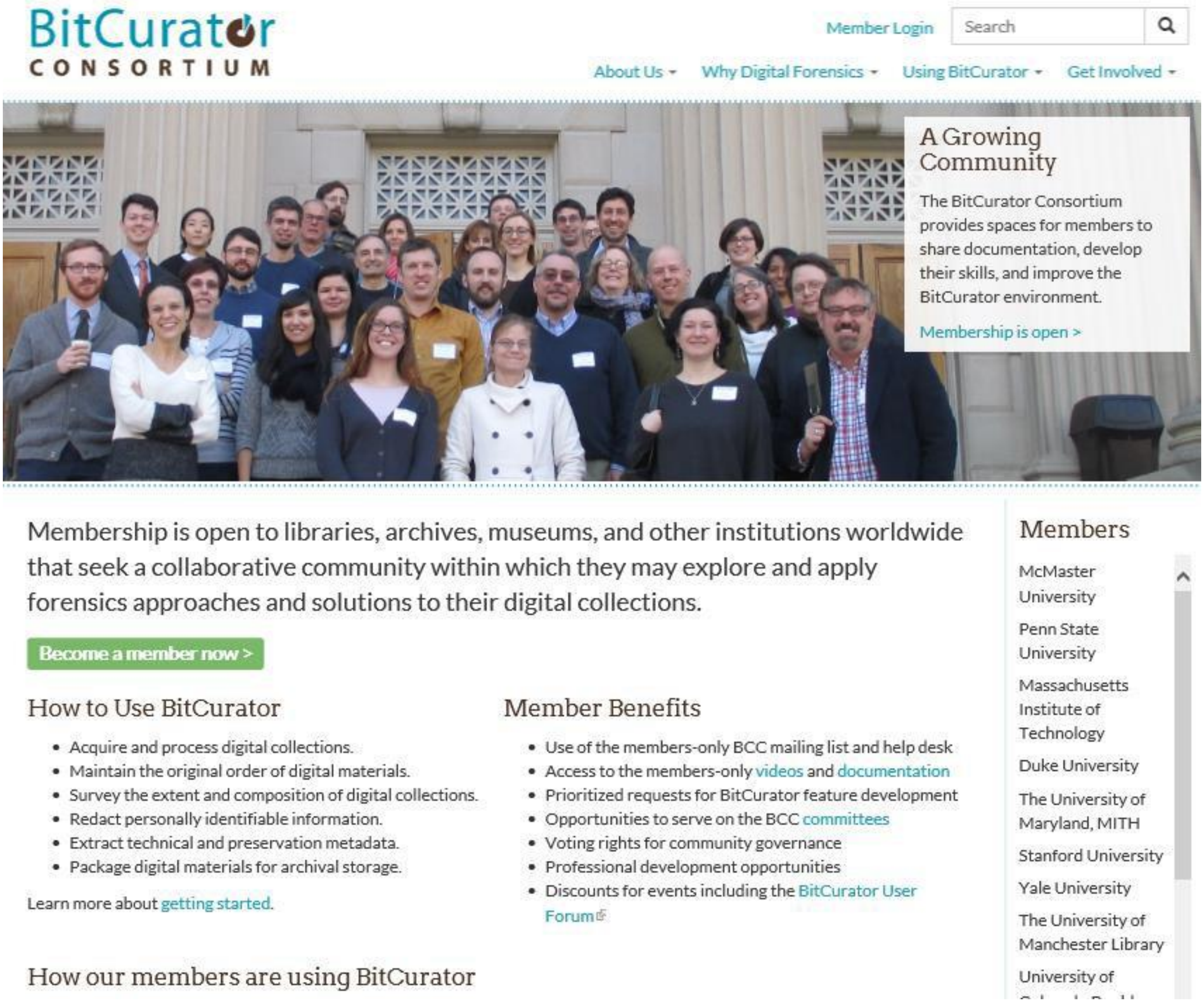
- Bundles, integrates and extends functionality of open source software
- Can be run as:
  - □ Self-contained environment running directly on a computer (download installation ISO)
  - □ Using “bootstrapping” installation scripts to turn any Ubuntu Linux machine into a BitCurator Environment
  - □ Self-contained Linux environment in a virtual machine using e.g. 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: https://github.com/BitCurator/bitcurator-distro

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)

https://bitcuratorconsortium.org/



A Growing Community

The BitCurator Consortium provides spaces for members to share documentation, develop their skills, and improve the BitCurator environment.

Membership is open >

Membership is open to libraries, archives, museums, and other institutions worldwide that seek a collaborative community within which they may explore and apply forensics approaches and solutions to their digital collections.

How to Use BitCurator

- Acquire and process digital collections.
- Maintain the original order of digital materials.
- Survey the extent and composition of digital collections.
- Redact personally identifiable information.
- Extract technical and preservation metadata.
- Package digital materials for archival storage.

Learn more about getting started.

Member Benefits

- Use of the members-only BCC mailing list and help desk
- Access to the members-only videos and documentation
- Prioritized requests for BitCurator feature development
- Opportunities to serve on the BCC committees
- Voting rights for community governance
- Professional development opportunities
- Discounts for events including the BitCurator User Forum

How our members are using BitCurator

Members

McMaster University

Penn State University

Massachusetts Institute of Technology

Duke University

The University of Maryland, MITH

Stanford University

Yale University

The University of Manchester Library

University of

BitCurator Consortium: Fostering Community

- Communication
  - Monthly community calls
  - Listserv
  - Maintains documentation feat. community scripts and data set libraries
- Active Subgroups
  - Software Development
  - Program
  - Membership Working Group
  - Executive Council
- Events
  - Mixers at various professional conferences
  - Annual User Forum

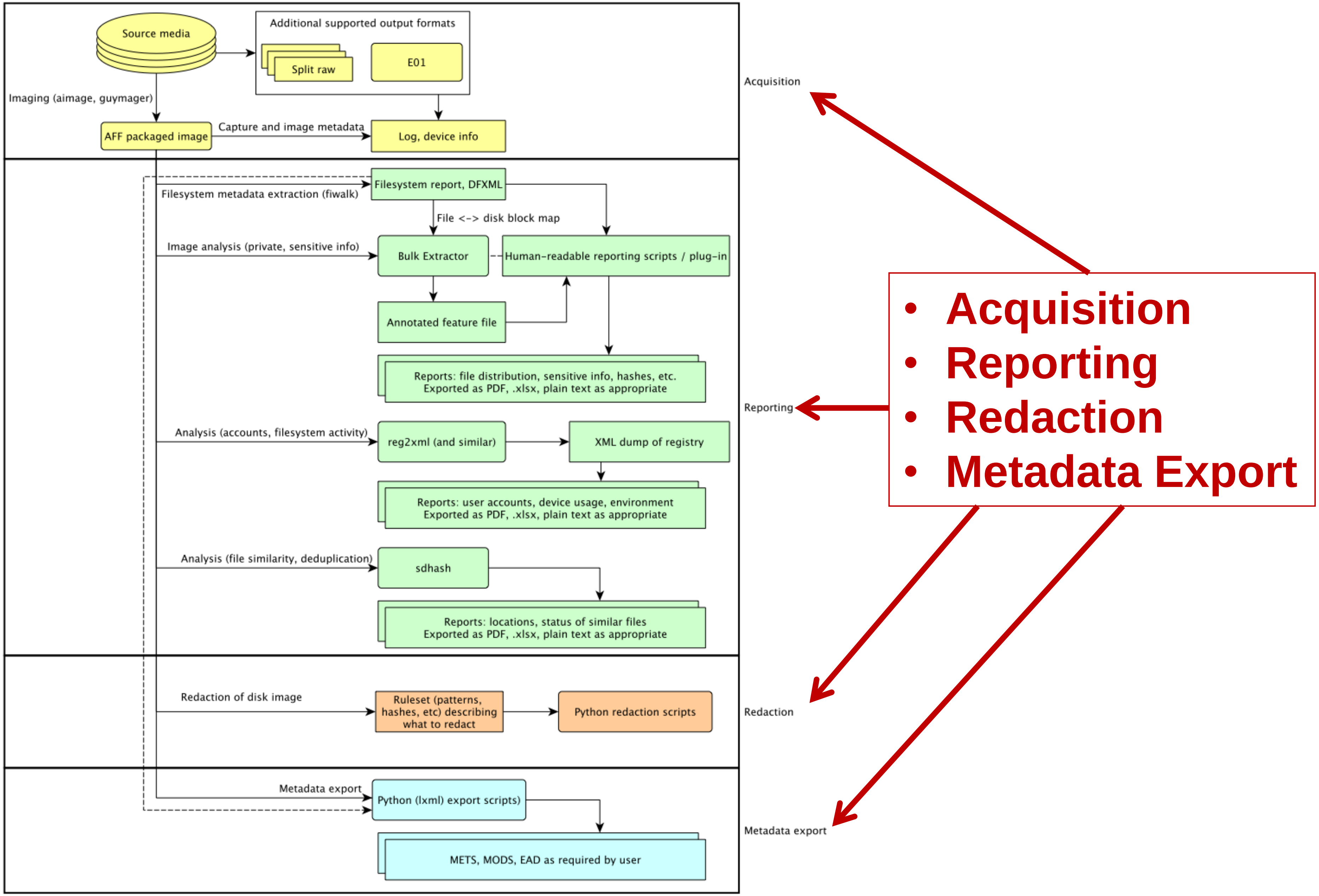
BitCuratorEdu (2018-2021)

- **Partners:** University of North Carolina at Chapel Hill School of Information and Library Science (UNC SILS), Educopia Institute, BitCurator Consortium, and the Council of State Archivists (CoSA)
- **Purpose:** study and advance adoption of digital forensics tools and methods in libraries and archives through professional education
- **Research Questions**:
    - What are the primary **institutional and technological factors** that influence adoption of digital forensics tools and methods in LIS classes in different educational settings?
    - What are the most viable mechanisms for **sustaining collaboration** among LIS programs on the adoption of digital forensics tools and methods?
- **Objectives:**
    - **produce and disseminate** learning materials
    - **investigate and report** on institutional factors to facilitate, hinder and shape adoption of educational offerings
    - **advance** community of practice around digital forensics education

Advisory Board

Catholic University	Jane Zhang, Associate Professor
Indiana University	Devan Donaldson, Assistant Professor
New York University	Howard Besser, Professor, Associate Director of MIAP, and Senior Scientist for Digital Library Initiatives for NYU Library
San Jose State University	Sandra Hirsh, Professor and Director of the School of Information; Alyce Scott, Lecturer
University of Illinois	Rhiannon Bettivia, Postdoctoral Research Associate
University of Maryland	Ricky Punzalan, Assistant Professor at iSchool, Affiliate Assistant Professor in Anthropology, and Co-Director of Museum Scholarship and Material Culture Program
University of Michigan	Paul Conway, Associate Professor
University of Texas	Patricia Galloway, Professor
Wayne State University	Kimberly Schroeder, Lecturer

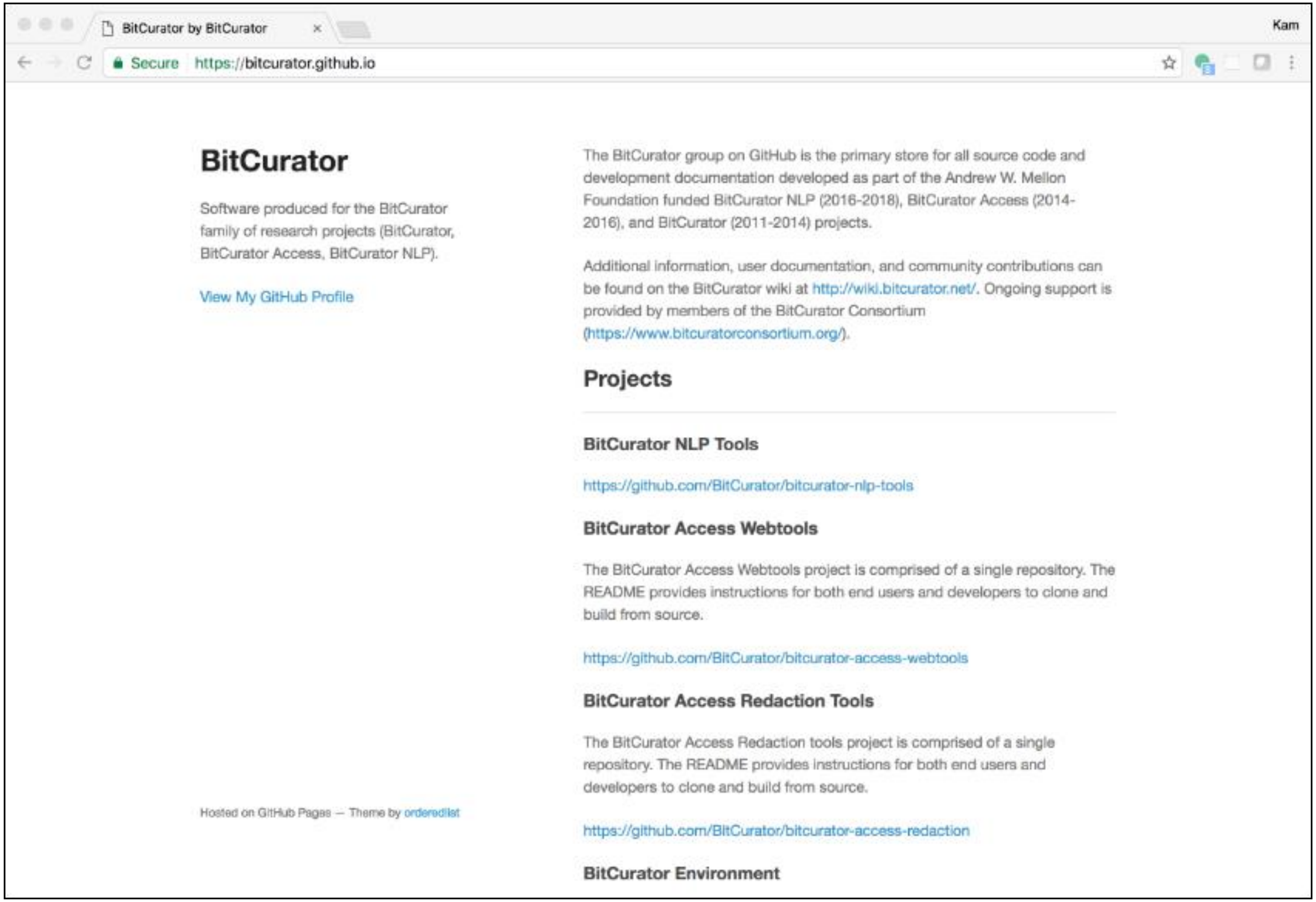
BitCurator-Supported Workflow


Source media
Additional supported output formats
Split raw
E01
Imaging (aimage, guymager)
AFF packaged image
Capture and image metadata
Log, device info
Acquisition
Filesystem report, DFXML
Filesystem metadata extraction (fiwalk)
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

For Further Information



BitCurator

Software produced for the BitCurator family of research projects (BitCurator, BitCurator Access, BitCurator NLP).

View My GitHub Profile

Hosted on GitHub Pages — Theme by orderedlist

The BitCurator group on GitHub is the primary store for all source code and development documentation developed as part of the Andrew W. Mellon Foundation funded BitCurator NLP (2016-2018), BitCurator Access (2014-2016), and BitCurator (2011-2014) projects.

Additional information, user documentation, and community contributions can be found on the BitCurator wiki at http://wiki.bitcurator.net/. Ongoing support is provided by members of the BitCurator Consortium (https://www.bitcuratorconsortium.org/).

Projects

BitCurator NLP Tools

https://github.com/BitCurator/bitcurator-nlp-tools

BitCurator Access Webtools

The BitCurator Access Webtools project is comprised of a single repository. The README provides instructions for both end users and developers to clone and build from source.

https://github.com/BitCurator/bitcurator-access-webtools

BitCurator Access Redaction Tools

The BitCurator Access Redaction tools project is comprised of a single repository. The README provides instructions for both end users and developers to clone and build from source.

https://github.com/BitCurator/bitcurator-access-redaction

BitCurator Environment

https://bitcurator.github.io/

Most of the tasks we cover in this class are explained in the Quick Start Guide. The most recent version is always available at:
https://github.com/BitCurator/bitcurator-distro/wiki/Releases

BitCurator

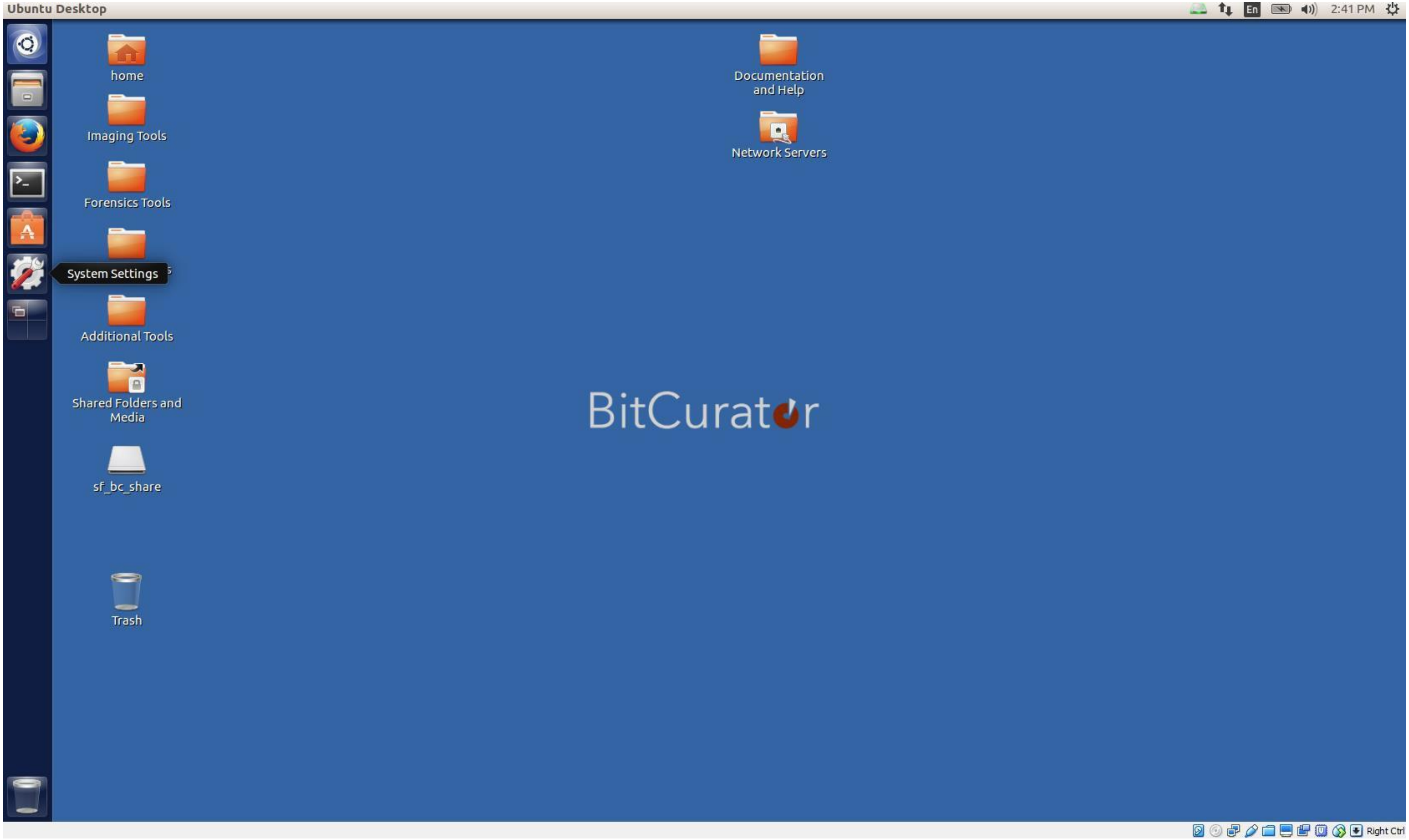
Quick Start Guide

Last updated: August 1, 2018

Release(s): 2.0.4 and later

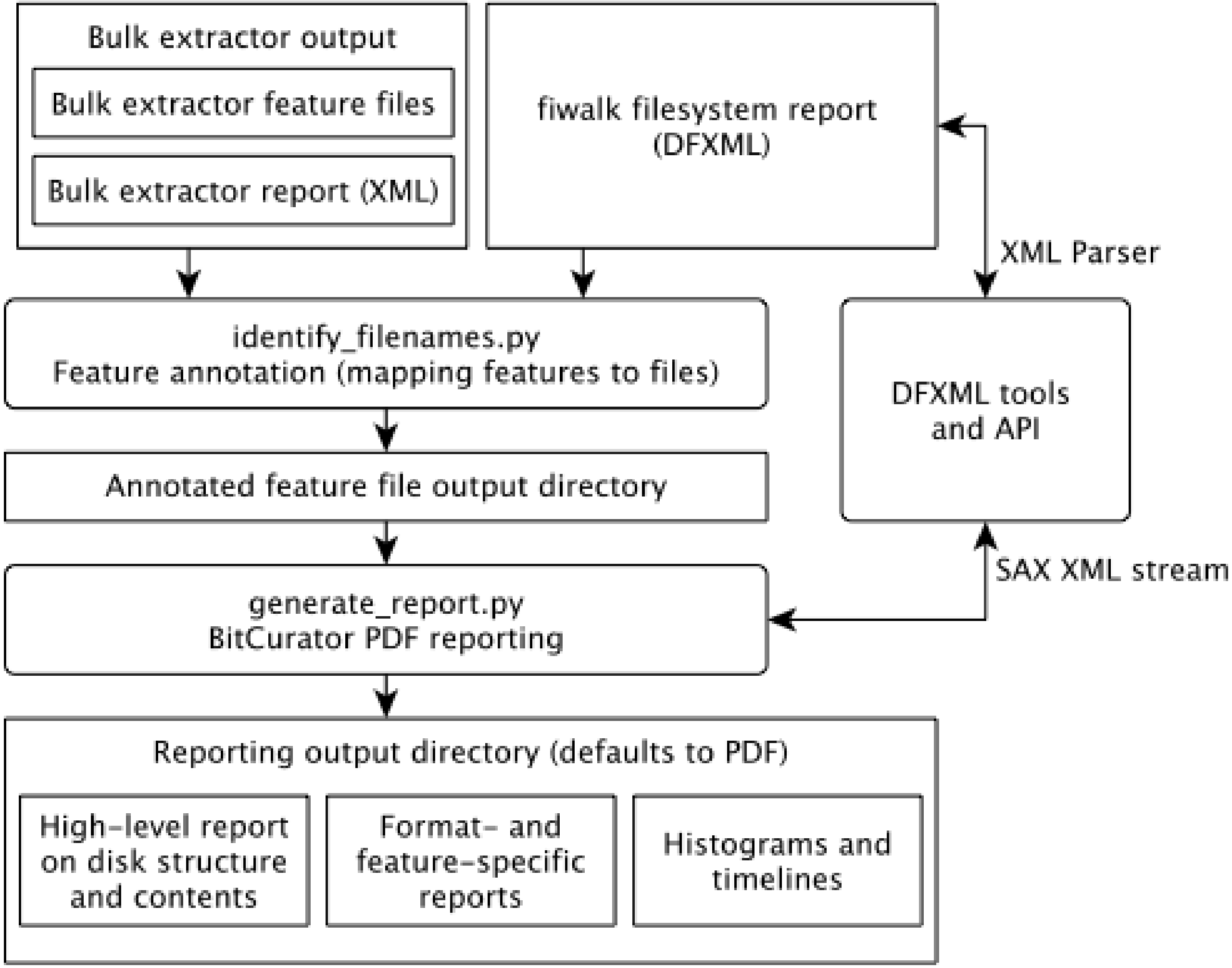
UNC
SCHOOL OF INFORMATION
AND LIBRARY SCIENCE

BitCurator
CONSORTIUM


Ubuntu Desktop
En
2:41 PM
home
Imaging Tools
Forensics Tools
System Settings
Additional Tools
Shared Folders and Media
sf_bc_share
Trash
Documentation and Help
Network Servers
BitCurator
Right Ctrl

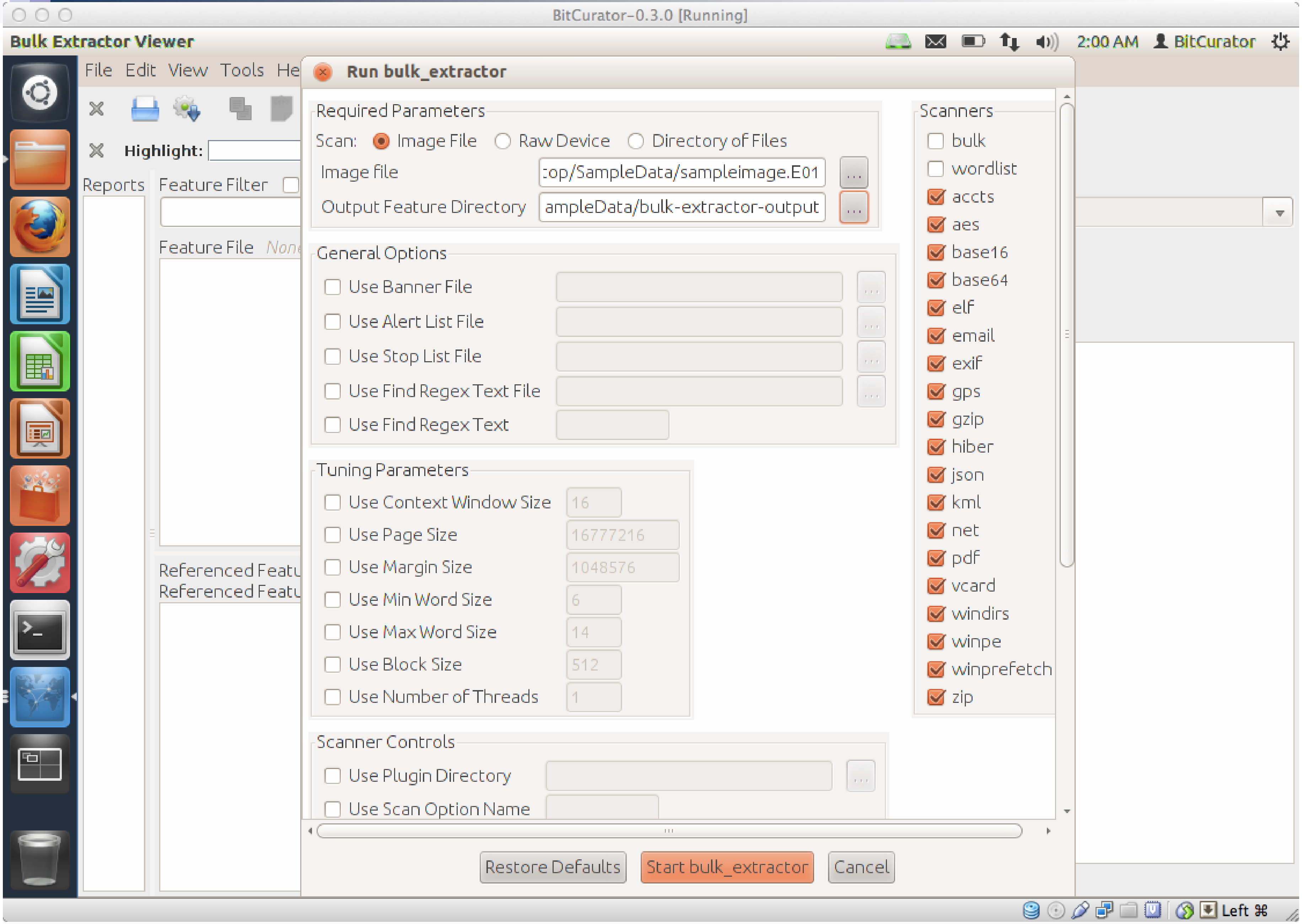
Creating and Extracting Forensic Metadata

High-Level view of Metadata Generation and Reporting



Bulk extractor output
Bulk extractor feature files
Bulk extractor report (XML)
fiwalk filesystem report (DFXML)
XML Parser
identify_filenames.py
Feature annotation (mapping features to files)
DFXML tools and API
Annotated feature file output directory
SAX XML stream
generate_report.py
BitCurator PDF reporting
Reporting output directory (defaults to PDF)
High-level report on disk structure and contents
Format- and feature-specific reports
Histograms and timelines

See: Woods, Kam, Christopher Lee, and Sunitha Misra. “Automated Analysis and Visualization of Disk Images and File Systems for Preservation.” In *Proceedings of Archiving 2013* (Springfield, VA: Society for Imaging Science and Technology, 2013), 239-244.



Run bulk_extractor

Required Parameters

Scan: ● Image File ○ Raw Device ○ Directory of Files

Image file :op/SampleData/sampleimage.E01 ...

Output Feature Directory ampleData/bulk-extractor-output ...

General Options

- □ Use Banner File
- □ Use Alert List File
- □ Use Stop List File
- □ Use Find Regex Text File
- □ Use Find Regex Text

Tuning Parameters

- □ Use Context Window Size 16
- □ Use Page Size 16777216
- □ Use Margin Size 1048576
- □ Use Min Word Size 6
- □ Use Max Word Size 14
- □ Use Block Size 512
- □ Use Number of Threads 1

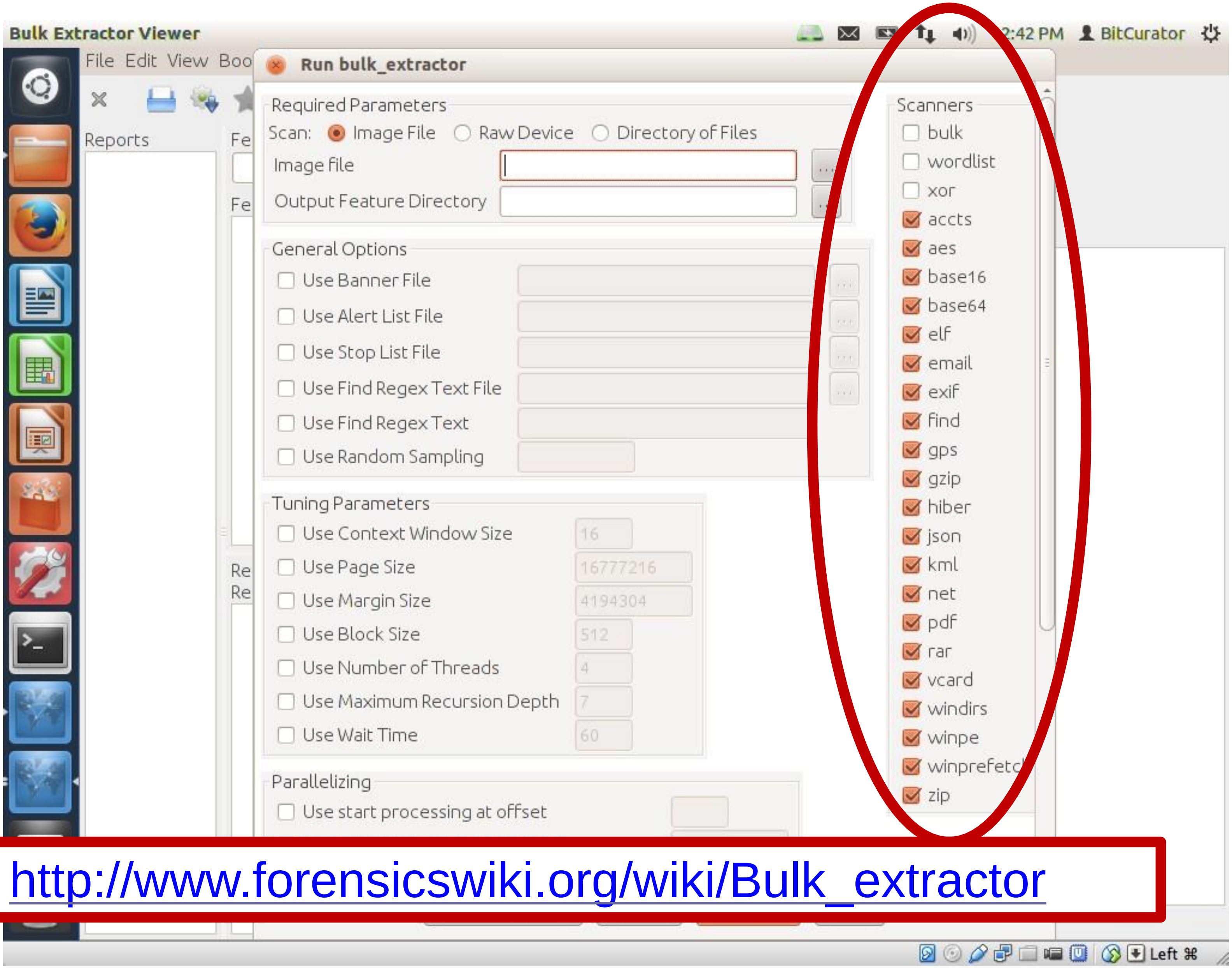
Scanner Controls

- □ Use Plugin Directory
- □ Use Scan Option Name

Scanners

- □ bulk
- □ wordlist
- ☑ accts
- ☑ aes
- ☑ base16
- ☑ base64
- ☑ elf
- ☑ email
- ☑ exif
- ☑ gps
- ☑ gzip
- ☑ hiber
- ☑ json
- ☑ kml
- ☑ net
- ☑ pdf
- ☑ vcard
- ☑ windirs
- ☑ winpe
- ☑ winprefetch
- ☑ zip

Bulk Extractor* – Identifying Potentially Sensitive Information



Bulk Extractor Viewer

Run bulk_extractor

Required Parameters

Scan: ● Image File ○ Raw Device ○ Directory of Files

Image file

Output Feature Directory

General Options

- □ Use Banner File
- □ Use Alert List File
- □ Use Stop List File
- □ Use Find Regex Text File
- □ Use Find Regex Text
- □ Use Random Sampling

Tuning Parameters

- □ Use Context Window Size 16
- □ Use Page Size 16777216
- □ Use Margin Size 4194304
- □ Use Block Size 512
- □ Use Number of Threads 4
- □ Use Maximum Recursion Depth 7
- □ Use Wait Time 60

Parallelizing

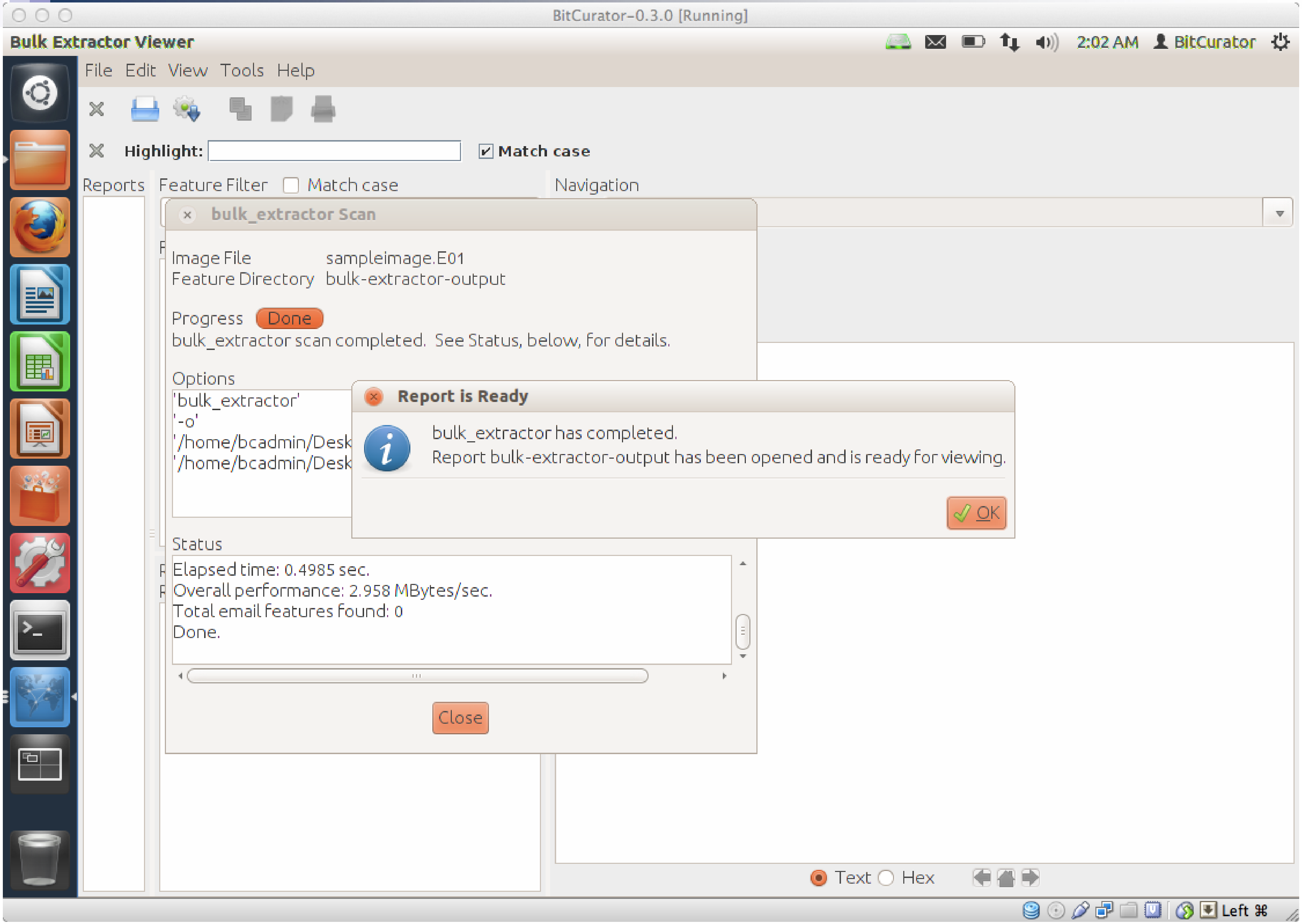
- □ Use start processing at offset

Scanners

- □ bulk
- □ wordlist
- □ xor
- ☑ accts
- ☑ aes
- ☑ base16
- ☑ base64
- ☑ elf
- ☑ email
- ☑ exif
- ☑ find
- ☑ gps
- ☑ gzip
- ☑ hiber
- ☑ json
- ☑ kml
- ☑ net
- ☑ pdf
- ☑ rar
- ☑ vcard
- ☑ windirs
- ☑ winpe
- ☑ winprefetc
- ☑ zip

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

*Developed by Simson Garfinkel



Highlight: ☑ Match case

Feature Filter □ Match case

bulk_extractor Scan

Image File sampleimage.E01
Feature Directory bulk-extractor-output

Progress Done
bulk_extractor scan completed. See Status, below, for details.

Options

```
'bulk_extractor'
'-o'
'/home/bcadmin/Desk
'/home/bcadmin/Desk
```

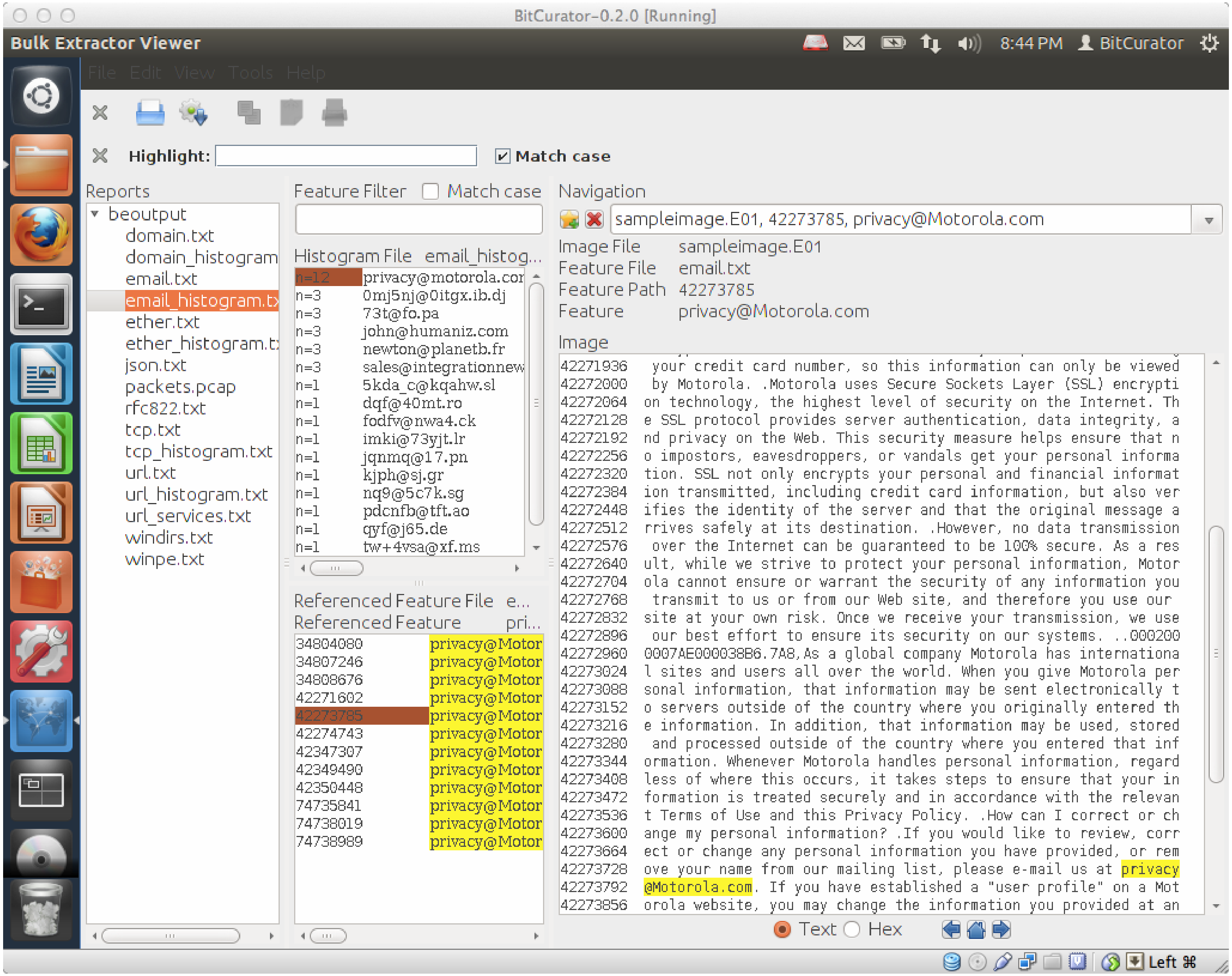
Status

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

Report is Ready

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

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



Histogram File email_histog...

```
n=12   privacy@motorola.cor
n=3    0mj5nj@0itgx.ib.dj
n=3    73t@fo.pa
n=3    john@humaniz.com
n=3    newton@planetb.fr
n=3    sales@integrationnew
n=1    5kda_c@kqahw.sl
n=1    dqf@40mt.ro
n=1    fodfv@nwa4.ck
n=1    imki@73yjt.lr
n=1    jqnmq@17.pn
n=1    kjph@sj.gr
n=1    nq9@5c7k.sg
n=1    pdcnfb@tft.ao
n=1    qyf@j65.de
n=1    tw+4vsa@xf.ms
```

Referenced Feature File e...
Referenced Feature pri...

```
34804080   privacy@Motor
34807246   privacy@Motor
34808676   privacy@Motor
42271602   privacy@Motor
42273785   privacy@Motor
42274743   privacy@Motor
42347307   privacy@Motor
42349490   privacy@Motor
42350448   privacy@Motor
74735841   privacy@Motor
74738019   privacy@Motor
74738989   privacy@Motor
```

sampleimage.E01, 42273785, privacy@Motorola.com

Image File sampleimage.E01
Feature File email.txt
Feature Path 42273785
Feature privacy@Motorola.com

Image

```
42271936   your credit card number, so this information can only be viewed
42272000   by Motorola. .Motorola uses Secure Sockets Layer (SSL) encrypti
42272064  on technology, the highest level of security on the Internet. Th
42272128  e SSL protocol provides server authentication, data integrity, a
42272192  nd privacy on the Web. This security measure helps ensure that n
42272256  o impostors, eavesdroppers, or vandals get your personal informa
42272320  tion. SSL not only encrypts your personal and financial informat
42272384  ion transmitted, including credit card information, but also ver
42272448  ifies the identity of the server and that the original message a
42272512  rrives safely at its destination. .However, no data transmission
42272576   over the Internet can be guaranteed to be 100% secure. As a res
42272640  ult, while we strive to protect your personal information, Motor
42272704  ola cannot ensure or warrant the security of any information you
42272768   transmit to us or from our Web site, and therefore you use our
42272832  site at your own risk. Once we receive your transmission, we use
42272896   our best effort to ensure its security on our systems. ..000200
42272960  0007AE000038B6.7A8,As a global company Motorola has internationa
42273024  l sites and users all over the world. When you give Motorola per
42273088  sonal information, that information may be sent electronically t
42273152  o servers outside of the country where you originally entered th
42273216  e information. In addition, that information may be used, stored
42273280   and processed outside of the country where you entered that inf
42273344  ormation. Whenever Motorola handles personal information, regard
42273408  less of where this occurs, it takes steps to ensure that your in
42273472  formation is treated securely and in accordance with the relevan
42273536  t Terms of Use and this Privacy Policy. .How can I correct or ch
42273600  ange my personal information? .If you would like to review, corr
42273664  ect or change any personal information you have provided, or rem
42273728  ove your name from our mailing list, please e-mail us at privacy
42273792  @Motorola.com. If you have established a "user profile" on a Mot
42273856  orola website, you may change the information you provided at an
```

Bulk Extractor Output*

File	Description
aes_keys.txt	AES encryption keys
alerts.txt	Processing errors
ccn.txt	Credit card numbers
ccn_track2.txt	Credit card “track 2” information, which has previously been found in some bank fraud cases
domain.txt	Internet domains found on the drive, including dotted-quad addresses found in text
email.txt	Email addresses
ether.txt	Ethernet MAC addresses found through IP packet carving of swap files and compressed system hibernation files and fragments
exif.txt	EXIF data from JPEG images and video segments
find.txt	Results of specific regular expression searches
gps.txt	Extracted GPS coordinates from Garmin XML and GPS-enabled JPEG files
ip.txt	IP addresses found through IP packet carving
json.txt	Extracted and validated JavaScript Object Notation fragments
kml.txt	Extracted KML files

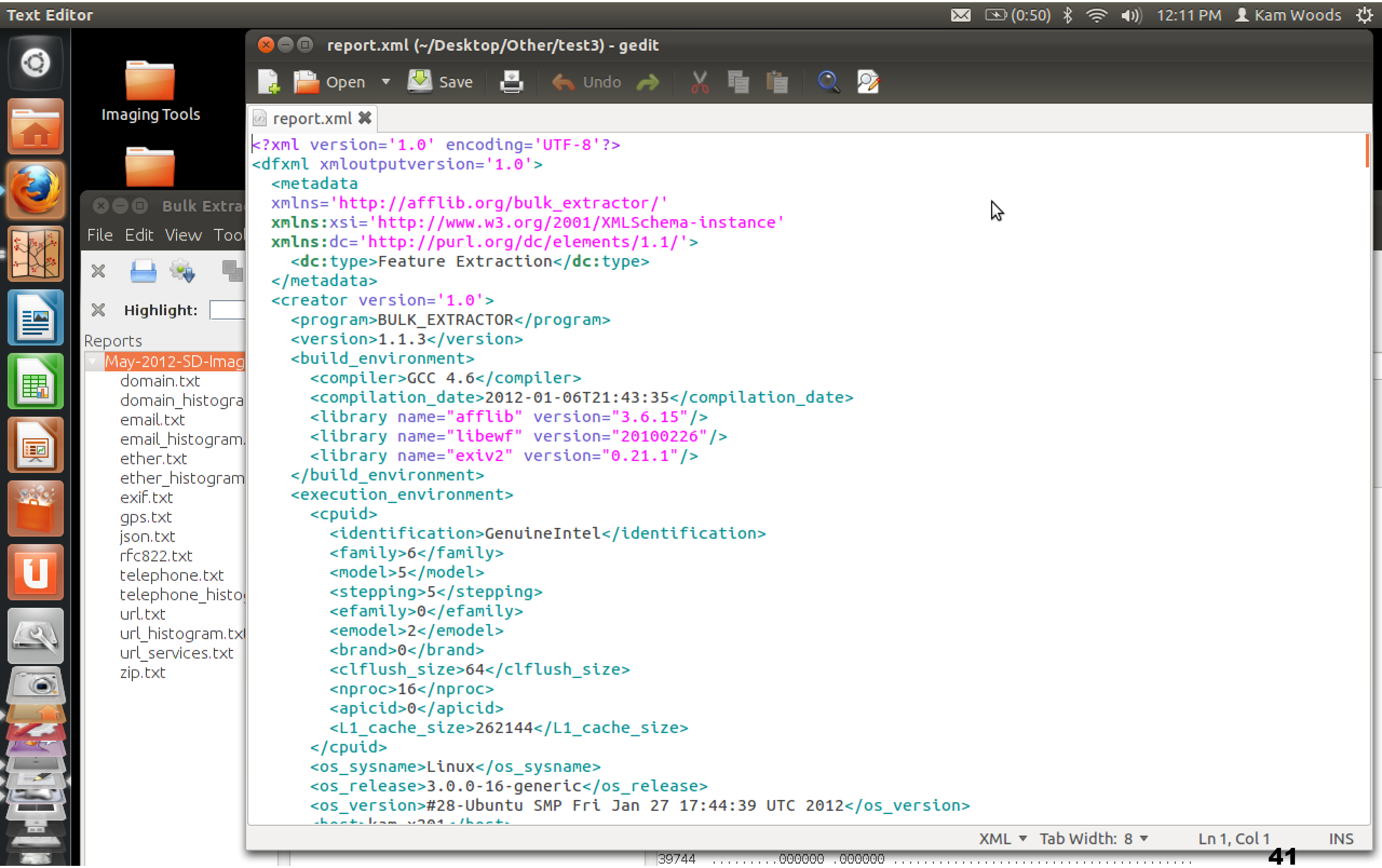
*https://www.forensicswiki.org/wiki/Bulk_extractor#Output_Feature_Files

Bulk Extractor Output (continued)*

File	Description
report.txt	DFMXL file that explains what happened
rfc822.txt	Email message headers including Date:, Subject:, and Message-ID: fields
tcp.txt	TCP flow information found through IP packet carving
telephone.txt	Phone numbers (US and other countries)
url.txt	URLs, typically found in browser caches, email messages, and pre-compiled into executables
url_searches.txt	Histogram of terms used in Internet searches
url_services.txt	Histogram of the domain name portion of all URLs found on the media
winpefect.txt	Windows prefetch files and fragments, recorded as XML
wordlist.txt	A list of all “words” extracted from the disk, useful for password cracking
wordlist_*.txt	The wordlist with duplicates removed, formatted to be imported into a popular password-cracking program
zip.txt	Information about ZIP file components found on media (including compound files such as MS Office documents)

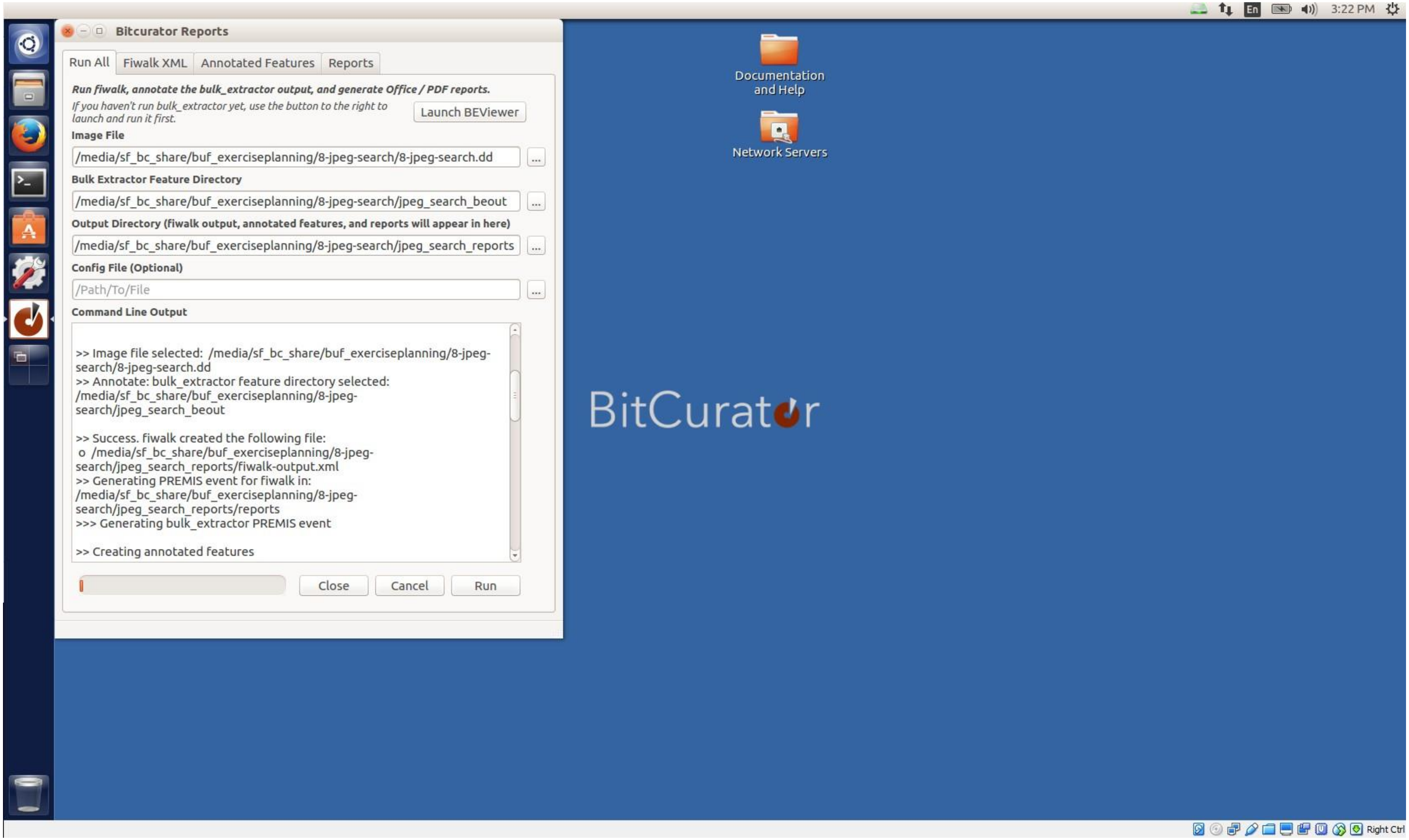
*https://www.forensicswiki.org/wiki/Bulk_extractor#Output_Feature_Files

Technical Metadata (about the System Used to do the Capture) in a Bulk Extractor Report



```
<?xml version='1.0' encoding='UTF-8'?>
<dfxml xmloutputversion='1.0'>
  <metadata
  xmlns='http://afflib.org/bulk_extractor/'
  xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'
  xmlns:dc='http://purl.org/dc/elements/1.1/'>
    <dc:type>Feature Extraction</dc:type>
  </metadata>
  <creator version='1.0'>
    <program>BULK_EXTRACTOR</program>
    <version>1.1.3</version>
    <build_environment>
      <compiler>GCC 4.6</compiler>
      <compilation_date>2012-01-06T21:43:35</compilation_date>
      <library name="afflib" version="3.6.15"/>
      <library name="libewf" version="20100226"/>
      <library name="exiv2" version="0.21.1"/>
    </build_environment>
    <execution_environment>
      <cpuid>
        <identification>GenuineIntel</identification>
        <family>6</family>
        <model>5</model>
        <stepping>5</stepping>
        <efamily>0</efamily>
        <emodel>2</emodel>
        <brand>0</brand>
        <clflush_size>64</clflush_size>
        <nproc>16</nproc>
        <apicid>0</apicid>
        <L1_cache_size>262144</L1_cache_size>
      </cpuid>
      <os_sysname>Linux</os_sysname>
      <os_release>3.0.0-16-generic</os_release>
      <os_version>#28-Ubuntu SMP Fri Jan 27 17:44:39 UTC 2012</os_version>
      <host>kam-x201</host>
```

BitCurator Reporting Tool



Run fiwalk, annotate the bulk_extractor output, and generate Office / PDF reports.

If you haven't run bulk_extractor yet, use the button to the right to launch and run it first.

Image File
/media/sf_bc_share/buf_exerciseplanning/8-jpeg-search/8-jpeg-search.dd

Bulk Extractor Feature Directory
/media/sf_bc_share/buf_exerciseplanning/8-jpeg-search/jpeg_search_beout

Output Directory (fiwalk output, annotated features, and reports will appear in here)
/media/sf_bc_share/buf_exerciseplanning/8-jpeg-search/jpeg_search_reports

Config File (Optional)
/Path/To/File

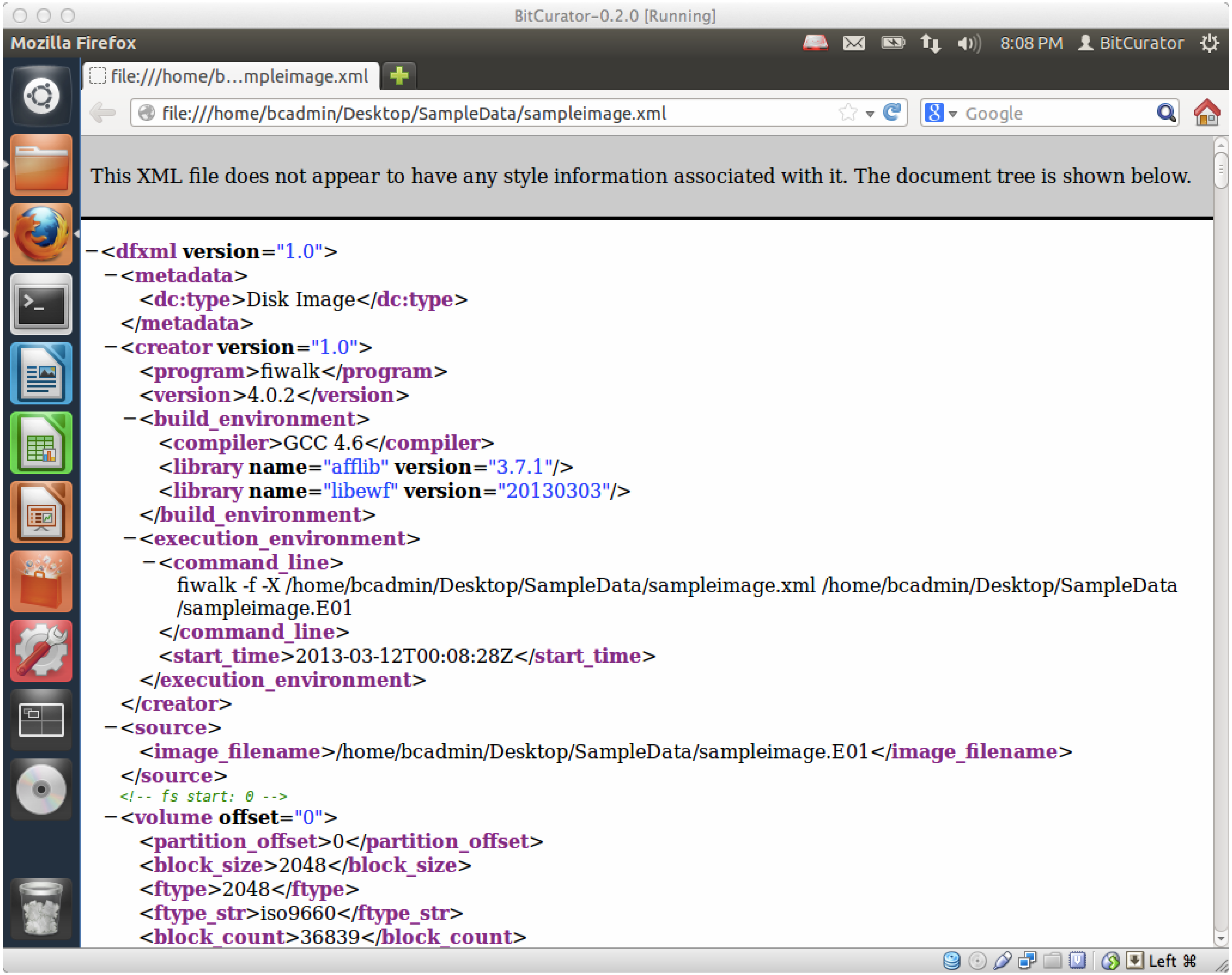
Command Line Output

```
>> Image file selected:  /media/sf_bc_share/buf_exerciseplanning/8-jpeg-
search/8-jpeg-search.dd
>> Annotate: bulk_extractor feature directory selected:
/media/sf_bc_share/buf_exerciseplanning/8-jpeg-
search/jpeg_search_beout

>> Success. fiwalk created the following file:
 o /media/sf_bc_share/buf_exerciseplanning/8-jpeg-
search/jpeg_search_reports/fiwalk-output.xml
>> Generating PREMIS event for fiwalk in:
/media/sf_bc_share/buf_exerciseplanning/8-jpeg-
search/jpeg_search_reports/reports
>>> Generating bulk_extractor PREMIS event

>> Creating annotated features
```

Provenance – DFXML Output from fiwalk

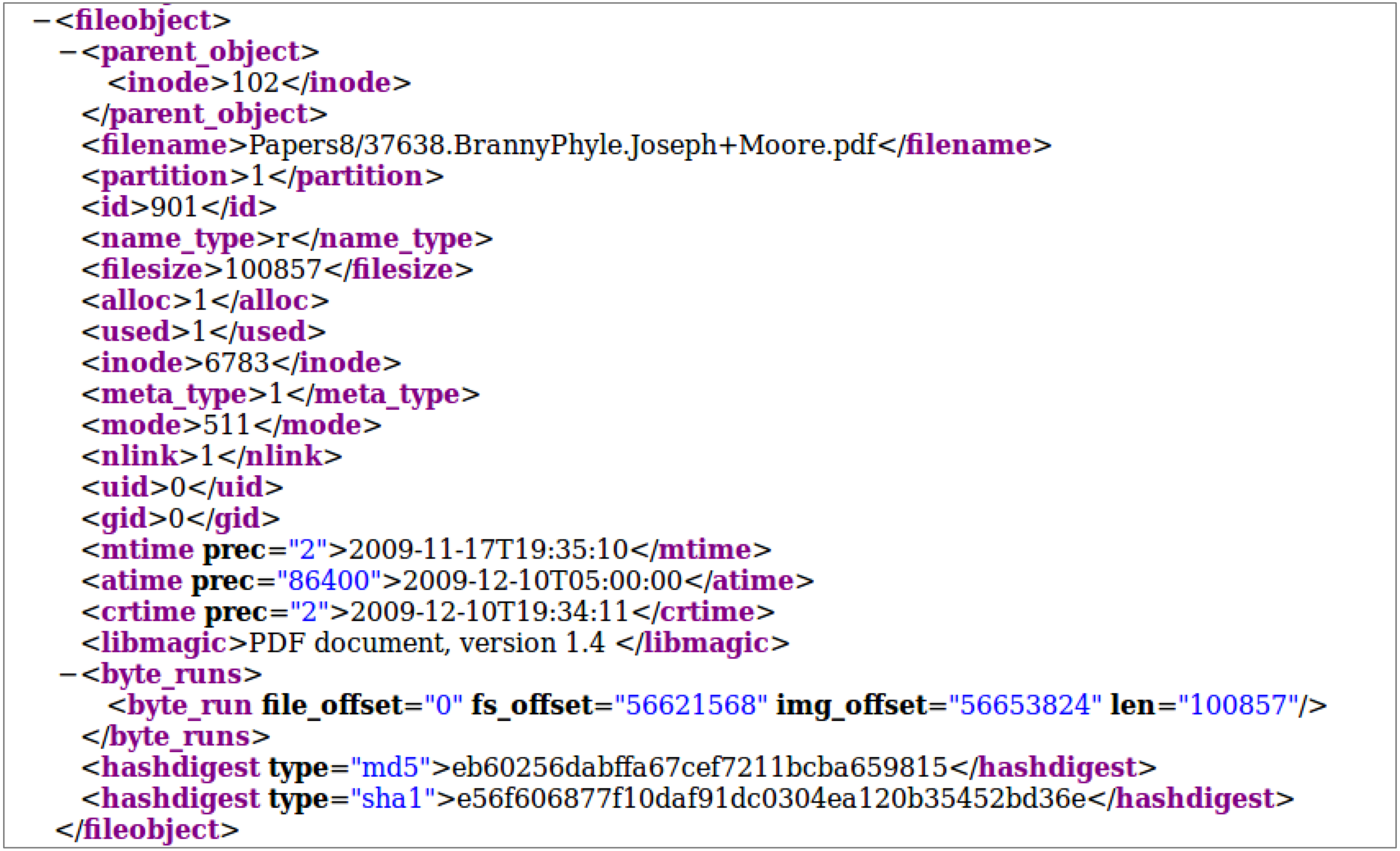


BitCurator-0.2.0 [Running]
Mozilla Firefox
8:08 PM
BitCurator
file:///home/b...mpleimage.xml
file:///home/bcadmin/Desktop/SampleData/sampleimage.xml
Google
Left ⌘

This XML file does not appear to have any style information associated with it. The document tree is shown below.

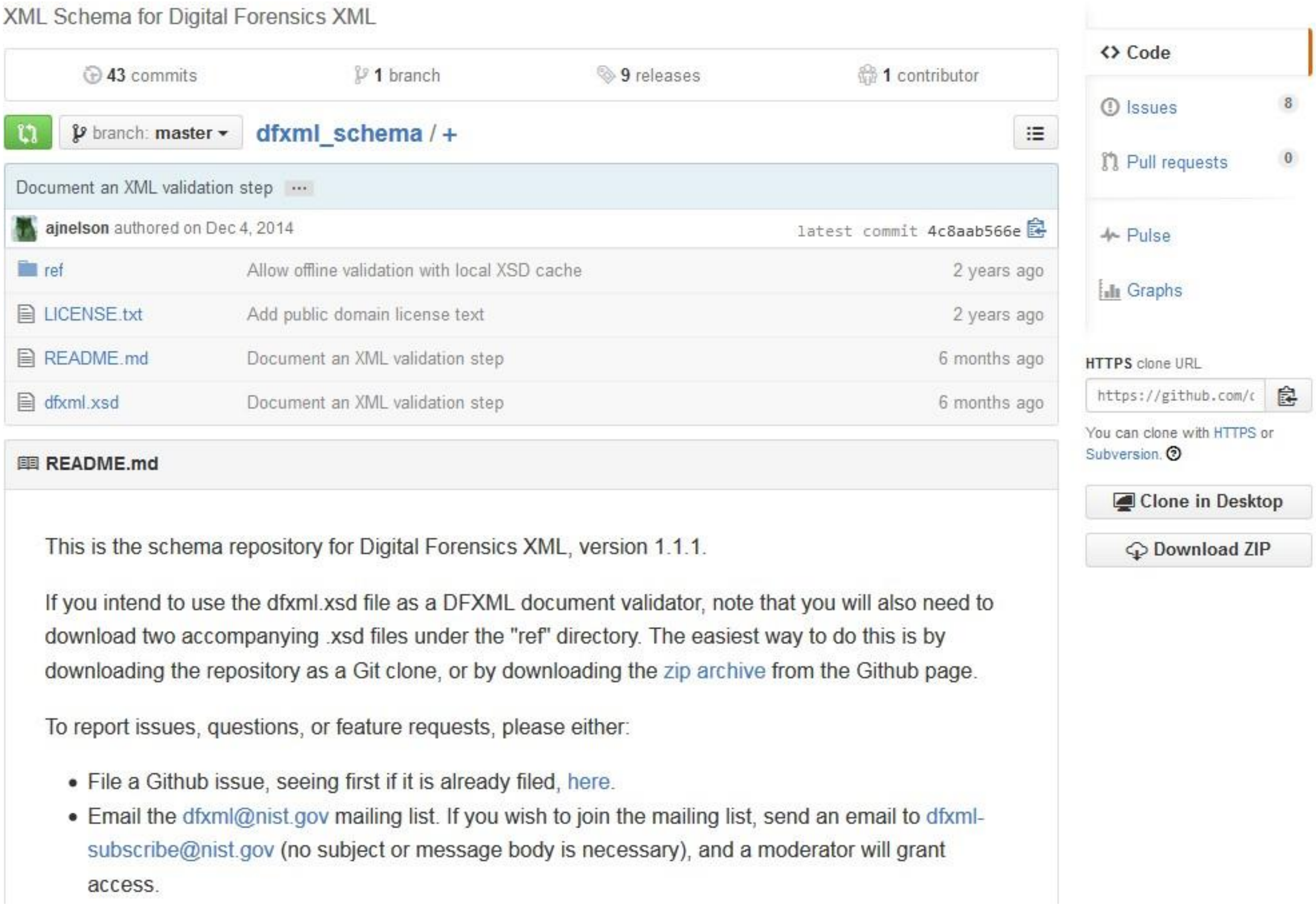
```
−<dfxml version="1.0">
  −<metadata>
      <dc:type>Disk Image</dc:type>
    </metadata>
  −<creator version="1.0">
      <program>fiwalk</program>
      <version>4.0.2</version>
    −<build_environment>
        <compiler>GCC 4.6</compiler>
        <library name="afflib" version="3.7.1"/>
        <library name="libewf" version="20130303"/>
      </build_environment>
    −<execution_environment>
      −<command_line>
          fiwalk -f -X /home/bcadmin/Desktop/SampleData/sampleimage.xml /home/bcadmin/Desktop/SampleData
          /sampleimage.E01
        </command_line>
        <start_time>2013-03-12T00:08:28Z</start_time>
      </execution_environment>
    </creator>
  −<source>
      <image_filename>/home/bcadmin/Desktop/SampleData/sampleimage.E01</image_filename>
    </source>
    <!-- fs start: 0 -->
  −<volume offset="0">
      <partition_offset>0</partition_offset>
      <block_size>2048</block_size>
      <ftype>2048</ftype>
      <ftype_str>iso9660</ftype_str>
      <block_count>36839</block_count>
```

Capturing Original Order - Filesystem Metadata Output from fiwalk*



```
 -<fileobject>
   -<parent_object>
      <inode>102</inode>
    </parent_object>
    <filename>Papers8/37638.BrannyPhyle.Joseph+Moore.pdf</filename>
    <partition>1</partition>
    <id>901</id>
    <name_type>r</name_type>
    <filesize>100857</filesize>
    <alloc>1</alloc>
    <used>1</used>
    <inode>6783</inode>
    <meta_type>1</meta_type>
    <mode>511</mode>
    <nlink>1</nlink>
    <uid>0</uid>
    <gid>0</gid>
    <mtime prec="2">2009-11-17T19:35:10</mtime>
    <atime prec="86400">2009-12-10T05:00:00</atime>
    <crtime prec="2">2009-12-10T19:34:11</crtime>
    <libmagic>PDF document, version 1.4 </libmagic>
   -<byte_runs>
      <byte_run file_offset="0" fs_offset="56621568" img_offset="56653824" len="100857"/>
    </byte_runs>
    <hashdigest type="md5">eb60256dabffa67cef7211bcba659815</hashdigest>
    <hashdigest type="sha1">e56f606877f10daf91dc0304ea120b35452bd36e</hashdigest>
  </fileobject>
```

*Developed by Simson Garfinkel



XML Schema for Digital Forensics XML

Document an XML validation step

ajnelson authored on Dec 4, 2014 latest commit 4c8aab566e

ref	Allow offline validation with local XSD cache	2 years ago
LICENSE.txt	Add public domain license text	2 years ago
README.md	Document an XML validation step	6 months ago
dfxml.xsd	Document an XML validation step	6 months ago

You can clone with HTTPS or Subversion.

README.md

This is the schema repository for Digital Forensics XML, version 1.1.1.

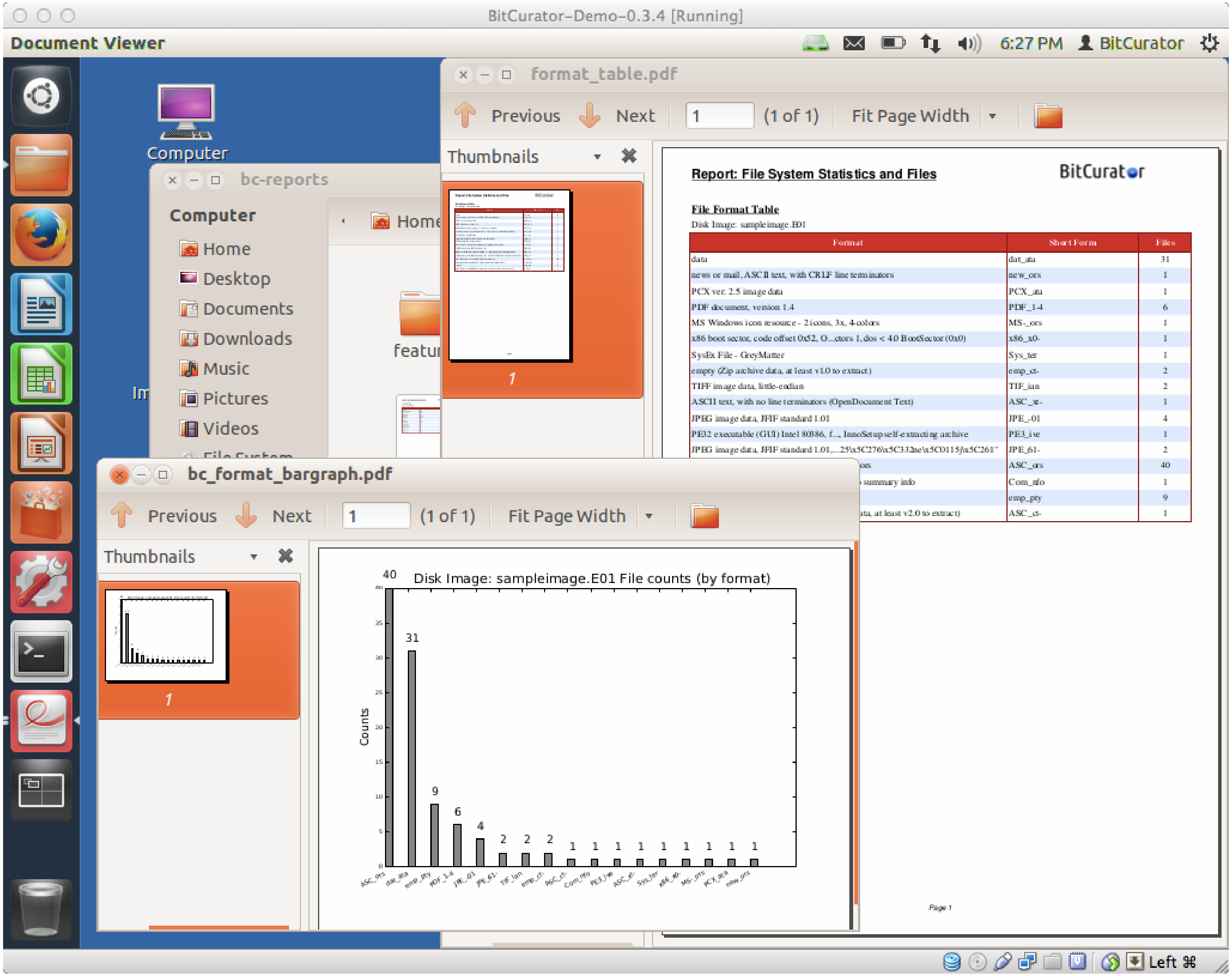
If you intend to use the dfxml.xsd file as a DFXML document validator, note that you will also need to download two accompanying .xsd files under the "ref" directory. The easiest way to do this is by downloading the repository as a Git clone, or by downloading the zip archive from the Github page.

To report issues, questions, or feature requests, please either:

- File a Github issue, seeing first if it is already filed, here.
- Email the dfxml@nist.gov mailing list. If you wish to join the mailing list, send an email to dfxml-subscribe@nist.gov (no subject or message body is necessary), and a moderator will grant access.

https://github.com/dfxml-working-group/dfxml_schema

Various Specialized BitCurator Reports



Report: File System Statistics and Files

BitCurator

File Format Table

Disk Image: sampleimage.E01

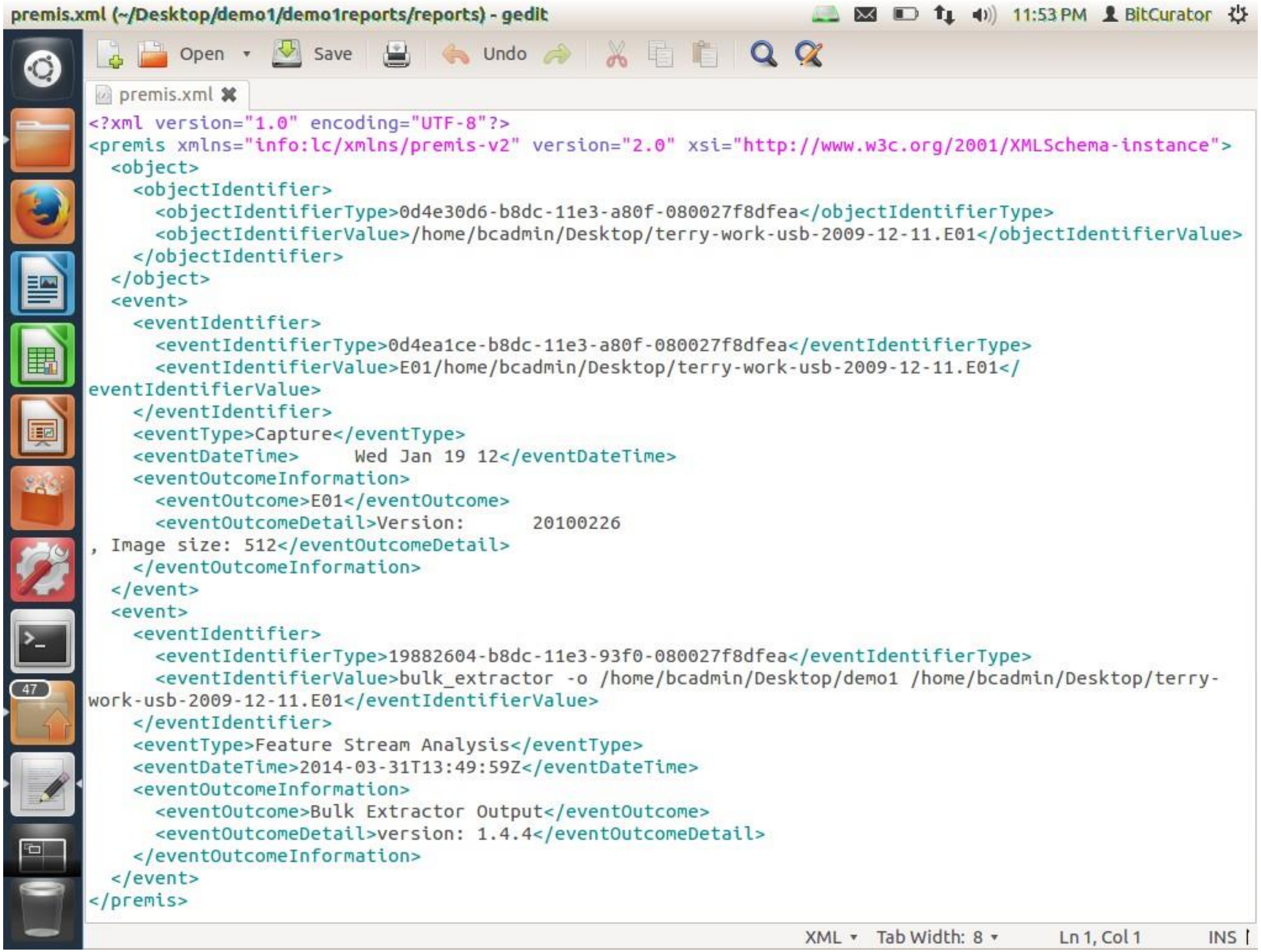
Format	Short Form	Files
data	dat_ata	31
news or mail, ASCII text, with CRLF line terminators	new_ors	1
PCX ver. 2.5 image data	PCX_ata	1
PDF document, version 1.4	PDF_1-4	6
MS Windows icon resource - 2 icons, 3x, 4-colors	MS-_ors	1
x86 boot sector, code offset 0x52, O...ctors 1, dos < 4.0 BootSector (0x0)	x86_x0-	1
SysEx File - GreyMatter	Sys_ter	1
empty (Zip archive data, at least v1.0 to extract)	emp_ct-	2
TIFF image data, little-endian	TIF_ian	2
ASCII text, with no line terminators (OpenDocument Text)	ASC_xt-	1
JPEG image data, JFIF standard 1.01	JPE_-01	4
PE32 executable (GUI) Intel 80386, f..., InnoSetup self-extracting archive	PE3_ive	1
JPEG image data, JFIF standard 1.01,...25\x5C276\x5C332ne\x5C0115f\x5C261"	JPE_61-	2
ors	ASC_ors	40
summary info	Com_nfo	1
	emp_pty	9
ata, at least v2.0 to extract)	ASC_ct-	1

Page 1

Specialized BitCurator Reports

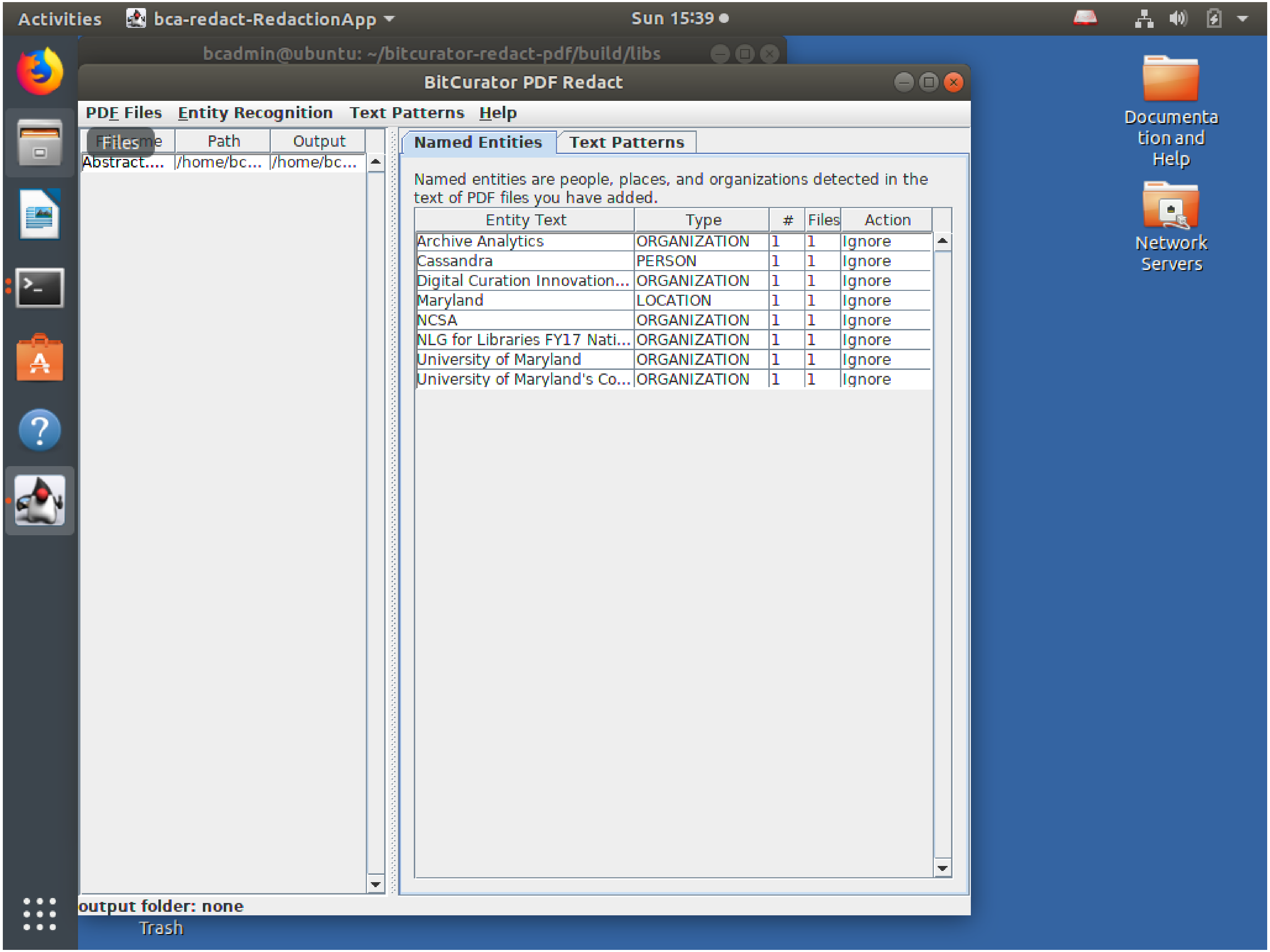
File	Content
bc_format_bargraph.pdf	histogram of file formats found on the volume
bulk_extractor_report.pdf	high-level overview of feature locations on disk
fiwalk_deleted_files.pdf	shows paths to any deleted materials found in a given partition
fiwalk-output.xml.xlsx	Excel converted DFXML output (file 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

PREMIS (Preservation) Metadata Generated from Running BitCurator Tools – Recorded as PREMIS Events



```
<?xml version="1.0" encoding="UTF-8"?>
<premis xmlns="info:lc/xmlns/premis-v2" version="2.0" xsi="http://www.w3c.org/2001/XMLSchema-instance">
  <object>
    <objectIdentifier>
      <objectIdentifierType>0d4e30d6-b8dc-11e3-a80f-080027f8dfea</objectIdentifierType>
      <objectIdentifierValue>/home/bcadmin/Desktop/terry-work-usb-2009-12-11.E01</objectIdentifierValue>
    </objectIdentifier>
  </object>
  <event>
    <eventIdentifier>
      <eventIdentifierType>0d4ea1ce-b8dc-11e3-a80f-080027f8dfea</eventIdentifierType>
      <eventIdentifierValue>E01/home/bcadmin/Desktop/terry-work-usb-2009-12-11.E01</
eventIdentifierValue>
    </eventIdentifier>
    <eventType>Capture</eventType>
    <eventDateTime>     Wed Jan 19 12</eventDateTime>
    <eventOutcomeInformation>
      <eventOutcome>E01</eventOutcome>
      <eventOutcomeDetail>Version:      20100226
, Image size: 512</eventOutcomeDetail>
    </eventOutcomeInformation>
  </event>
  <event>
    <eventIdentifier>
      <eventIdentifierType>19882604-b8dc-11e3-93f0-080027f8dfea</eventIdentifierType>
      <eventIdentifierValue>bulk_extractor -o /home/bcadmin/Desktop/demo1 /home/bcadmin/Desktop/terry-
work-usb-2009-12-11.E01</eventIdentifierValue>
    </eventIdentifier>
    <eventType>Feature Stream Analysis</eventType>
    <eventDateTime>2014-03-31T13:49:59Z</eventDateTime>
    <eventOutcomeInformation>
      <eventOutcome>Bulk Extractor Output</eventOutcome>
      <eventOutcomeDetail>version: 1.4.4</eventOutcomeDetail>
    </eventOutcomeInformation>
  </event>
</premis>
```

BitCurator PDF Redaction Tool

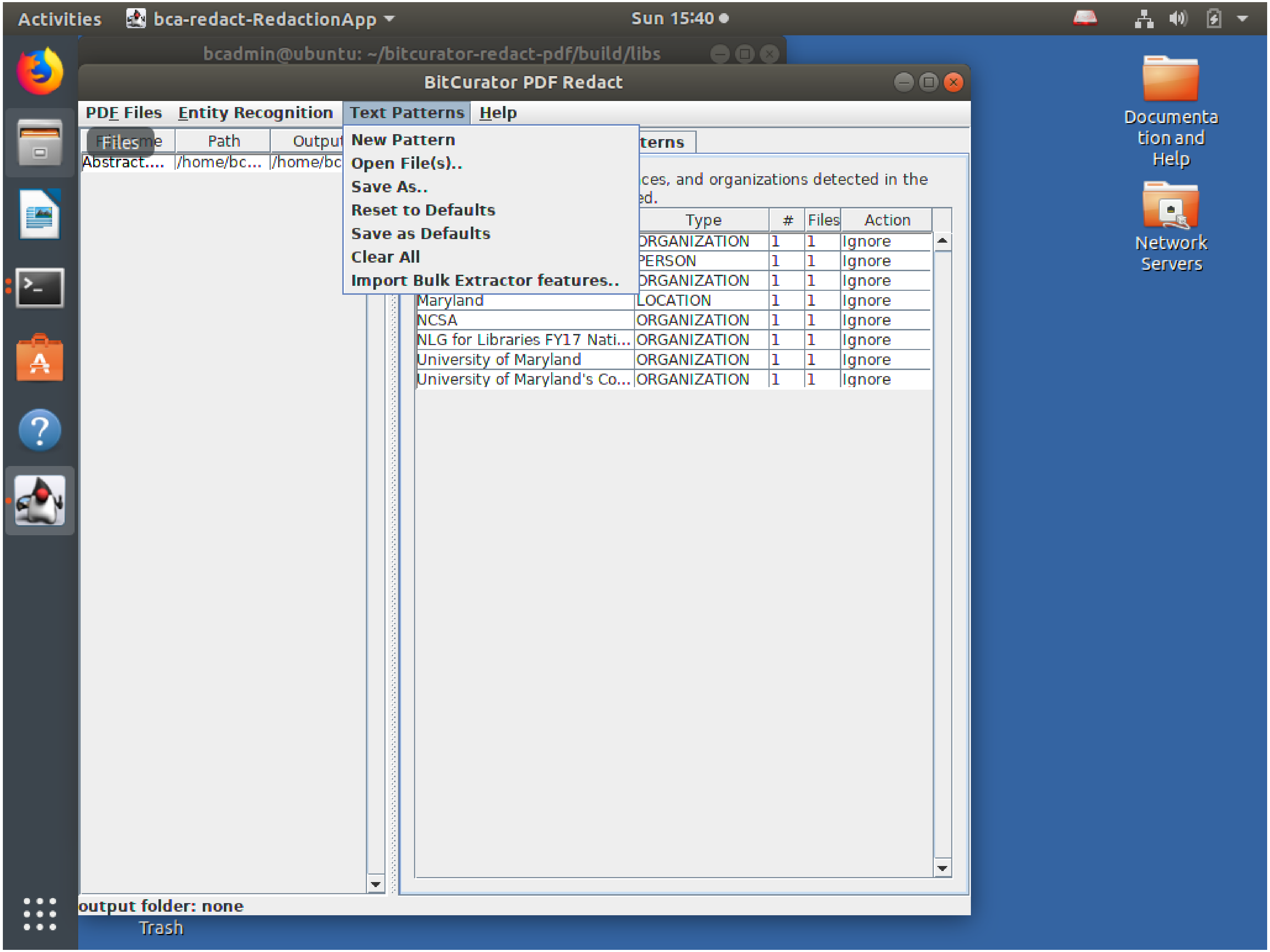


Filename	Path	Output
Abstract....	/home/bc...	/home/bc...

Named entities are people, places, and organizations detected in the text of PDF files you have added.

Entity Text	Type	#	Files	Action
Archive Analytics	ORGANIZATION	1	1	Ignore
Cassandra	PERSON	1	1	Ignore
Digital Curation Innovation...	ORGANIZATION	1	1	Ignore
Maryland	LOCATION	1	1	Ignore
NCSA	ORGANIZATION	1	1	Ignore
NLG for Libraries FY17 Nati...	ORGANIZATION	1	1	Ignore
University of Maryland	ORGANIZATION	1	1	Ignore
University of Maryland's Co...	ORGANIZATION	1	1	Ignore

output folder: none

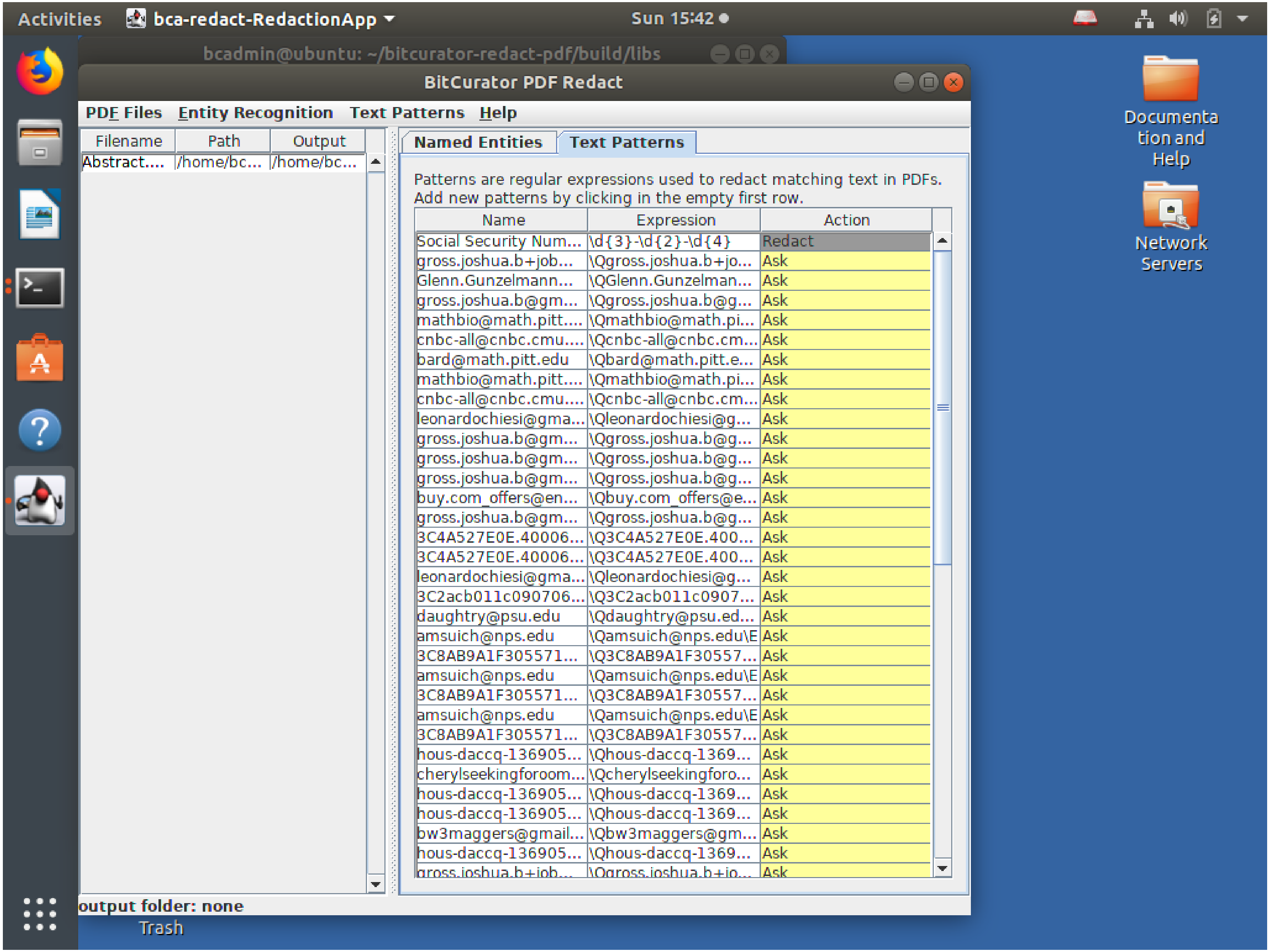


Filename	Path	Output
Abstract....	/home/bc...	/home/bc

ces, and organizations detected in the
ed.

	Type	#	Files	Action
	ORGANIZATION	1	1	Ignore
	PERSON	1	1	Ignore
	ORGANIZATION	1	1	Ignore
Maryland	LOCATION	1	1	Ignore
NCSA	ORGANIZATION	1	1	Ignore
NLG for Libraries FY17 Nati...	ORGANIZATION	1	1	Ignore
University of Maryland	ORGANIZATION	1	1	Ignore
University of Maryland's Co...	ORGANIZATION	1	1	Ignore

output folder: none

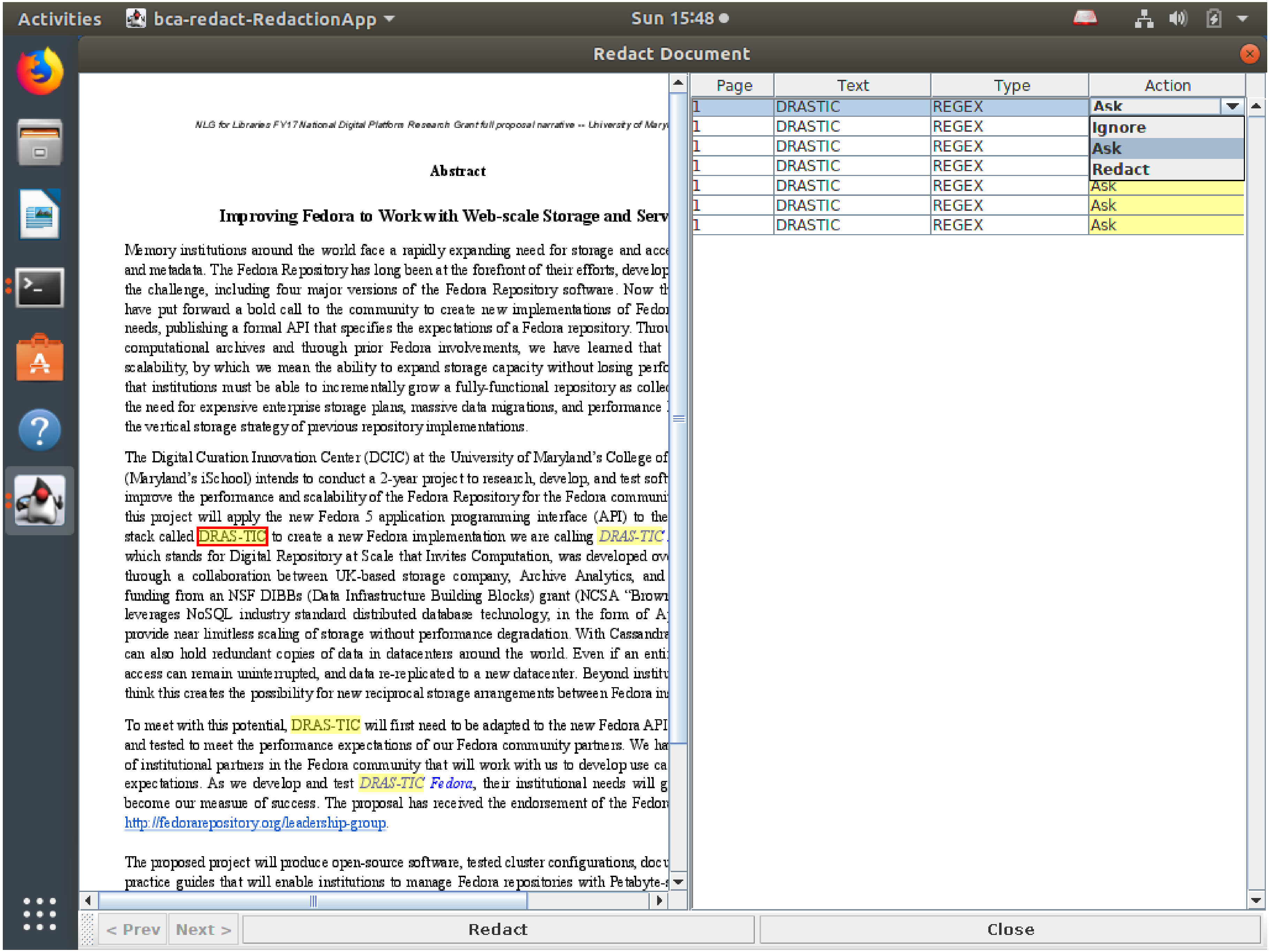


Filename	Path	Output
Abstract....	/home/bc...	/home/bc...

Patterns are regular expressions used to redact matching text in PDFs.
Add new patterns by clicking in the empty first row.

Name	Expression	Action
Social Security Num...	\d{3}-\d{2}-\d{4}	Redact
gross.joshua.b+job...	\Qgross.joshua.b+jo...	Ask
Glenn.Gunzelmann...	\QGlenn.Gunzelman...	Ask
gross.joshua.b@gm...	\Qgross.joshua.b@g...	Ask
mathbio@math.pitt....	\Qmathbio@math.pi...	Ask
cnbc-all@cnbc.cmu....	\Qcnbc-all@cnbc.cm...	Ask
bard@math.pitt.edu	\Qbard@math.pitt.e...	Ask
mathbio@math.pitt....	\Qmathbio@math.pi...	Ask
cnbc-all@cnbc.cmu....	\Qcnbc-all@cnbc.cm...	Ask
leonardochiesi@gma...	\Qleonardochiesi@g...	Ask
gross.joshua.b@gm...	\Qgross.joshua.b@g...	Ask
gross.joshua.b@gm...	\Qgross.joshua.b@g...	Ask
gross.joshua.b@gm...	\Qgross.joshua.b@g...	Ask
buy.com_offers@en...	\Qbuy.com_offers@e...	Ask
gross.joshua.b@gm...	\Qgross.joshua.b@g...	Ask
3C4A527E0E.40006...	\Q3C4A527E0E.400...	Ask
3C4A527E0E.40006...	\Q3C4A527E0E.400...	Ask
leonardochiesi@gma...	\Qleonardochiesi@g...	Ask
3C2acb011c090706...	\Q3C2acb011c0907...	Ask
daughtry@psu.edu	\Qdaughtry@psu.ed...	Ask
amsuich@nps.edu	\Qamsuich@nps.edu\E	Ask
3C8AB9A1F305571...	\Q3C8AB9A1F30557...	Ask
amsuich@nps.edu	\Qamsuich@nps.edu\E	Ask
3C8AB9A1F305571...	\Q3C8AB9A1F30557...	Ask
amsuich@nps.edu	\Qamsuich@nps.edu\E	Ask
3C8AB9A1F305571...	\Q3C8AB9A1F30557...	Ask
hous-daccq-136905...	\Qhous-daccq-1369...	Ask
cherylseekingforoom...	\Qcherylseekingforo...	Ask
hous-daccq-136905...	\Qhous-daccq-1369...	Ask
hous-daccq-136905...	\Qhous-daccq-1369...	Ask
bw3maggers@gmail...	\Qbw3maggers@gm...	Ask
hous-daccq-136905...	\Qhous-daccq-1369...	Ask
gross.joshua.b+job...	\Qgross.joshua.b+jo...	Ask

output folder: none



NLG for Libraries FY17 National Digital Platform Research Grant full proposal narrative -- University of Mary

Abstract

Improving Fedora to Work with Web-scale Storage and Serv

Memory institutions around the world face a rapidly expanding need for storage and acce
and metadata. The Fedora Repository has long been at the forefront of their efforts, develop
the challenge, including four major versions of the Fedora Repository software. Now th
have put forward a bold call to the community to create new implementations of Fedor
needs, publishing a formal API that specifies the expectations of a Fedora repository. Throu
computational archives and through prior Fedora involvements, we have learned that
scalability, by which we mean the ability to expand storage capacity without losing perfo
that institutions must be able to incrementally grow a fully-functional repository as colle
the need for expensive enterprise storage plans, massive data migrations, and performance l
the vertical storage strategy of previous repository implementations.

The Digital Curation Innovation Center (DCIC) at the University of Maryland's College of
(Maryland's iSchool) intends to conduct a 2-year project to research, develop, and test soft
improve the performance and scalability of the Fedora Repository for the Fedora communi
this project will apply the new Fedora 5 application programming interface (API) to the
stack called DRAS-TIC to create a new Fedora implementation we are calling *DRAS-TIC*.
which stands for Digital Repository at Scale that Invites Computation, was developed ov
through a collaboration between UK-based storage company, Archive Analytics, and
funding from an NSF DIBBs (Data Infrastructure Building Blocks) grant (NCSA "Brown
leverages NoSQL industry standard distributed database technology, in the form of A
provide near limitless scaling of storage without performance degradation. With Cassandra
can also hold redundant copies of data in datacenters around the world. Even if an enti
access can remain uninterrupted, and data re-replicated to a new datacenter. Beyond institu
think this creates the possibility for new reciprocal storage arrangements between Fedora in

To meet with this potential, DRAS-TIC will first need to be adapted to the new Fedora API
and tested to meet the performance expectations of our Fedora community partners. We ha
of institutional partners in the Fedora community that will work with us to develop use ca
expectations. As we develop and test *DRAS-TIC Fedora*, their institutional needs will g
become our measure of success. The proposal has received the endorsement of the Fedor
http://fedorarepository.org/leadership-group.

The proposed project will produce open-source software, tested cluster configurations, docu
practice guides that will enable institutions to manage Fedora repositories with Petabyte-s

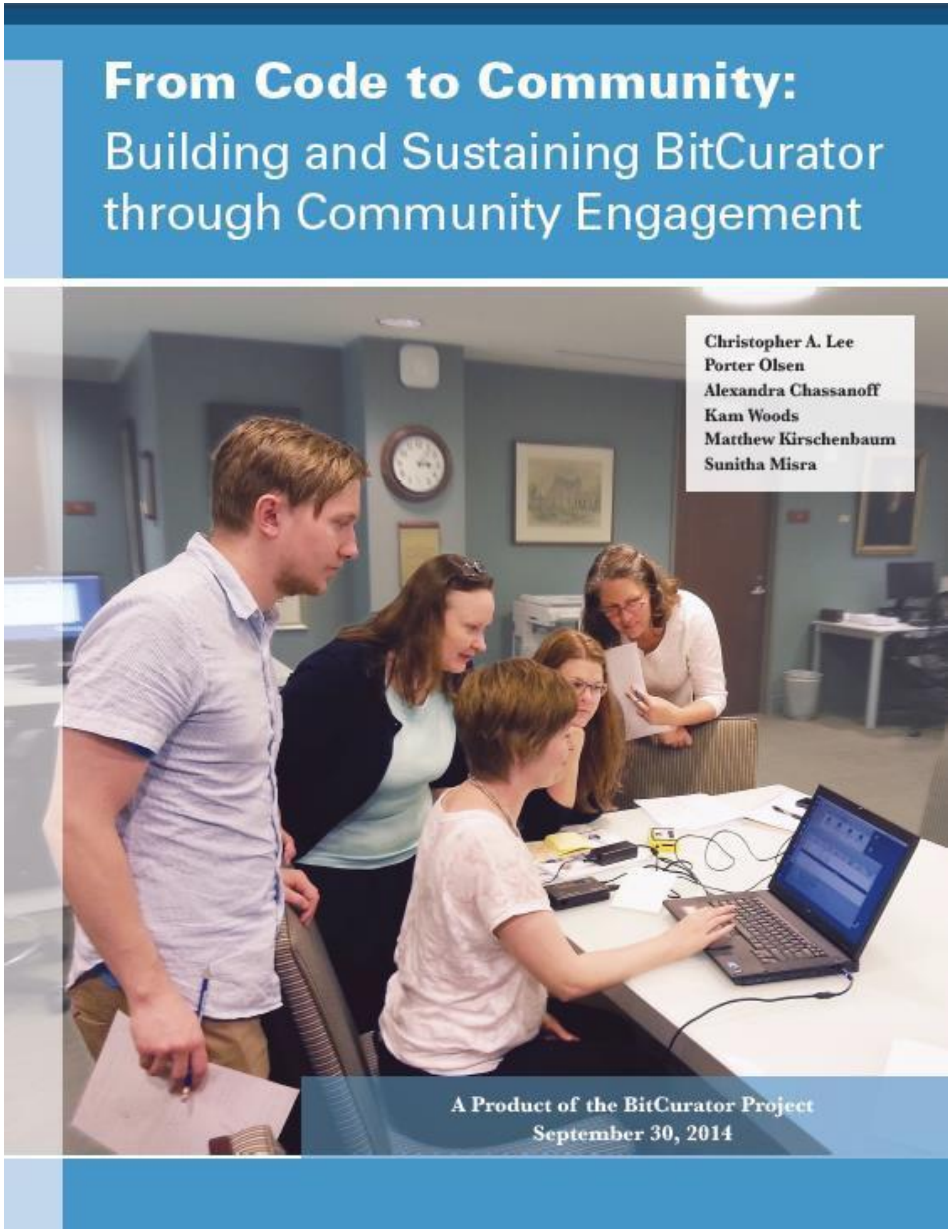
Page	Text	Type	Action
1	DRASTIC	REGEX	Ask
1	DRASTIC	REGEX	Ignore
1	DRASTIC	REGEX	Ask
1	DRASTIC	REGEX	Redact
1	DRASTIC	REGEX	Ask
1	DRASTIC	REGEX	Ask
1	DRASTIC	REGEX	Ask

Other Functionality to Meet Identified User Needs:

Function	Tool(s)
Identify duplicate files	FSLint
Characterize files	FIDO, Siegfried, Brunnhilde
Scan for viruses	ClamTK
Examine, copy and extract information from old Mac disks	HFS Utilities (including HFS Explorer)
Capture AV file metadata	MediaInfo, FFProbe
Extract text from older binary (.doc) Word files	antiword
Read contents of Mircosoft Outlook PST files	readpst
Examine embedded header information in images	pyExifToolGUI
Generate images of problematic disks or particular disk types (I addition to Guymager	dd, dcfldd, ddrescue, cdrdao (for audio CDs)
Extract and analyze data from Windows Registry files	regripper
Identify files that are partially similar but not identical	sdhash, ssdeep
Package files for storage and/or transfer	BagIt (Java) library, Bagger
File preview (left-click on file then hit space bar)	gnome-sushi

Other Functionality to Meet Identified User Needs (Continued):

Function	Tool(s)
Play and examine metadata from AV media files	VLC media player
Damaged/lost partition recovery	TestDisk
Damaged/lost file recovery	PhotoRec
Identify the filesystem on a disk	disktype
Index and search for keywords in documents	recoll
Find blacklist data by using hashes calculated from hash blocks	hashdb
Generate hashes of files and blocks	GTK Hash, md5deep, md5sum
Compare hashes of files to hashes in the National Software Reference Library (NSRL) of known system files	nsrllookup
View and edit bytestreams (hex editor)	Bless Hex Editor, GHex



From Code to Community:

Building and Sustaining BitCurator through Community Engagement

Christopher A. Lee
Porter Olsen
Alexandra Chassanoff
Kam Woods
Matthew Kirschenbaum
Sunitha Misra

A Product of the BitCurator Project
September 30, 2014

http://www.bitcurator.net/wp-content/uploads/2014/11/code-to-community.pdf

Storage Media Acquisition and Handling Profile for Digital Repositories*

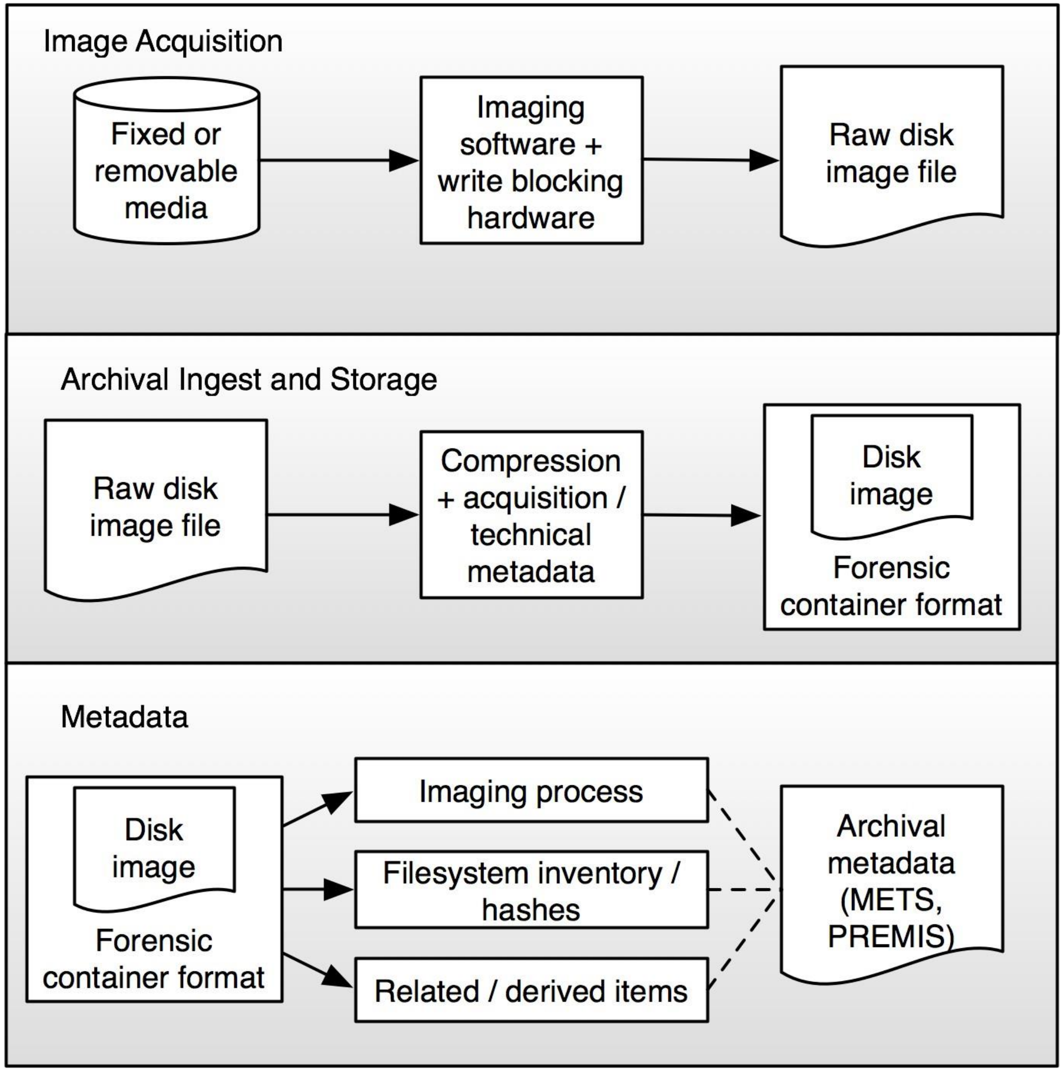
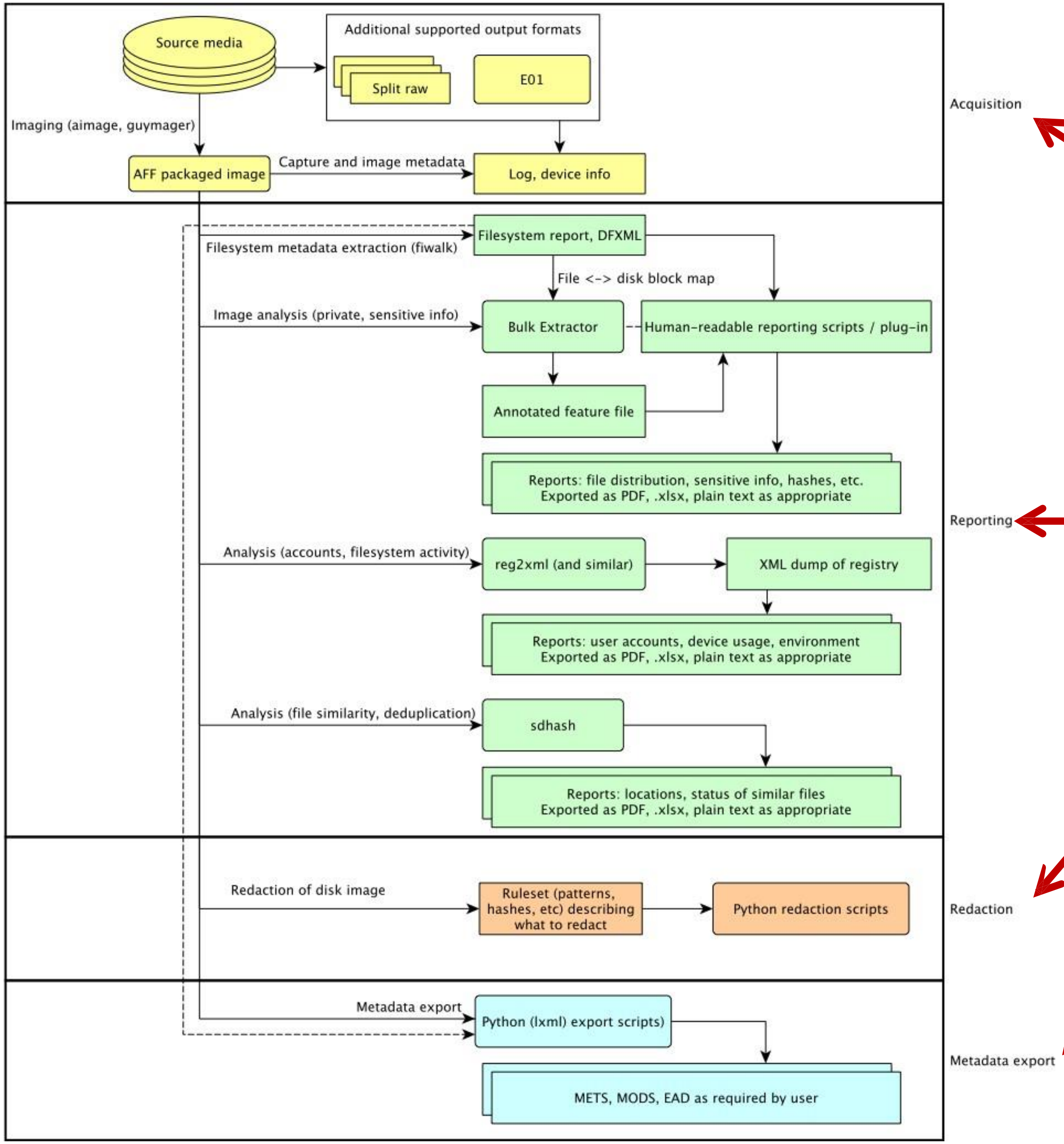


Image Acquisition
Fixed or removable media
Imaging software + write blocking hardware
Raw disk image file
Archival Ingest and Storage
Raw disk image file
Compression + acquisition / technical metadata
Disk image
Forensic container format
Metadata
Disk image
Forensic container format
Imaging process
Filesystem inventory / hashes
Related / derived items
Archival metadata (METS, PREMIS)

*Woods, Kam, Christopher A. Lee, and Simson Garfinkel. "Extending Digital Repository Architectures to Support Disk Image Preservation and Access." In *JCDL '11: Proceeding of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, 57-66. New York, NY: ACM Press, 2011.

BitCurator-Supported Workflow


Source media
Additional supported output formats
Split raw
E01
Acquisition
Imaging (aimage, guymager)
AFF packaged image
Capture and image metadata
Log, device info
Filesystem report, DFXML
Filesystem metadata extraction (fiwalk)
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

Five Sources of Workflow Examples

Martin J. Gengenbach, "'The Way We Do it Here': Mapping Digital Forensics Workflows in Collecting Institutions," A Master's Paper for the M.S. in L.S degree. August 2012. http://digitalcurationexchange.org/system/files/gengenbach-forensic-workflows-2012.pdf

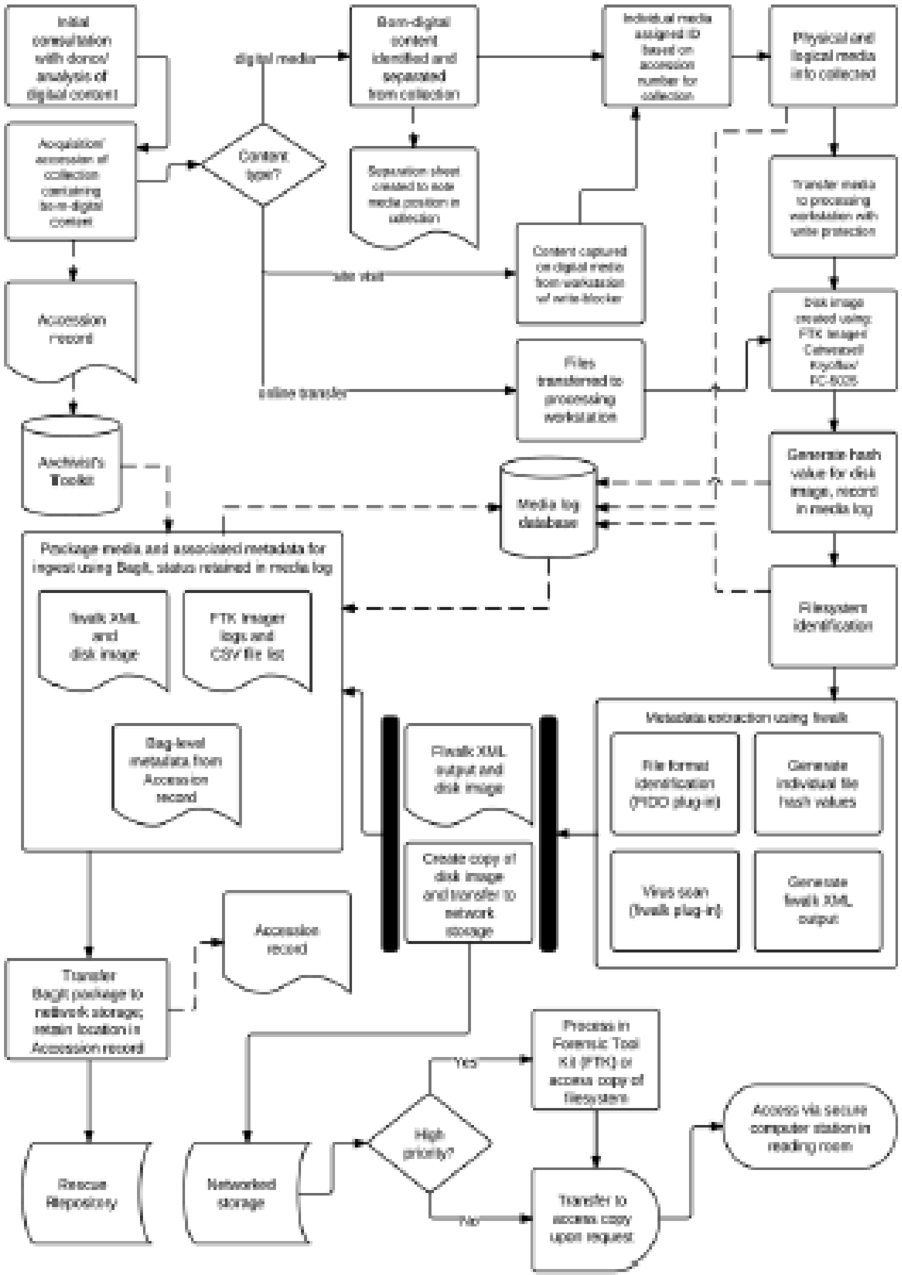
AIMS Work Group, "AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship," January 2012. http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf

Digital Sustainability Lab – Massachusetts Institute of Technology
http://www.dpworkshop.org/sites/default/files/DCM-Pipeline_28Apr2015.pdf

Workflows, BitCurator Consortium
https://bitcuratorconsortium.org/workflows

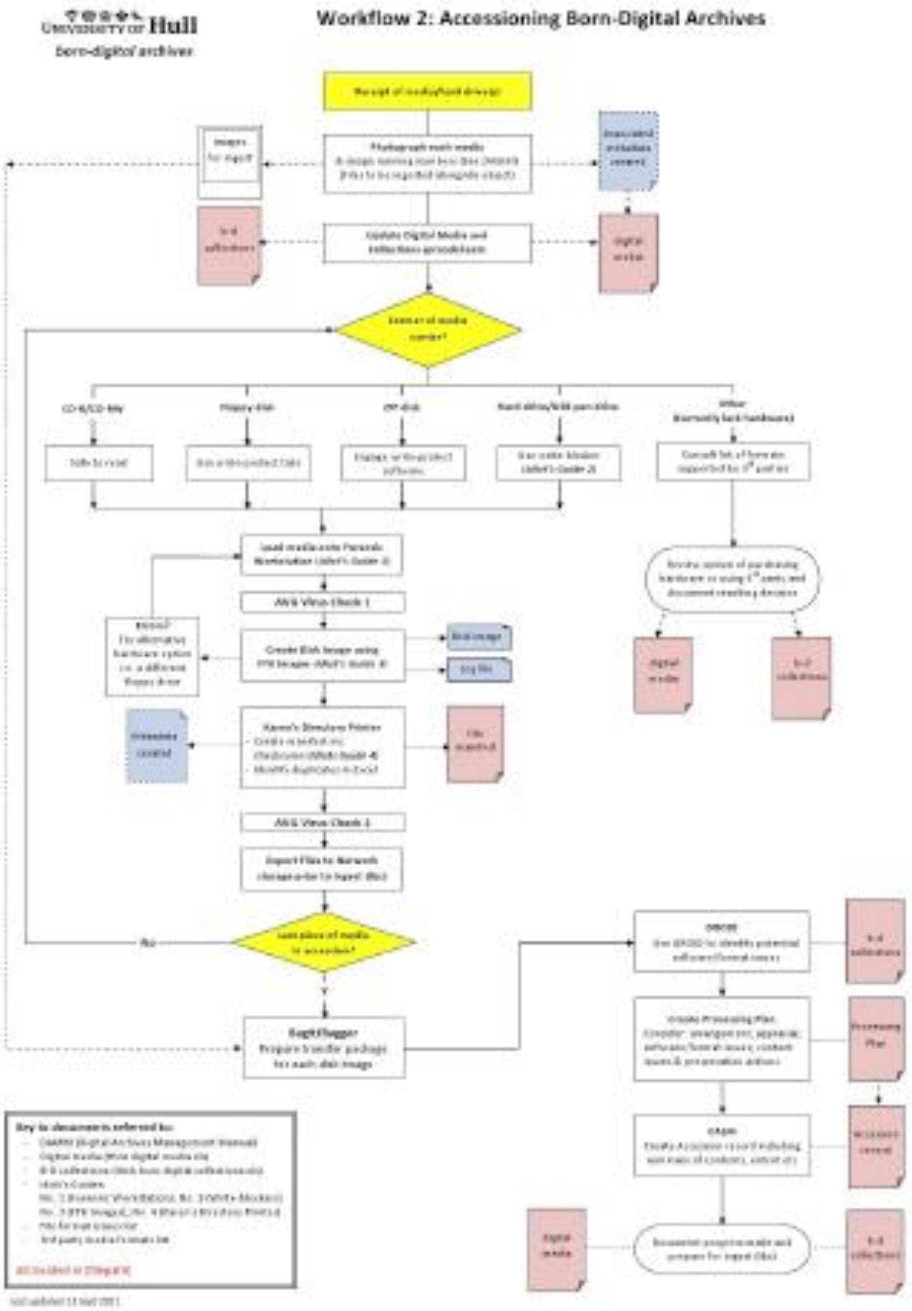
OSSArcFlow Project - https://educopia.org/research/ossarcflow

Figure 1. Beinecke Rare Book and Manuscript Library, Yale University

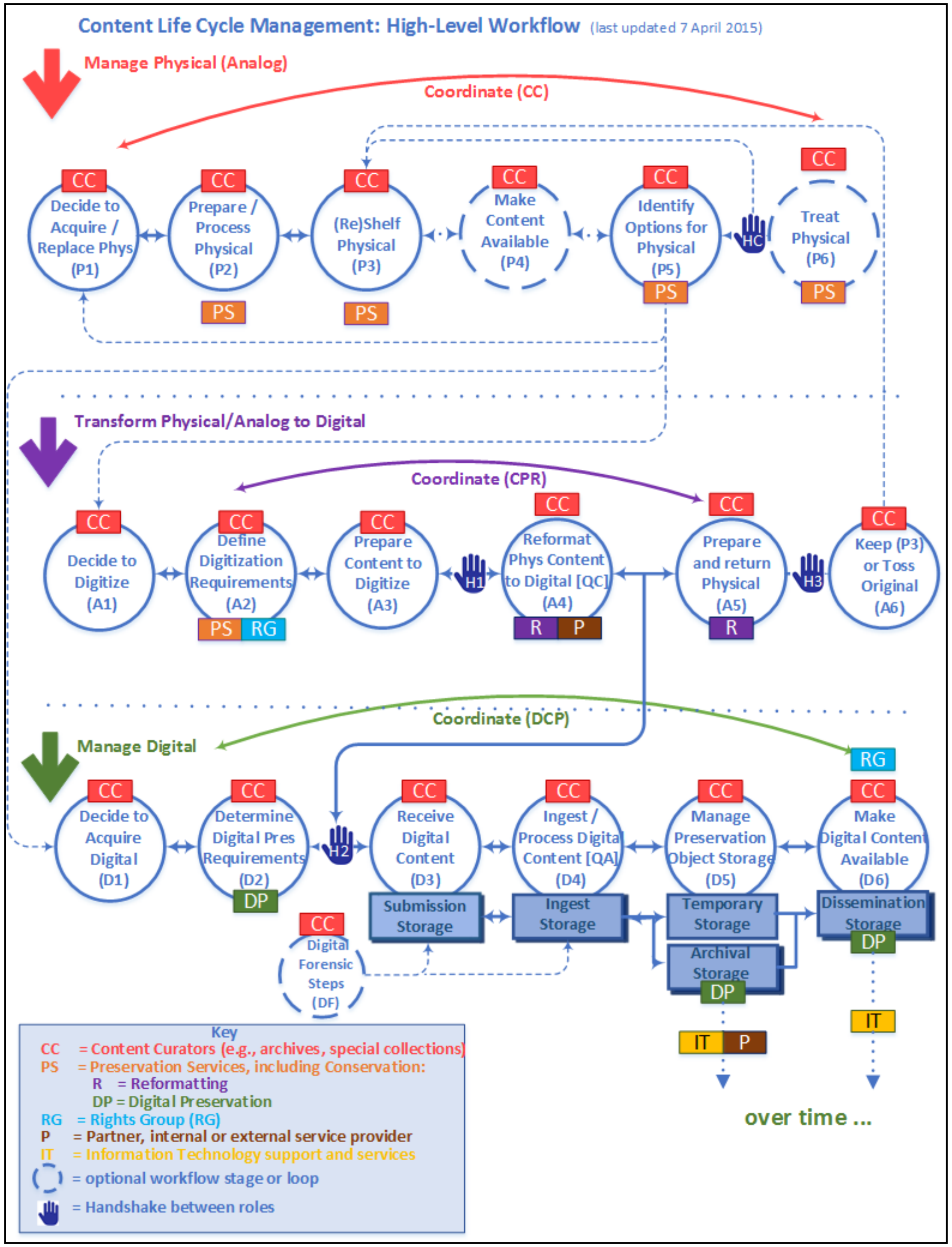


Initial consultation with donor/ analysis of digital content
Acquisition/ accession of collection containing born-digital content
Accession record
Archivist's Toolkit
Content type?
digital media
site visit
online transfer
Born-digital content identified and separated from collection
Separation sheet created to note media position in collection
Individual media assigned ID based on accession number for collection
Physical and logical media into collected
Content captured on digital media from workstation w/ write-blocker
Files transferred to processing workstation
Transfer media to processing workstation with write protection
Disk image created using: FTK Imager/ Catweasel/ KryoFlux/ FC-5025
Generate hash value for disk image, record in media log
Media log database
Filesystem identification
Package media and associated metadata for ingest using BagIt, status retained in media log
fiwalk XML and disk image
FTK Imager logs and CSV file list
Bag-level metadata from Accession record
Fiwalk XML output and disk image
Create copy of disk image and transfer to network storage
Metadata extraction using fiwalk
File format identification (PIDO plug-in)
Generate individual file hash values
Virus scan (fiwalk plug-in)
Generate fiwalk XML output
Transfer BagIt package to network storage; retain location in Accession record
Accession record
Rescue Repository
Networked storage
High priority?
Yes
No
Process in Forensic Tool Kit (FTK) or access copy of filesystem
Transfer to access copy upon request
Access via secure computer station in reading room

Martin J. Gengenbach, "'The Way We Do it Here': Mapping Digital Forensics Workflows in Collecting Institutions," A Master's Paper for the M.S. in L.S degree. August, 2012.



AIMS Work Group, “AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship,” January 2012.


Content Life Cycle Management: High-Level Workflow (last updated 7 April 2015)
Manage Physical (Analog)
Coordinate (CC)
CC
Decide to Acquire / Replace Phys (P1)
CC
Prepare / Process Physical (P2)
PS
CC
(Re)Shelf Physical (P3)
PS
CC
Make Content Available (P4)
CC
Identify Options for Physical (P5)
PS
HC
CC
Treat Physical (P6)
PS
Transform Physical/Analog to Digital
Coordinate (CPR)
CC
Decide to Digitize (A1)
CC
Define Digitization Requirements (A2)
PS
RG
CC
Prepare Content to Digitize (A3)
H1
CC
Reformat Phys Content to Digital [QC] (A4)
R
P
CC
Prepare and return Physical (A5)
R
H3
CC
Keep (P3) or Toss Original (A6)
Manage Digital
Coordinate (DCP)
RG
CC
Decide to Acquire Digital (D1)
CC
Determine Digital Pres Requirements (D2)
DP
H2
CC
Receive Digital Content (D3)
CC
Ingest / Process Digital Content [QA] (D4)
CC
Manage Preservation Object Storage (D5)
CC
Make Digital Content Available (D6)
Submission Storage
Ingest Storage
Temporary Storage
Archival Storage
DP
Dissemination Storage
DP
CC
Digital Forensic Steps (DF)
IT
P
IT
over time ...

Key

CC = Content Curators (e.g., archives, special collections)

PS = Preservation Services, including Conservation:

R = Reformatting

DP = Digital Preservation

RG = Rights Group (RG)

P = Partner, internal or external service provider

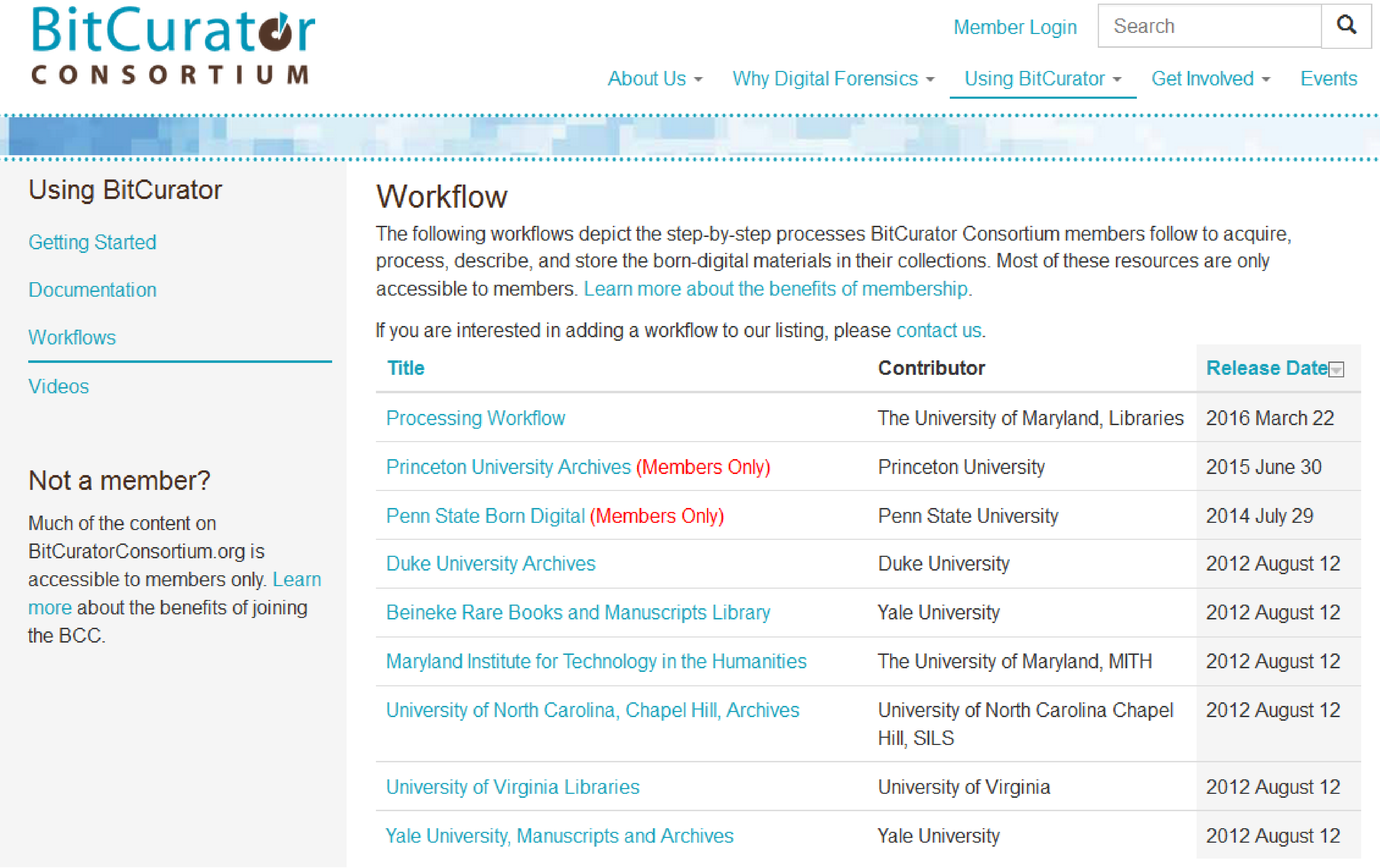
IT = Information Technology support and services

◌ = optional workflow stage or loop

✋ = Handshake between roles

Kari Smith, Massachusetts Institute of Technology.

http://www.dpworkshop.org/sites/default/files/DCM-Pipeline_28Apr2015.pdf



Using BitCurator

- Getting Started
- Documentation
- Workflows
- Videos

Not a member?

Much of the content on BitCuratorConsortium.org is accessible to members only. Learn more about the benefits of joining the BCC.

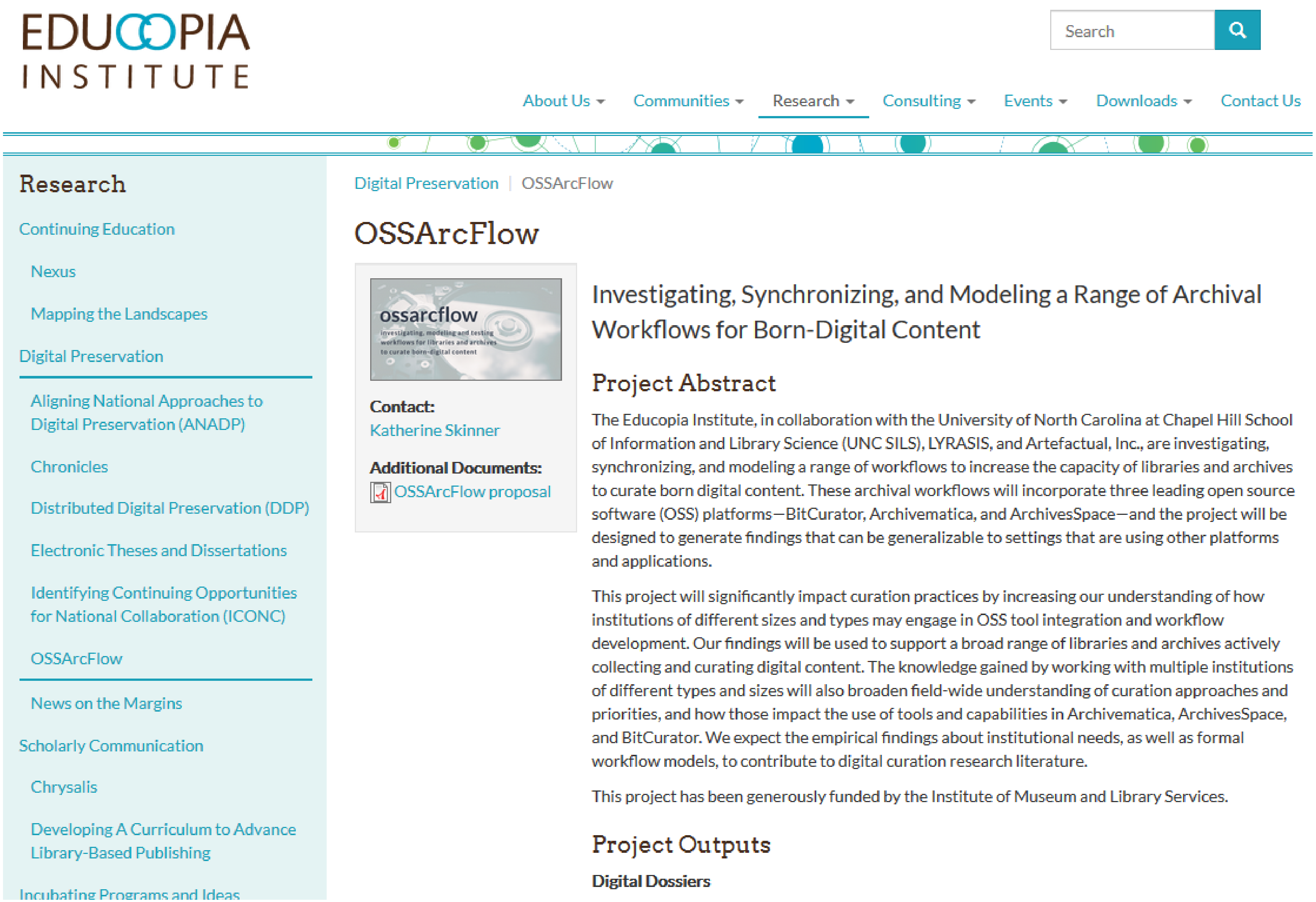
Workflow

The following workflows depict the step-by-step processes BitCurator Consortium members follow to acquire, process, describe, and store the born-digital materials in their collections. Most of these resources are only accessible to members. Learn more about the benefits of membership.

If you are interested in adding a workflow to our listing, please contact us.

Title	Contributor	Release Date
Processing Workflow	The University of Maryland, Libraries	2016 March 22
Princeton University Archives (Members Only)	Princeton University	2015 June 30
Penn State Born Digital (Members Only)	Penn State University	2014 July 29
Duke University Archives	Duke University	2012 August 12
Beineke Rare Books and Manuscripts Library	Yale University	2012 August 12
Maryland Institute for Technology in the Humanities	The University of Maryland, MITH	2012 August 12
University of North Carolina, Chapel Hill, Archives	University of North Carolina Chapel Hill, SILS	2012 August 12
University of Virginia Libraries	University of Virginia	2012 August 12
Yale University, Manuscripts and Archives	Yale University	2012 August 12

https://bitcuratorconsortium.org/workflows


EDUCOPIA INSTITUTE
Search
About Us
Communities
Research
Consulting
Events
Downloads
Contact Us
Research
Continuing Education
Nexus
Mapping the Landscapes
Digital Preservation
Aligning National Approaches to Digital Preservation (ANADP)
Chronicles
Distributed Digital Preservation (DDP)
Electronic Theses and Dissertations
Identifying Continuing Opportunities for National Collaboration (ICONC)
OSSArcFlow
News on the Margins
Scholarly Communication
Chrysalis
Developing A Curriculum to Advance Library-Based Publishing
Incubating Programs and Ideas
Digital Preservation | OSSArcFlow
ossarcflow
investigating, modeling and testing workflows for libraries and archives to curate born-digital content

OSSArcFlow

Contact:
Katherine Skinner

Additional Documents:
OSSArcFlow proposal

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

Project Abstract

The Educopia Institute, in collaboration with the University of North Carolina at Chapel Hill School of Information and Library Science (UNC SILS), LYRASIS, and Artefactual, Inc., are investigating, synchronizing, and modeling a range of workflows to increase the capacity of libraries and archives to curate born digital content. These archival workflows will incorporate three leading open source software (OSS) platforms—BitCurator, Archivematica, and ArchivesSpace—and the project will be designed to generate findings that can be generalizable to settings that are using other platforms and applications.

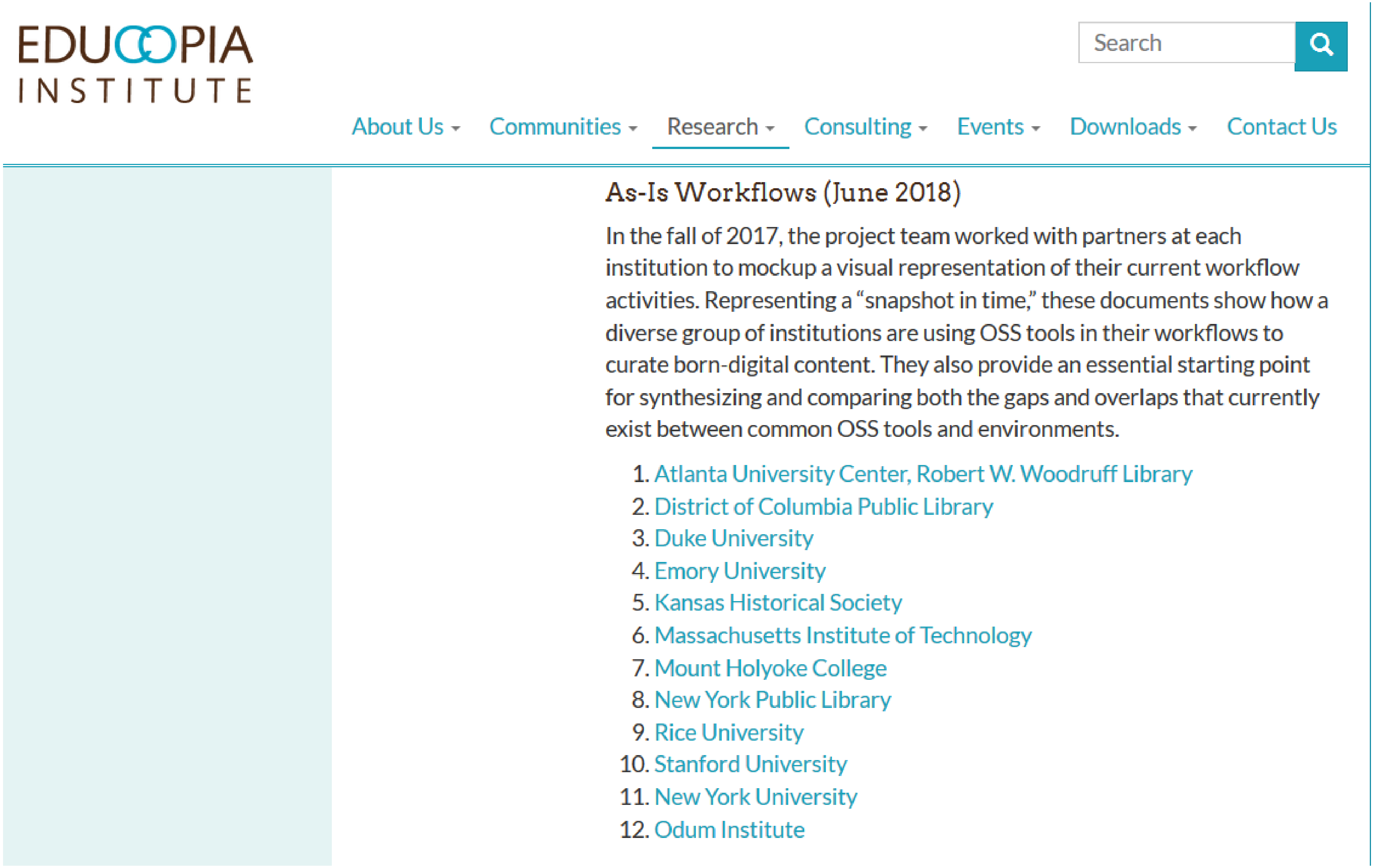
This project will significantly impact curation practices by increasing our understanding of how institutions of different sizes and types may engage in OSS tool integration and workflow development. Our findings will be used to support a broad range of libraries and archives actively collecting and curating digital content. The knowledge gained by working with multiple institutions of different types and sizes will also broaden field-wide understanding of curation approaches and priorities, and how those impact the use of tools and capabilities in Archivematica, ArchivesSpace, and BitCurator. We expect the empirical findings about institutional needs, as well as formal workflow models, to contribute to digital curation research literature.

This project has been generously funded by the Institute of Museum and Library Services.

Project Outputs

Digital Dossiers

https://educopia.org/research/ossarcflow



As-Is Workflows (June 2018)

In the fall of 2017, the project team worked with partners at each institution to mockup a visual representation of their current workflow activities. Representing a "snapshot in time," these documents show how a diverse group of institutions are using OSS tools in their workflows to curate born-digital content. They also provide an essential starting point for synthesizing and comparing both the gaps and overlaps that currently exist between common OSS tools and environments.

1. Atlanta University Center, Robert W. Woodruff Library
2. District of Columbia Public Library
3. Duke University
4. Emory University
5. Kansas Historical Society
6. Massachusetts Institute of Technology
7. Mount Holyoke College
8. New York Public Library
9. Rice University
10. Stanford University
11. New York University
12. Odum Institute

https://educopia.org/research/ossarcflow

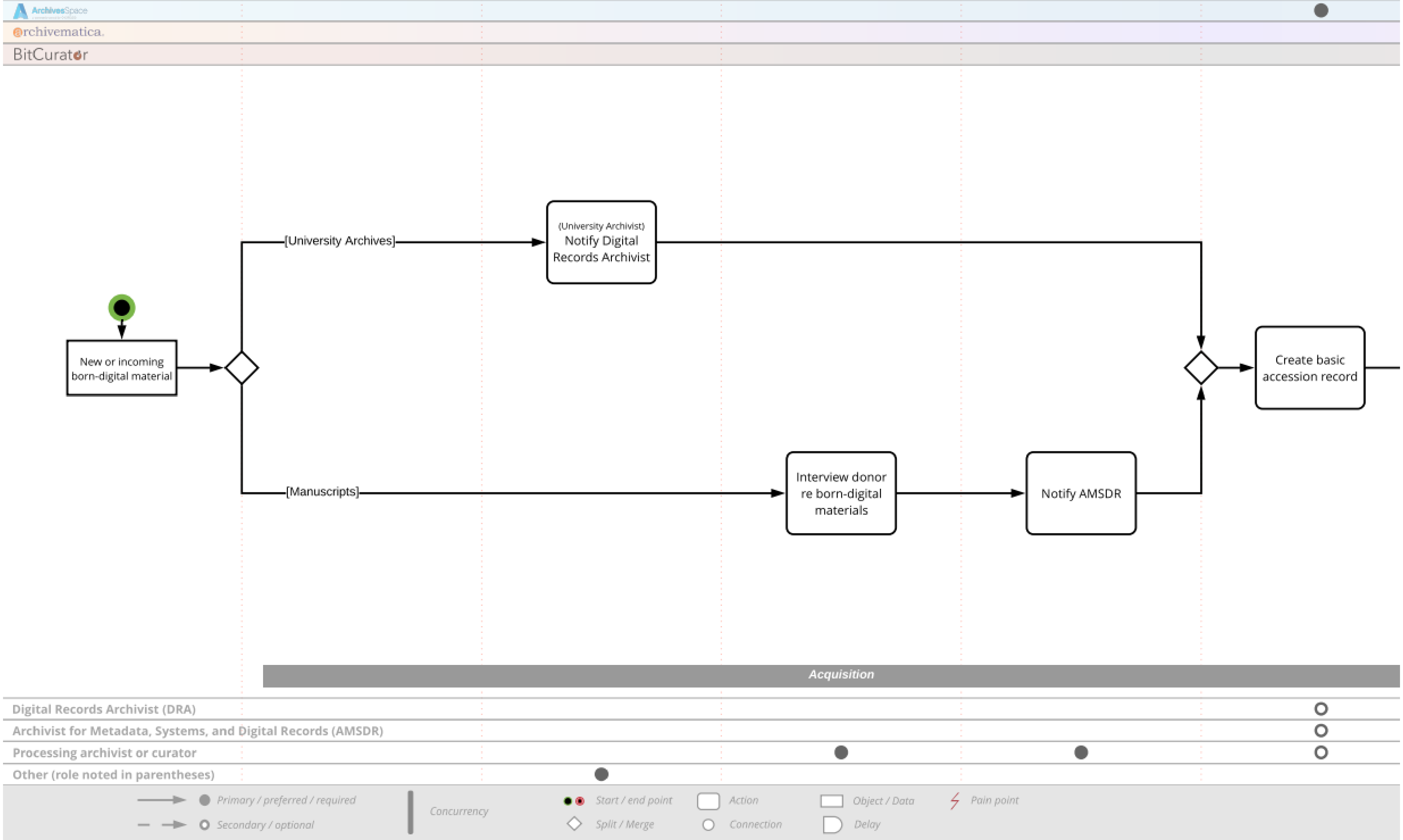
As-Is Digital Curation Workflow, August 2017

Born-digital content

Duke University

OssArcFlow Project

https://educopia.org/research/ossarcflow



ArchivesSpace

archivematica.

BitCurator

New or incoming born-digital material

[University Archives]

(University Archivist) Notify Digital Records Archivist

[Manuscripts]

Interview donor re born-digital materials

Notify AMSDR

Create basic accession record

Acquisition

Digital Records Archivist (DRA)

Archivist for Metadata, Systems, and Digital Records (AMSDR)

Processing archivist or curator

Other (role noted in parentheses)

● Primary / preferred / required

○ Secondary / optional

Concurrency

Start / end point

Split / Merge

Action

Connection

Object / Data

Delay

Pain point

End User Access Scenarios*

- Virtualization and emulation
- Mounting the original filesystem
- Accessing (but not mounting) disk images using forensics software
- Remote, dynamic access to disk image contents
- Cross-drive analysis

*Note: The first three were discussed earlier

BitCurator Access

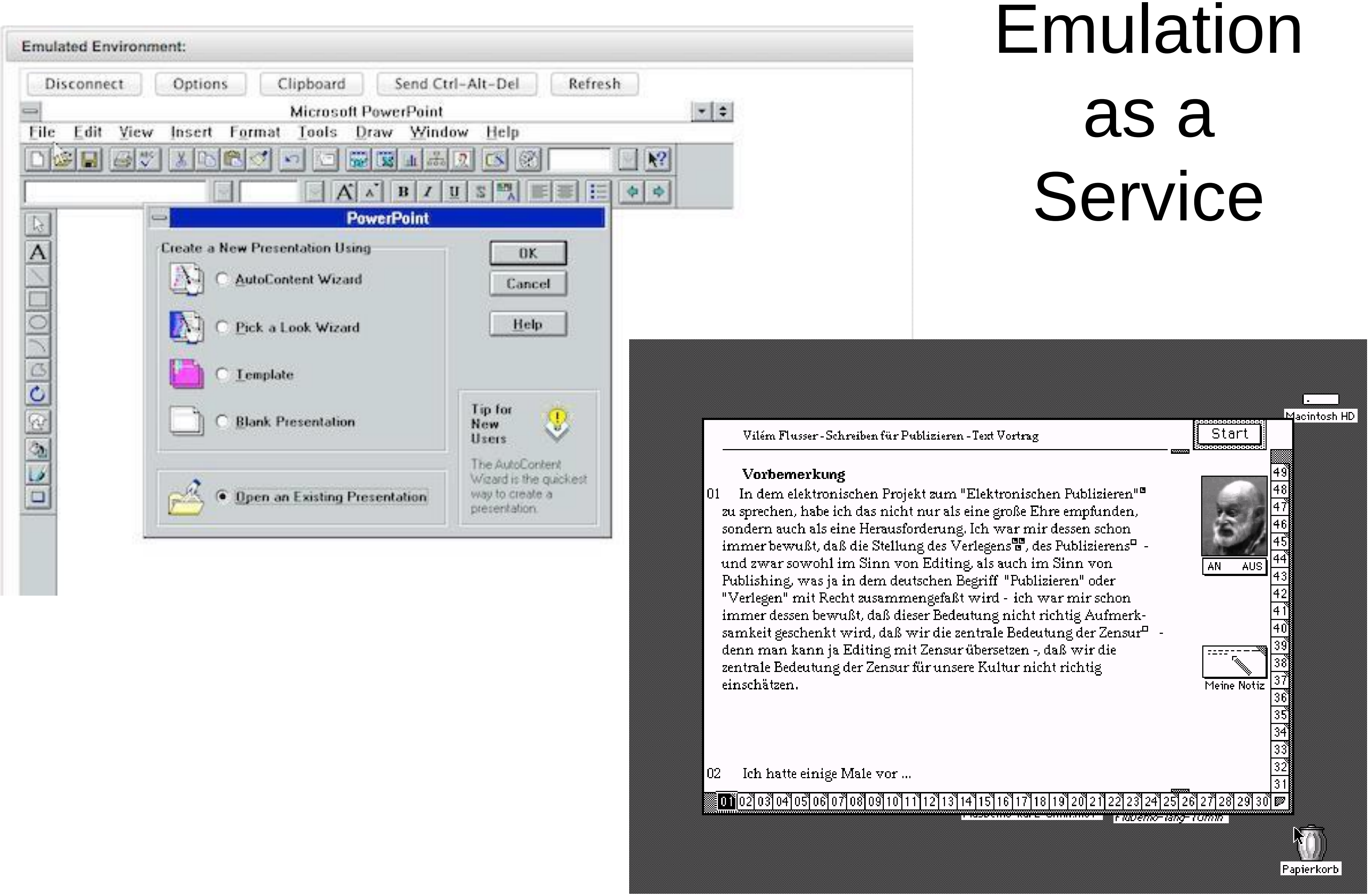
- Two-year project (October 1, 2014 – September 30, 2016) at School of Information and Library Science, University of North Carolina at Chapel Hill
- Funded by Andrew W. Mellon Foundation
- Developing open-source software to support access to disk images. Core areas of focus:
  - Tools and reusable libraries to support web access services for disk images
  - Analyzing contents of file systems and associated metadata
  - Redacting complex born-digital objects (disk images)
  - Emulated access to data from disk images

BitCurator Access Redaction Tools

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

https://github.com/bitcurator/bitcurator-access-redaction

Emulation as a Service



Create a New Presentation Using

○ AutoContent Wizard

○ Pick a Look Wizard

○ Template

○ Blank Presentation

● Open an Existing Presentation

Tip for New Users

The AutoContent Wizard is the quickest way to create a presentation.

Vilém Flusser - Schreiben für Publizieren - Text Vortrag

Vorbemerkung

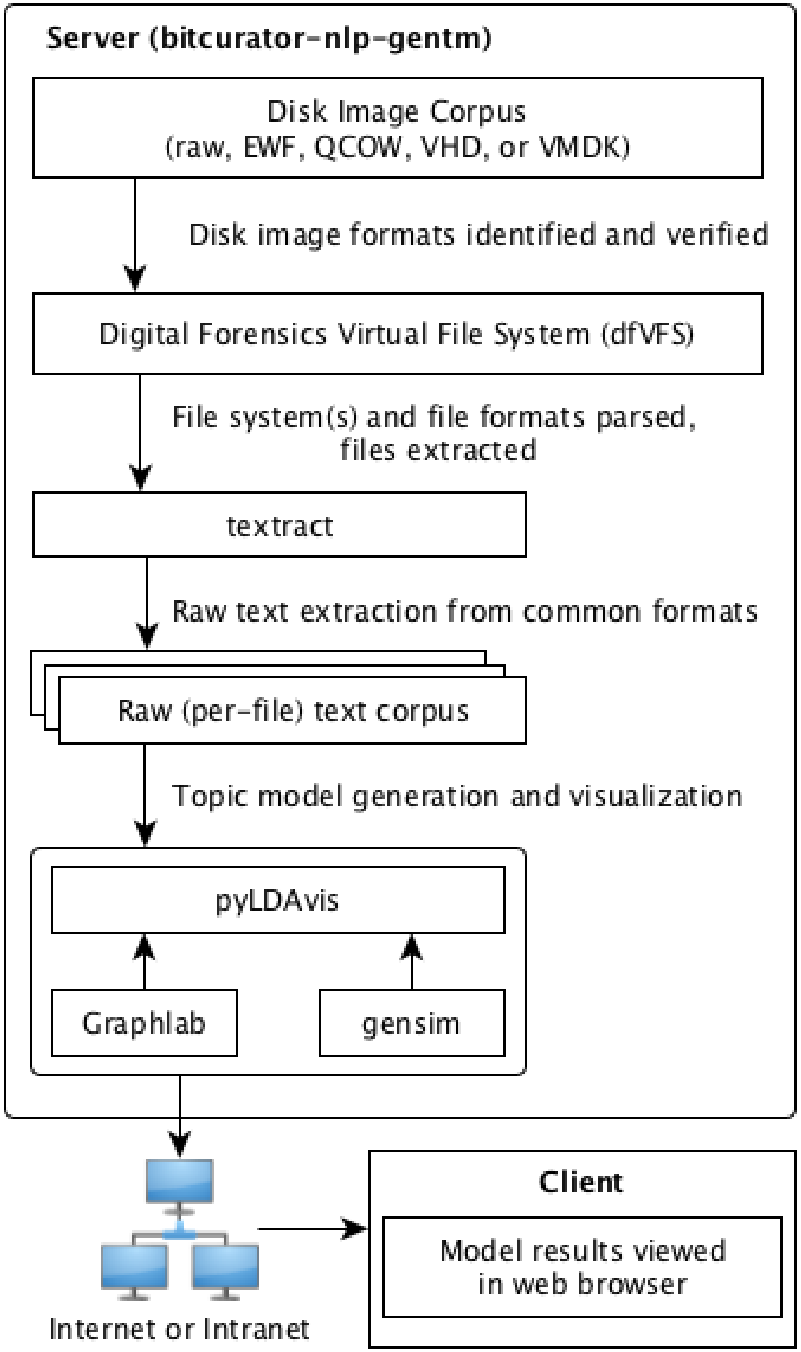
01 In dem elektronischen Projekt zum "Elektronischen Publizieren"□ zu sprechen, habe ich das nicht nur als eine große Ehre empfunden, sondern auch als eine Herausforderung. Ich war mir dessen schon immer bewußt, daß die Stellung des Verlegens□, des Publizierens□ - und zwar sowohl im Sinn von Editing, als auch im Sinn von Publishing, was ja in dem deutschen Begriff "Publizieren" oder "Verlegen" mit Recht zusammengefaßt wird - ich war mir schon immer dessen bewußt, daß dieser Bedeutung nicht richtig Aufmerksamkeit geschenkt wird, daß wir die zentrale Bedeutung der Zensur□ - denn man kann ja Editing mit Zensur übersetzen -, daß wir die zentrale Bedeutung der Zensur für unsere Kultur nicht richtig einschätzen.

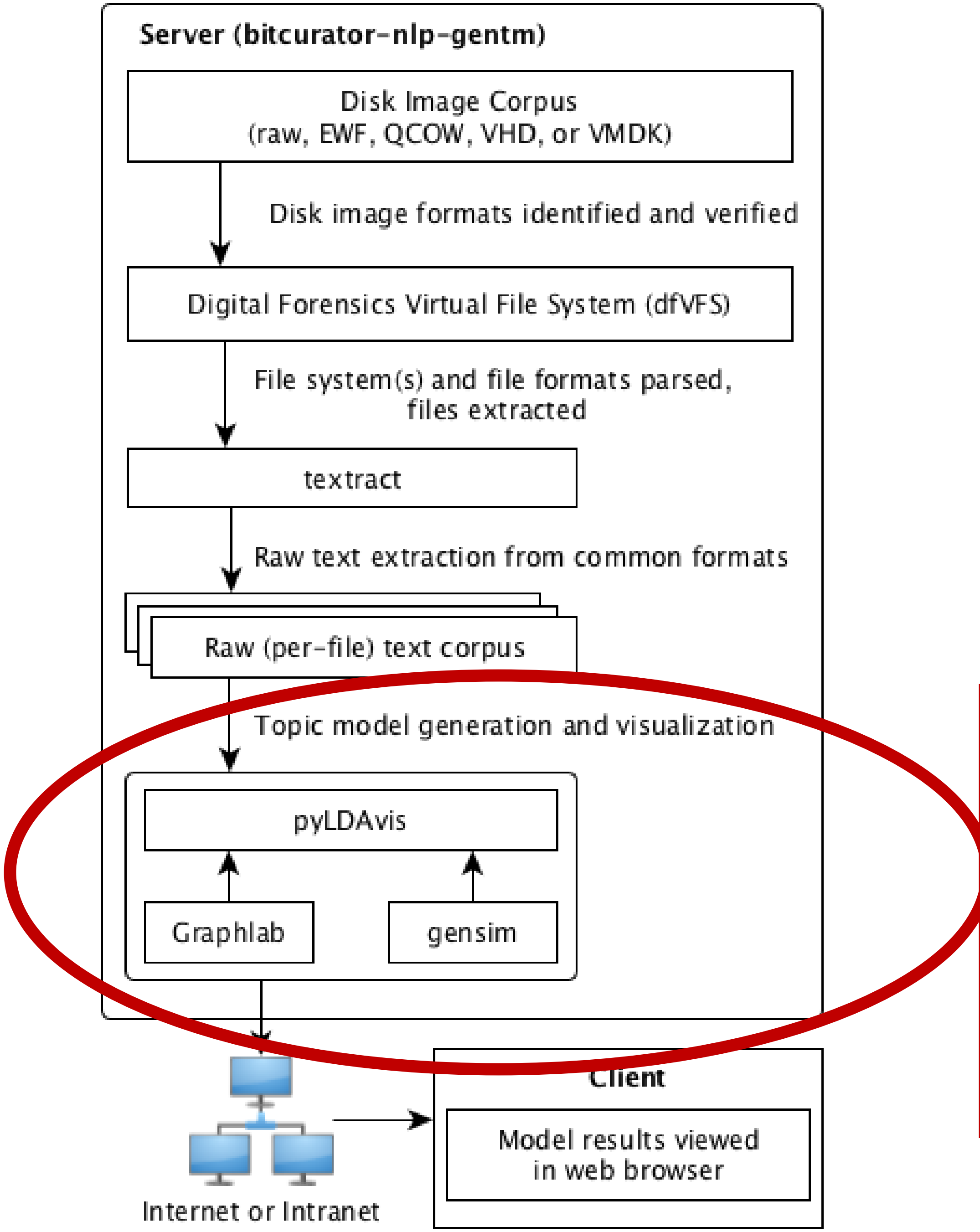
02 Ich hatte einige Male vor ...

http://bw-fla.uni-freiburg.de/demos.html

BitCurator NLP

- Funded by Andrew W. Mellon Foundation: October 1, 2016 – December 31, 2018
- Develop software for collecting institutions to extract, analyze, and produce reports on features of interest in text extracted from born-digital materials
- Use existing natural language processing software libraries to identify and report on those items likely to be relevant to ongoing preservation, information organization, and access activities
- May include entities (e.g. persons, places, and organizations), potential relationships among entities (e.g. appear together within documents or set of documents), and topic models to provide insight into how concepts are naturally clustered within the documents.


Server (bitcurator-nlp-gentm)
Disk Image Corpus
(raw, EWF, QCOW, VHD, or VMDK)
Disk image formats identified and verified
Digital Forensics Virtual File System (dfVFS)
File system(s) and file formats parsed,
files extracted
textract
Raw text extraction from common formats
Raw (per-file) text corpus
Topic model generation and visualization
pyLDAvis
Graphlab
gensim
Internet or Intranet
Client
Model results viewed
in web browser



Server (bitcurator-nlp-gentm)

Disk Image Corpus
(raw, EWF, QCOW, VHD, or VMDK)

Disk image formats identified and verified

Digital Forensics Virtual File System (dfVFS)

File system(s) and file formats parsed,
files extracted

textract

Raw text extraction from common formats

Raw (per-file) text corpus

Topic model generation and visualization

pyLDAvis

Graphlab

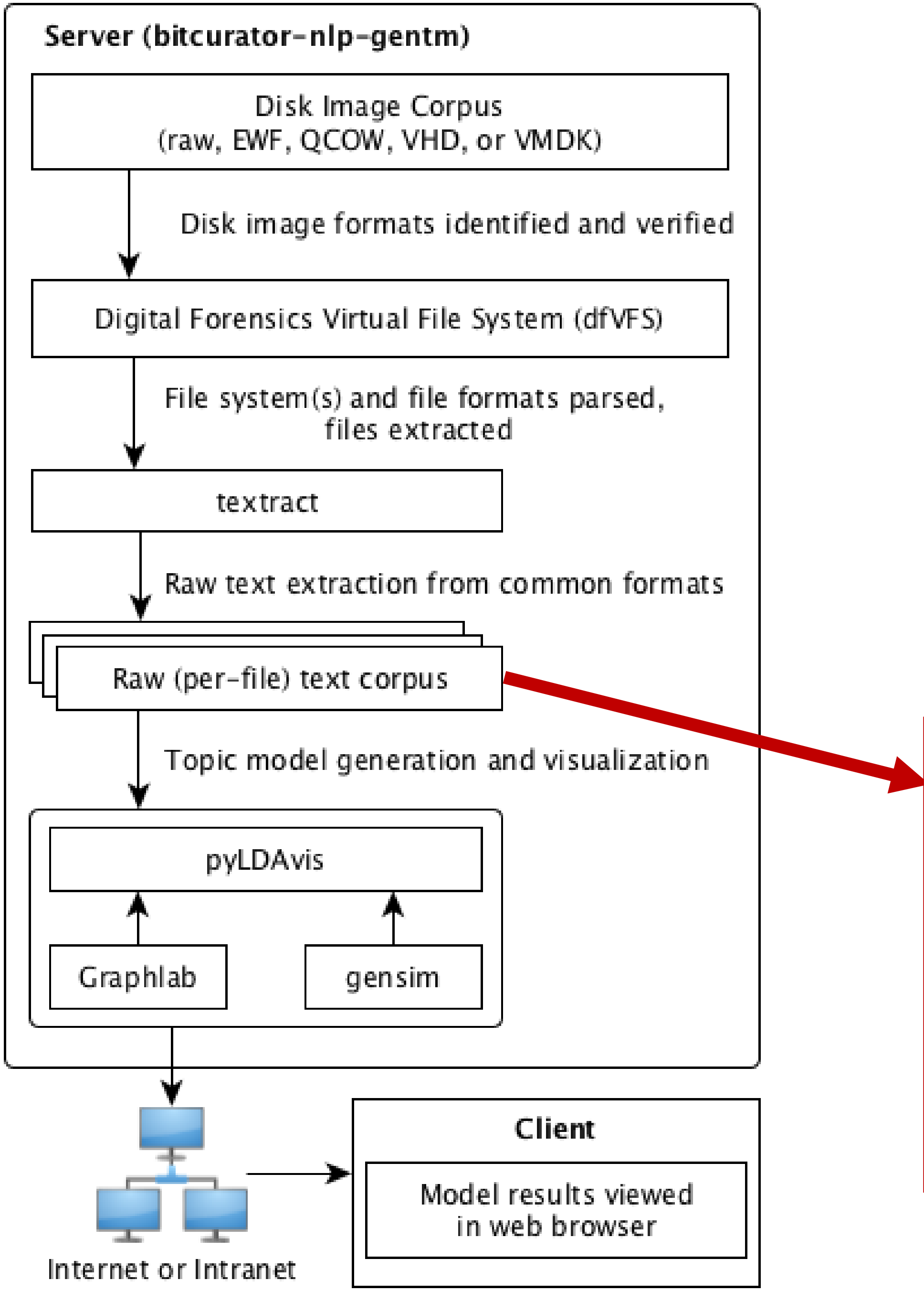
gensim

Internet or Intranet

Client

Model results viewed
in web browser

This is showing topic modeling, which we'll look at in more detail soon.



Server (bitcurator-nlp-gentm)

Disk Image Corpus
(raw, EWF, QCOW, VHD, or VMDK)

Disk image formats identified and verified

Digital Forensics Virtual File System (dfVFS)

File system(s) and file formats parsed,
files extracted

textract

Raw text extraction from common formats

Raw (per-file) text corpus

Topic model generation and visualization

pyLDAvis

Graphlab

gensim

Internet or Intranet

Client

Model results viewed
in web browser

Another path at this point is to feed the text into spacy for named-entity recognition

Available Toolsets

BitCurator Access Webtools

- Browse file systems contained in disk images on the web
- Recently refactored and updated to improve support for collections of arbitrary size, extract text from common formats, build full-text index
- https://www.github.com/bitcurator/bitcurator-access-webtools

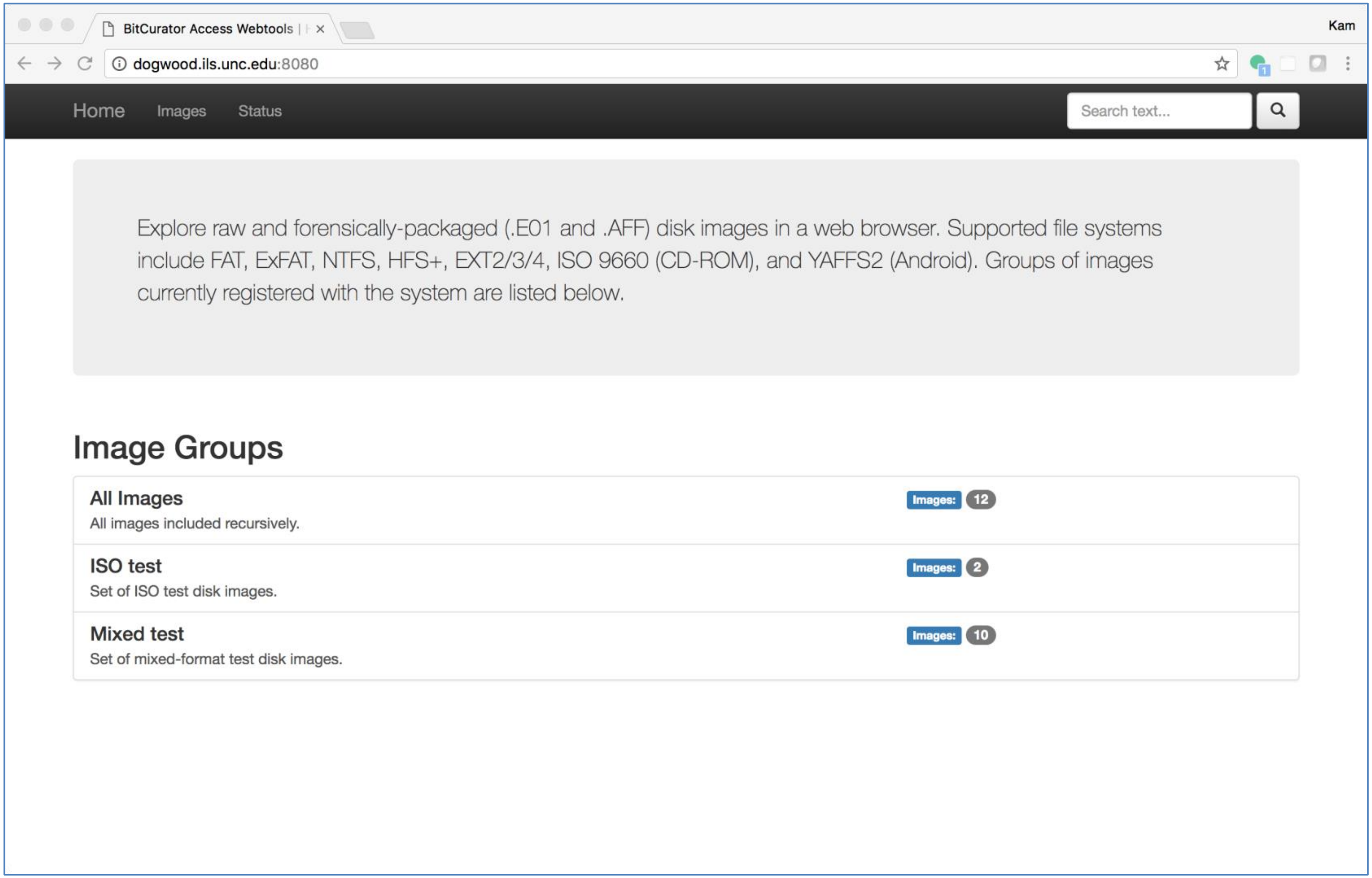
Topic modeling of disk image contents (bitcurator-nlp-gentm)

- Automated file extraction, text extraction and postprocessing (stemming, lemmatization, stopword removal, etc) via The Sleuth Kit, textract and GraphLab
- LDA (Latent Dirichlet Allocation) for topic discovery and visualization via pyLDAvis
- https://www.github.com/bitcurator/bitcurator-nlp-gentm

Entity identification and reporting for heterogeneous file collections (bitcurator-nlp-entspan)

- Automated text extraction and postprocessing via textract and spaCy
- Rapid query and reporting on per-document entity presence and entity distribution
- https://www.github.com/bitcurator/bitcurator-nlp-entspan

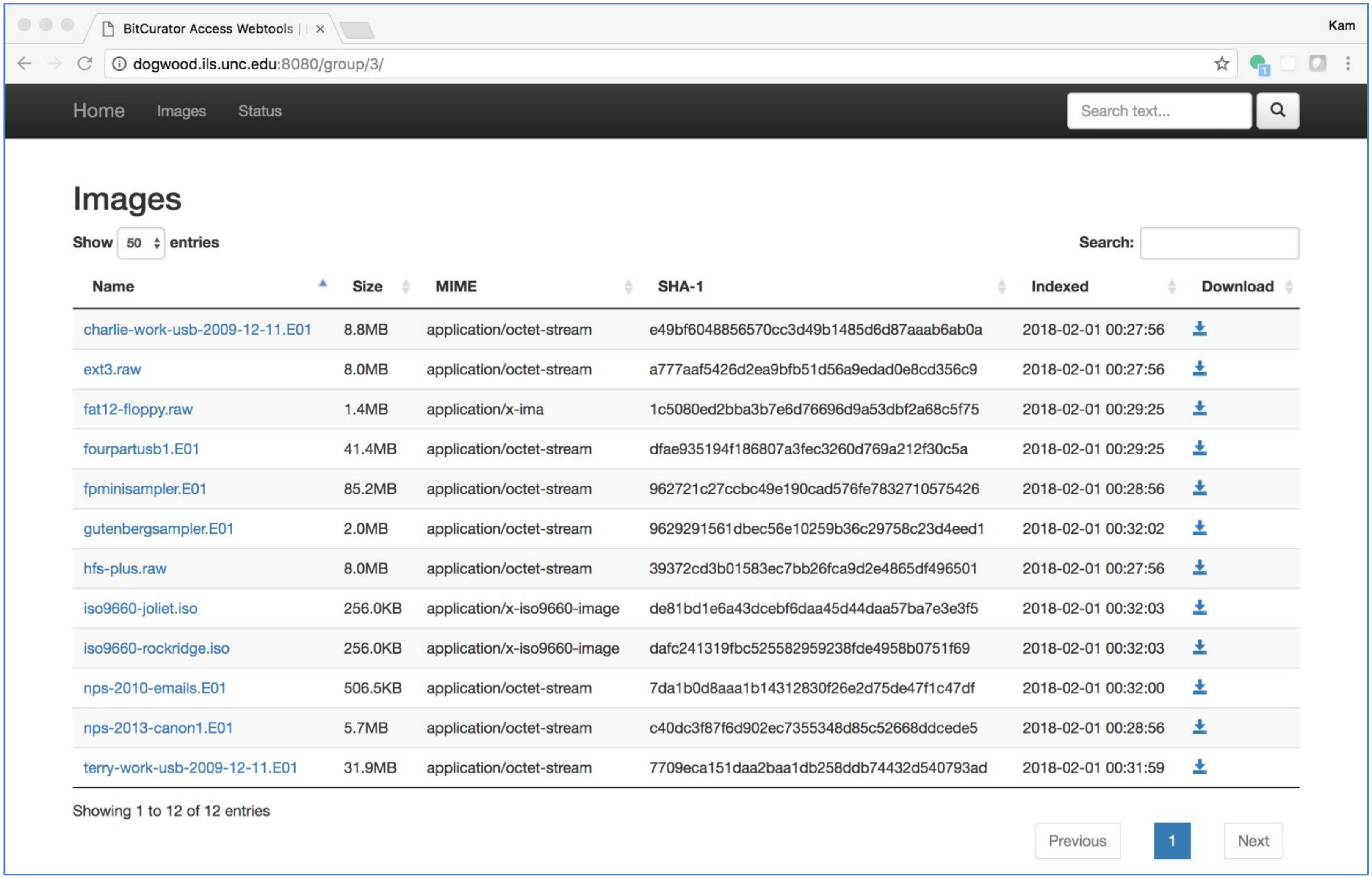
BitCurator Access Webtools



Explore raw and forensically-packaged (.E01 and .AFF) disk images in a web browser. Supported file systems include FAT, ExFAT, NTFS, HFS+, EXT2/3/4, ISO 9660 (CD-ROM), and YAFFS2 (Android). Groups of images currently registered with the system are listed below.

Image Groups

All Images All images included recursively.	Images: 12
ISO test Set of ISO test disk images.	Images: 2
Mixed test Set of mixed-format test disk images.	Images: 10



BitCurator Access Webtools | ×
Kam
dogwood.ils.unc.edu:8080/group/3/
Home
Images
Status
Search text...

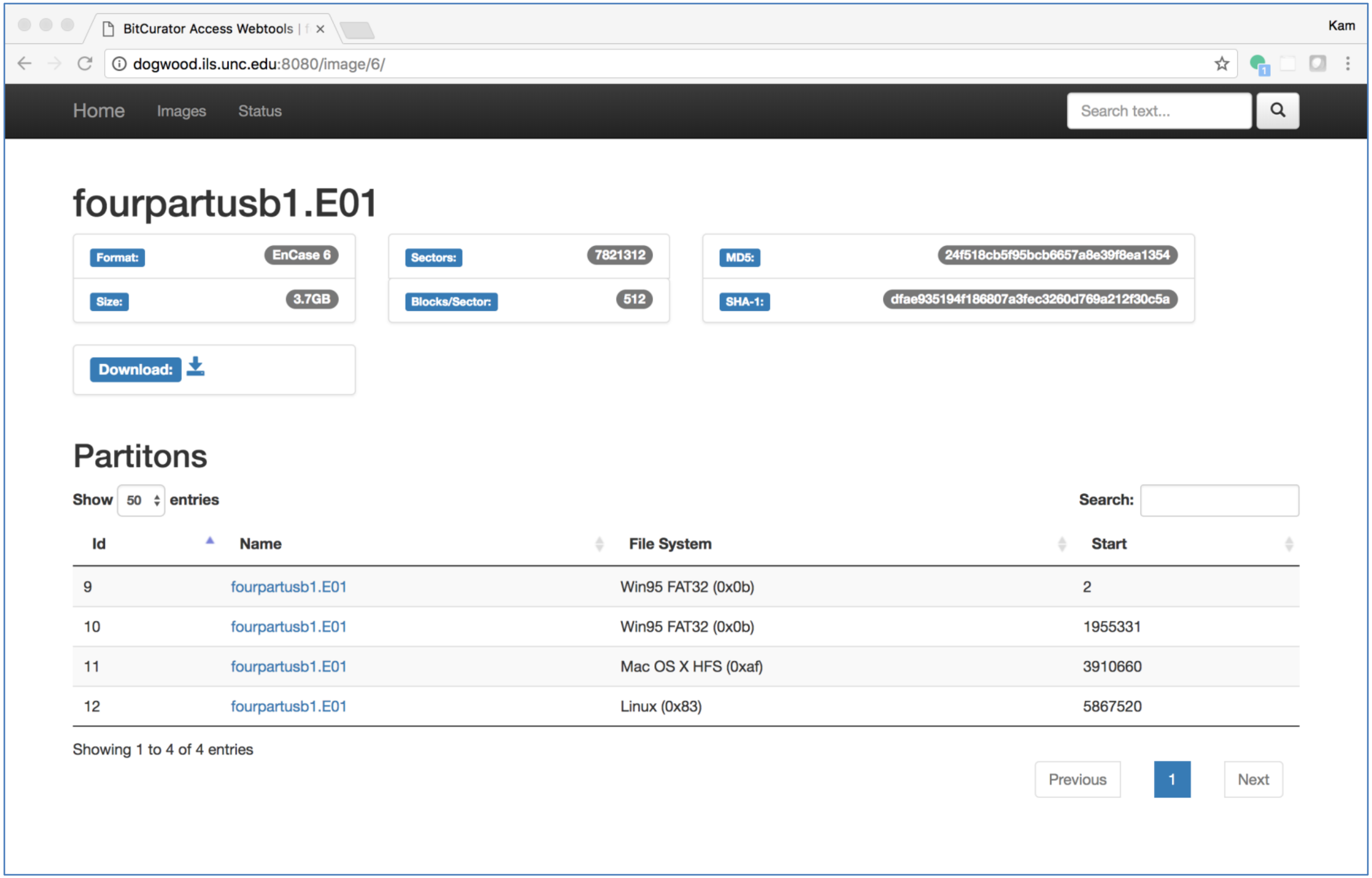
Images

Show 50 entries

Search:

Name	Size	MIME	SHA-1	Indexed	Download
charlie-work-usb-2009-12-11.E01	8.8MB	application/octet-stream	e49bf6048856570cc3d49b1485d6d87aaab6ab0a	2018-02-01 00:27:56	⤓
ext3.raw	8.0MB	application/octet-stream	a777aaf5426d2ea9bfb51d56a9edad0e8cd356c9	2018-02-01 00:27:56	⤓
fat12-floppy.raw	1.4MB	application/x-ima	1c5080ed2bba3b7e6d76696d9a53dbf2a68c5f75	2018-02-01 00:29:25	⤓
fourpartusb1.E01	41.4MB	application/octet-stream	dfae935194f186807a3fec3260d769a212f30c5a	2018-02-01 00:29:25	⤓
fpminisampler.E01	85.2MB	application/octet-stream	962721c27ccbc49e190cad576fe7832710575426	2018-02-01 00:28:56	⤓
gutenbergsampler.E01	2.0MB	application/octet-stream	9629291561dbec56e10259b36c29758c23d4eed1	2018-02-01 00:32:02	⤓
hfs-plus.raw	8.0MB	application/octet-stream	39372cd3b01583ec7bb26fca9d2e4865df496501	2018-02-01 00:27:56	⤓
iso9660-joliet.iso	256.0KB	application/x-iso9660-image	de81bd1e6a43dcebf6daa45d44daa57ba7e3e3f5	2018-02-01 00:32:03	⤓
iso9660-rockridge.iso	256.0KB	application/x-iso9660-image	dafc241319fbc525582959238fde4958b0751f69	2018-02-01 00:32:03	⤓
nps-2010-emails.E01	506.5KB	application/octet-stream	7da1b0d8aaa1b14312830f26e2d75de47f1c47df	2018-02-01 00:32:00	⤓
nps-2013-canon1.E01	5.7MB	application/octet-stream	c40dc3f87f6d902ec7355348d85c52668ddcede5	2018-02-01 00:28:56	⤓
terry-work-usb-2009-12-11.E01	31.9MB	application/octet-stream	7709eca151daa2baa1db258ddb74432d540793ad	2018-02-01 00:31:59	⤓

Showing 1 to 12 of 12 entries


BitCurator Access Webtools | f ×
Kam
dogwood.ils.unc.edu:8080/image/6/
Home
Images
Status
Search text...

fourpartusb1.E01

Format:	EnCase 6
Size:	3.7GB
Sectors:	7821312
Blocks/Sector:	512
MD5:	24f518cb5f95bcb6657a8e39f8ea1354
SHA-1:	dfae935194f186807a3fec3260d769a212f30c5a

Download:

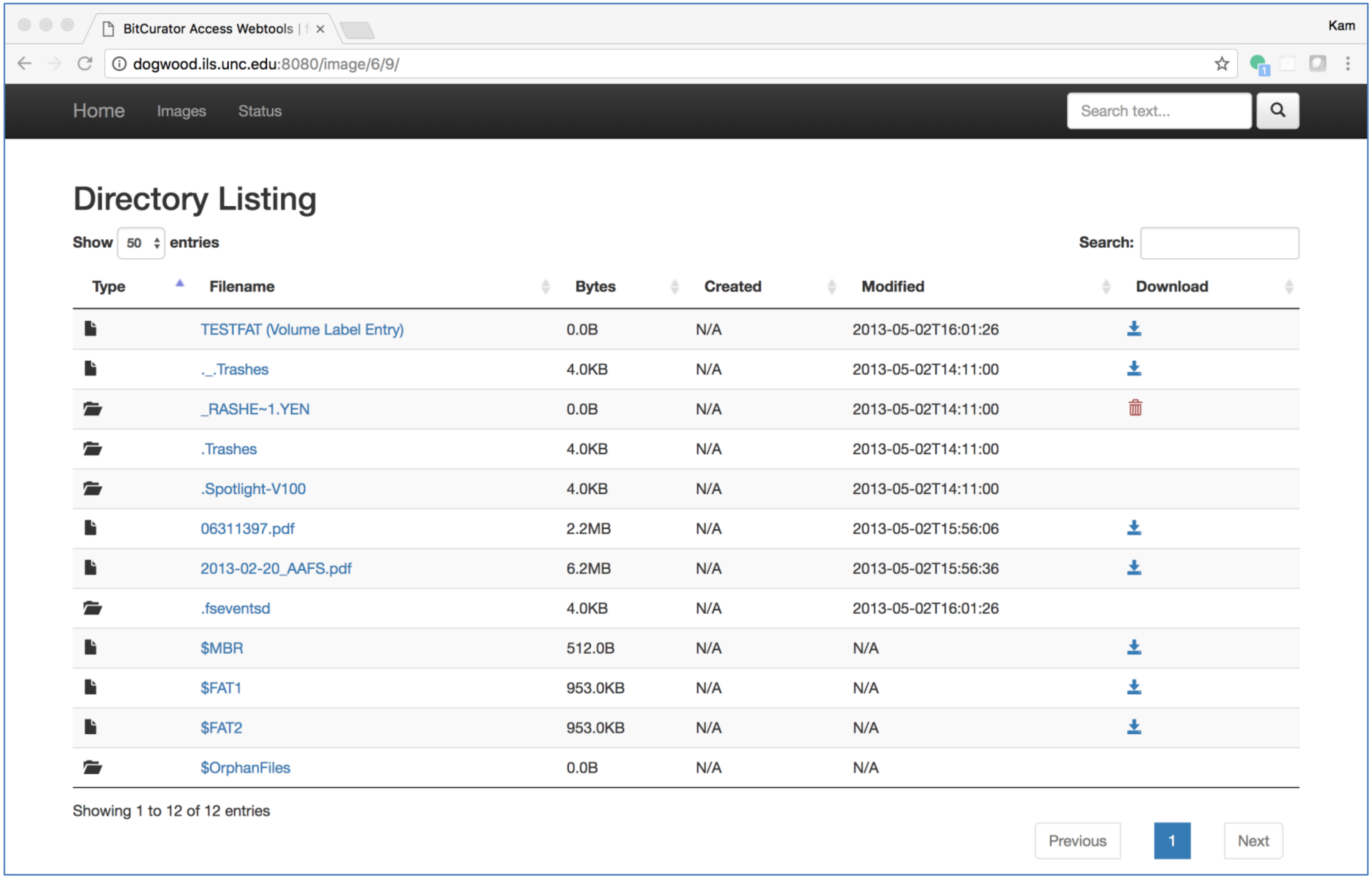
Partitons

Show 50 entries

Search:

Id	Name	File System	Start
9	fourpartusb1.E01	Win95 FAT32 (0x0b)	2
10	fourpartusb1.E01	Win95 FAT32 (0x0b)	1955331
11	fourpartusb1.E01	Mac OS X HFS (0xaf)	3910660
12	fourpartusb1.E01	Linux (0x83)	5867520

Showing 1 to 4 of 4 entries



BitCurator Access Webtools | f
Kam
dogwood.ils.unc.edu:8080/image/6/9/
Home
Images
Status
Search text...

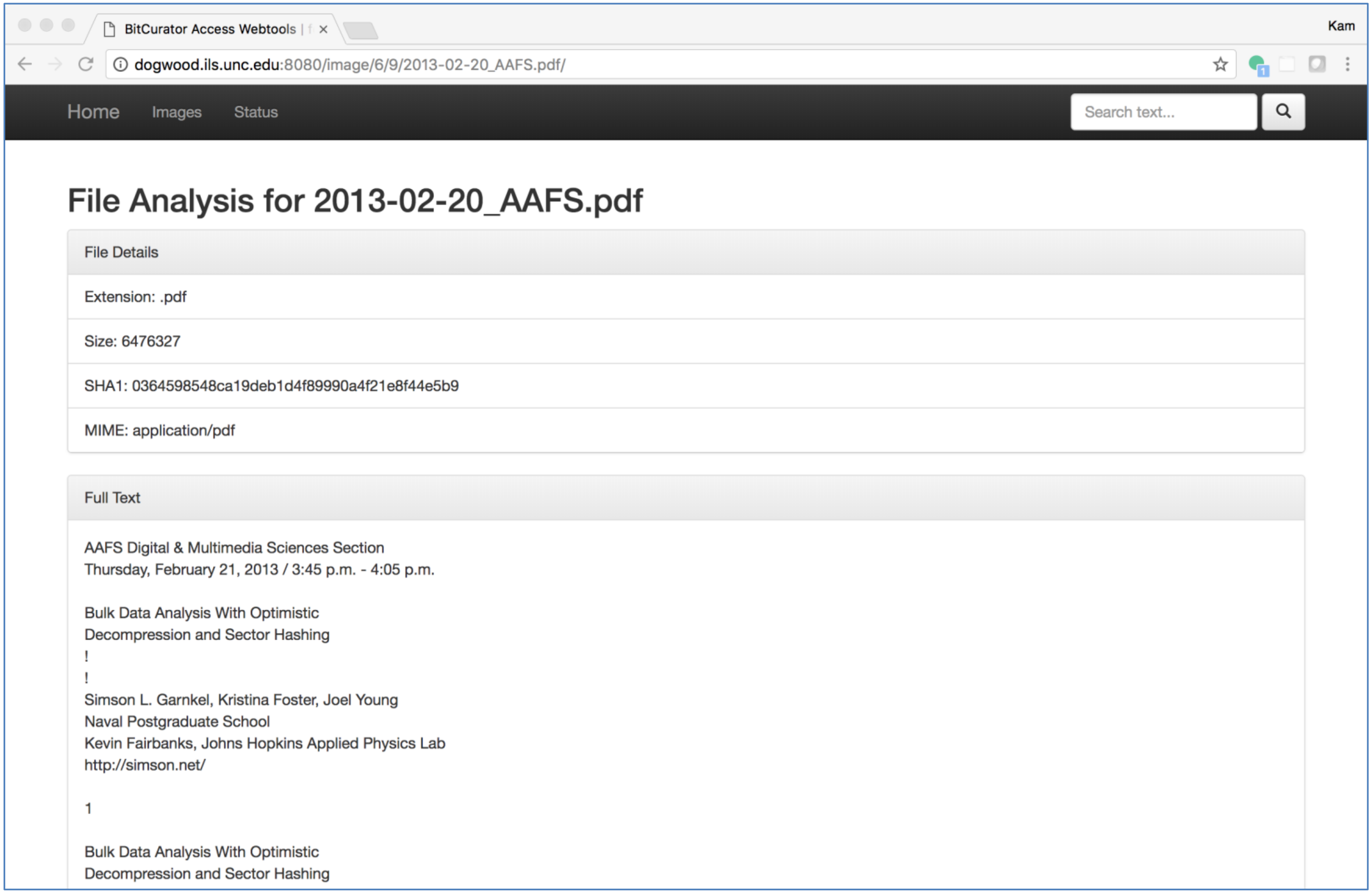
Directory Listing

Show 50 entries

Search:

Type	Filename	Bytes	Created	Modified	Download
📄	TESTFAT (Volume Label Entry)	0.0B	N/A	2013-05-02T16:01:26	⬇
📄	._.Trashes	4.0KB	N/A	2013-05-02T14:11:00	⬇
📂	_RASHE~1.YEN	0.0B	N/A	2013-05-02T14:11:00	🗑
📂	.Trashes	4.0KB	N/A	2013-05-02T14:11:00	
📂	.Spotlight-V100	4.0KB	N/A	2013-05-02T14:11:00	
📄	06311397.pdf	2.2MB	N/A	2013-05-02T15:56:06	⬇
📄	2013-02-20_AAFS.pdf	6.2MB	N/A	2013-05-02T15:56:36	⬇
📂	.fseventsd	4.0KB	N/A	2013-05-02T16:01:26	
📄	$MBR	512.0B	N/A	N/A	⬇
📄	$FAT1	953.0KB	N/A	N/A	⬇
📄	$FAT2	953.0KB	N/A	N/A	⬇
📂	$OrphanFiles	0.0B	N/A	N/A	

Showing 1 to 12 of 12 entries


BitCurator Access Webtools | f
dogwood.ils.unc.edu:8080/image/6/9/2013-02-20_AAFS.pdf/
Kam
Home
Images
Status
Search text...

File Analysis for 2013-02-20_AAFS.pdf

File Details

Extension: .pdf

Size: 6476327

SHA1: 0364598548ca19deb1d4f89990a4f21e8f44e5b9

MIME: application/pdf

Full Text

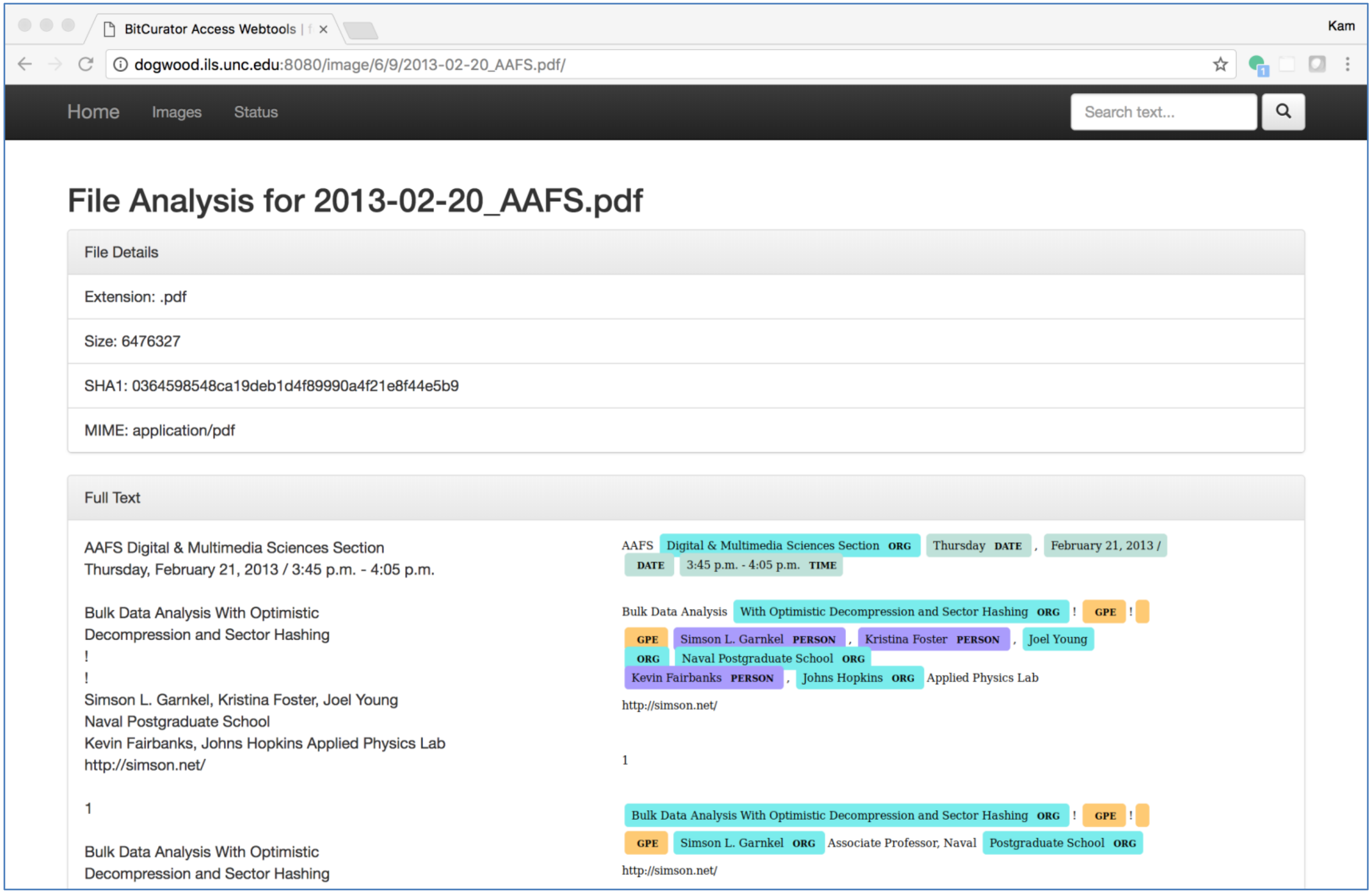
AAFS Digital & Multimedia Sciences Section
Thursday, February 21, 2013 / 3:45 p.m. - 4:05 p.m.

Bulk Data Analysis With Optimistic
Decompression and Sector Hashing
!
!
Simson L. Garnkel, Kristina Foster, Joel Young
Naval Postgraduate School
Kevin Fairbanks, Johns Hopkins Applied Physics Lab
http://simson.net/

1

Bulk Data Analysis With Optimistic
Decompression and Sector Hashing

Entity type	Description
PERSON	People
NORP	Nationalities, religious, and political groups.
FACILITY	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, and institutions.
GPE	Countries, cities, and states.
LOC	Locations other than GPE (e.g. mountain ranges, bodies of water)
PRODUCT	Objects other than services (e.g. devices, foods)
EVENT	Historical events (e.g. cultural, weather, conflicts)
WORK_OF_ART	Titles of works of art
LANGUAGE	Named languages
Additional feature types	Description
DATE	Dates or periods (absolute / relative)
TIME	Time periods less than a day
PERCENT	Percentages (also marked by ‘%’)
MONEY	Monetary values, including by unit
QUANTITY	Weight, distance, other measurements
ORDINAL	E.g ‘first’, ‘second’
CARDINAL	Numeral identifiers other than those typed above



BitCurator Access Webtools | f
Kam
dogwood.ils.unc.edu:8080/image/6/9/2013-02-20_AAFS.pdf/
Home
Images
Status
Search text...

File Analysis for 2013-02-20_AAFS.pdf

File Details

Extension: .pdf

Size: 6476327

SHA1: 0364598548ca19deb1d4f89990a4f21e8f44e5b9

MIME: application/pdf

Full Text

AAFS Digital & Multimedia Sciences Section
Thursday, February 21, 2013 / 3:45 p.m. - 4:05 p.m.

Bulk Data Analysis With Optimistic
Decompression and Sector Hashing
!
!
Simson L. Garnkel, Kristina Foster, Joel Young
Naval Postgraduate School
Kevin Fairbanks, Johns Hopkins Applied Physics Lab
http://simson.net/

1

Bulk Data Analysis With Optimistic
Decompression and Sector Hashing

AAFS Digital & Multimedia Sciences Section ORG Thursday DATE , February 21, 2013 / DATE 3:45 p.m. - 4:05 p.m. TIME

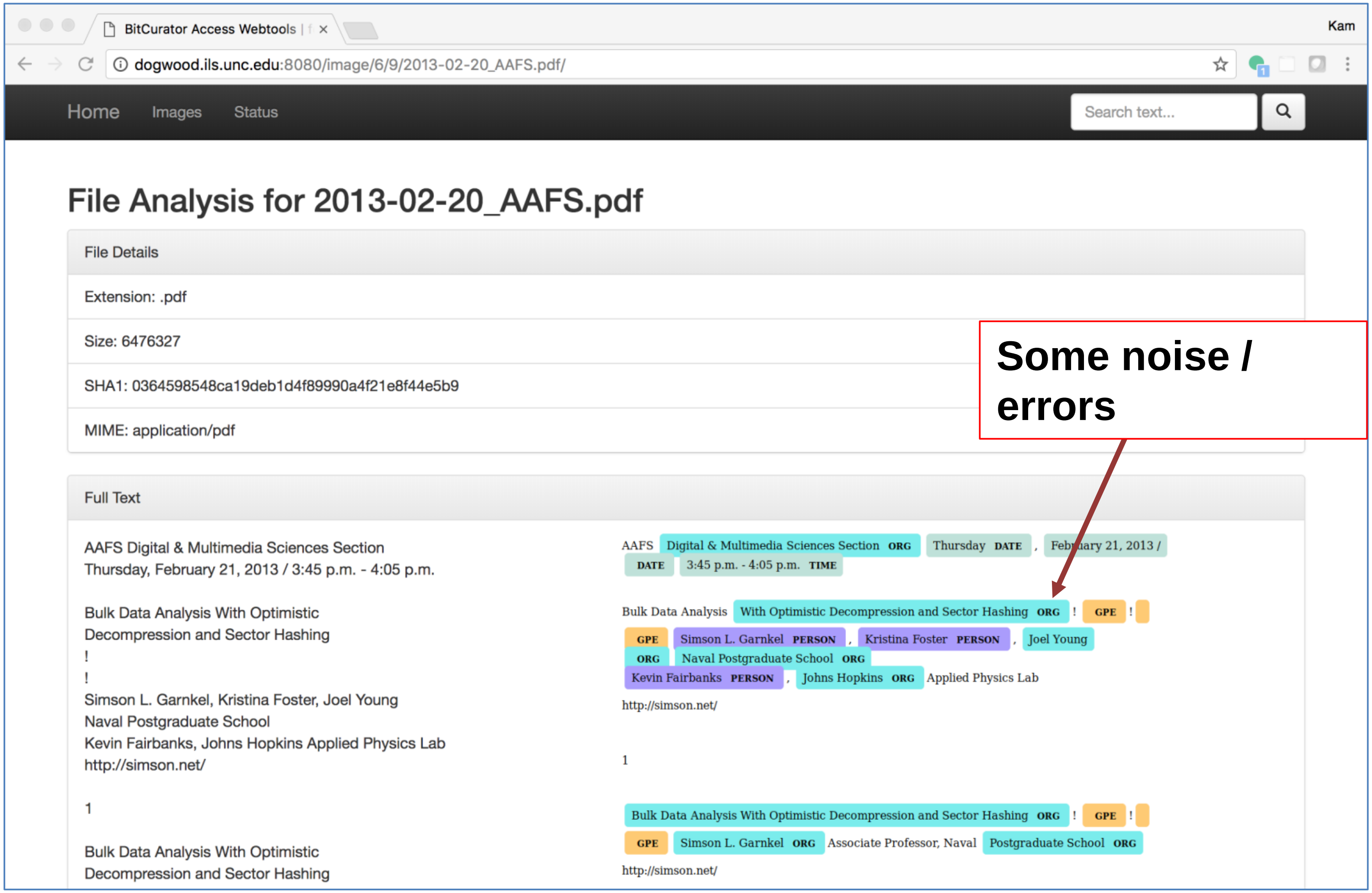
Bulk Data Analysis With Optimistic Decompression and Sector Hashing ORG ! GPE ! GPE Simson L. Garnkel PERSON , Kristina Foster PERSON , Joel Young ORG Naval Postgraduate School ORG Kevin Fairbanks PERSON , Johns Hopkins ORG Applied Physics Lab

http://simson.net/

1

Bulk Data Analysis With Optimistic Decompression and Sector Hashing ORG ! GPE ! GPE Simson L. Garnkel ORG Associate Professor, Naval Postgraduate School ORG

http://simson.net/



BitCurator Access Webtools | f ×
Kam
dogwood.ils.unc.edu:8080/image/6/9/2013-02-20_AAFS.pdf/
Home
Images
Status
Search text...

File Analysis for 2013-02-20_AAFS.pdf

File Details

Extension: .pdf

Size: 6476327

SHA1: 0364598548ca19deb1d4f89990a4f21e8f44e5b9

MIME: application/pdf

Full Text

AAFS Digital & Multimedia Sciences Section
Thursday, February 21, 2013 / 3:45 p.m. - 4:05 p.m.

Bulk Data Analysis With Optimistic
Decompression and Sector Hashing
!
!
Simson L. Garnkel, Kristina Foster, Joel Young
Naval Postgraduate School
Kevin Fairbanks, Johns Hopkins Applied Physics Lab
http://simson.net/

1

Bulk Data Analysis With Optimistic
Decompression and Sector Hashing

AAFS Digital & Multimedia Sciences Section ORG Thursday DATE , February 21, 2013 / DATE 3:45 p.m. - 4:05 p.m. TIME

Bulk Data Analysis With Optimistic Decompression and Sector Hashing ORG ! GPE ! GPE Simson L. Garnkel PERSON , Kristina Foster PERSON , Joel Young ORG Naval Postgraduate School ORG Kevin Fairbanks PERSON , Johns Hopkins ORG Applied Physics Lab

http://simson.net/

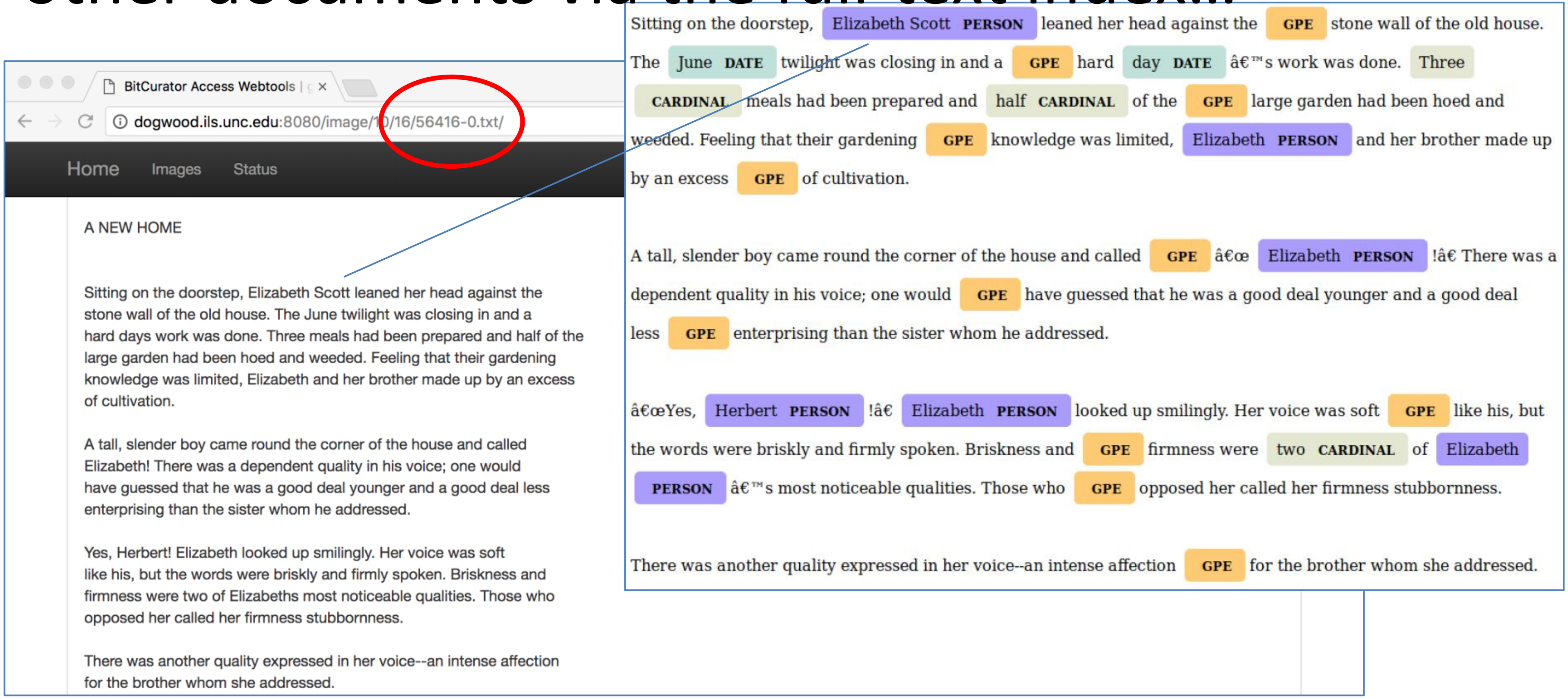
1

Bulk Data Analysis With Optimistic Decompression and Sector Hashing ORG ! GPE ! GPE Simson L. Garnkel ORG Associate Professor, Naval Postgraduate School ORG

http://simson.net/

Some noise / errors

Entities tagged by DisplaCy can be located in other documents via the full-text index…



BitCurator Access Webtools | ×
dogwood.ils.unc.edu:8080/image/10/16/56416-0.txt/
Home
Images
Status

A NEW HOME

Sitting on the doorstep, Elizabeth Scott leaned her head against the stone wall of the old house. The June twilight was closing in and a hard days work was done. Three meals had been prepared and half of the large garden had been hoed and weeded. Feeling that their gardening knowledge was limited, Elizabeth and her brother made up by an excess of cultivation.

A tall, slender boy came round the corner of the house and called Elizabeth! There was a dependent quality in his voice; one would have guessed that he was a good deal younger and a good deal less enterprising than the sister whom he addressed.

Yes, Herbert! Elizabeth looked up smilingly. Her voice was soft like his, but the words were briskly and firmly spoken. Briskness and firmness were two of Elizabeths most noticeable qualities. Those who opposed her called her firmness stubbornness.

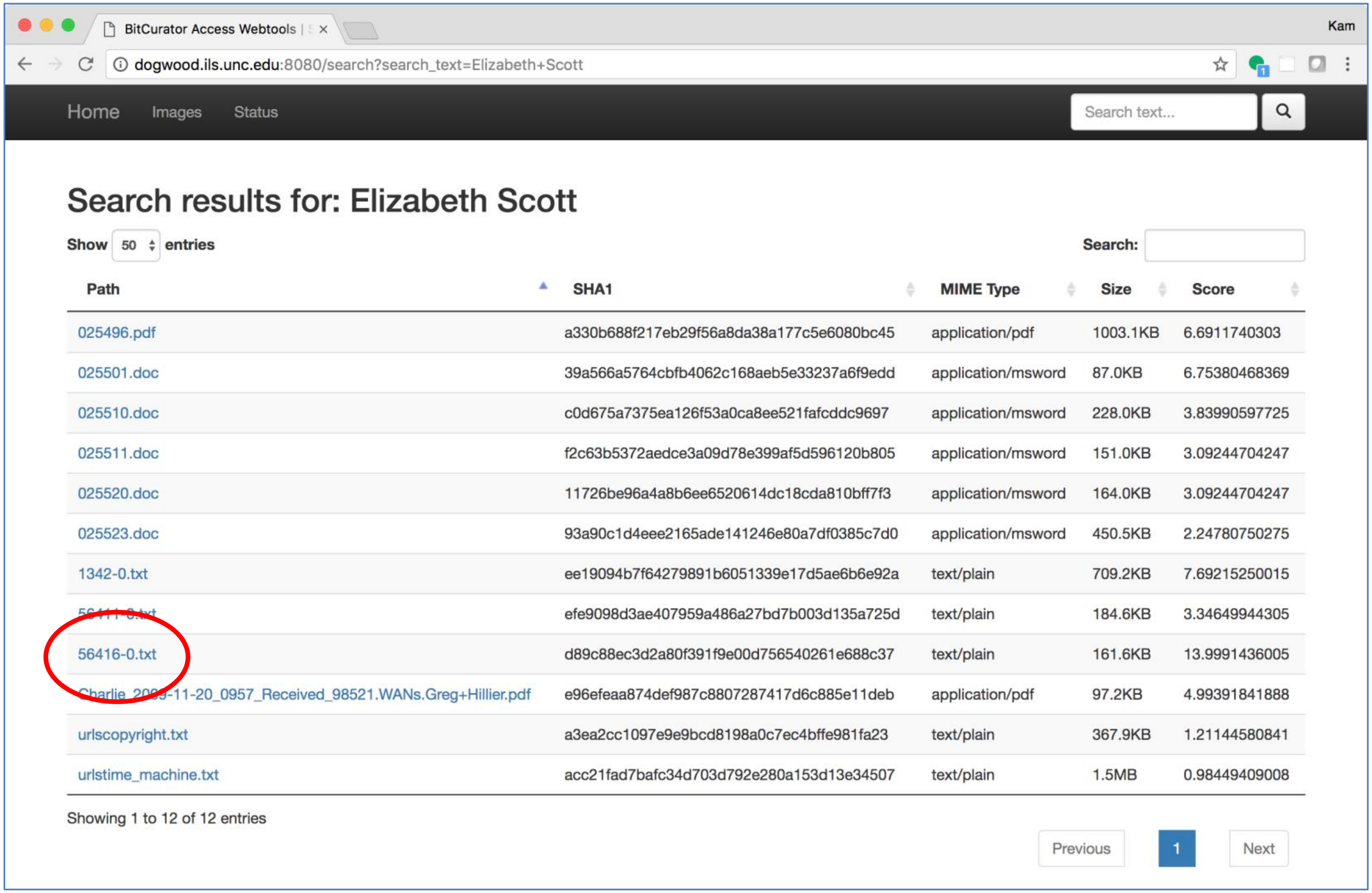
There was another quality expressed in her voice--an intense affection for the brother whom she addressed.

Sitting on the doorstep, Elizabeth Scott **PERSON** leaned her head against the **GPE** stone wall of the old house. The June **DATE** twilight was closing in and a **GPE** hard day **DATE** â€™s work was done. Three **CARDINAL** meals had been prepared and half **CARDINAL** of the **GPE** large garden had been hoed and weeded. Feeling that their gardening **GPE** knowledge was limited, Elizabeth **PERSON** and her brother made up by an excess **GPE** of cultivation.

A tall, slender boy came round the corner of the house and called **GPE** â€œ Elizabeth **PERSON** !â€ There was a dependent quality in his voice; one would **GPE** have guessed that he was a good deal younger and a good deal less **GPE** enterprising than the sister whom he addressed.

â€œYes, Herbert **PERSON** !â€ Elizabeth **PERSON** looked up smilingly. Her voice was soft **GPE** like his, but the words were briskly and firmly spoken. Briskness and **GPE** firmness were two **CARDINAL** of Elizabeth **PERSON** â€™s most noticeable qualities. Those who **GPE** opposed her called her firmness stubbornness.

There was another quality expressed in her voice--an intense affection **GPE** for the brother whom she addressed.


BitCurator Access Webtools | S ×
Kam
dogwood.ils.unc.edu:8080/search?search_text=Elizabeth+Scott
Home
Images
Status
Search text...

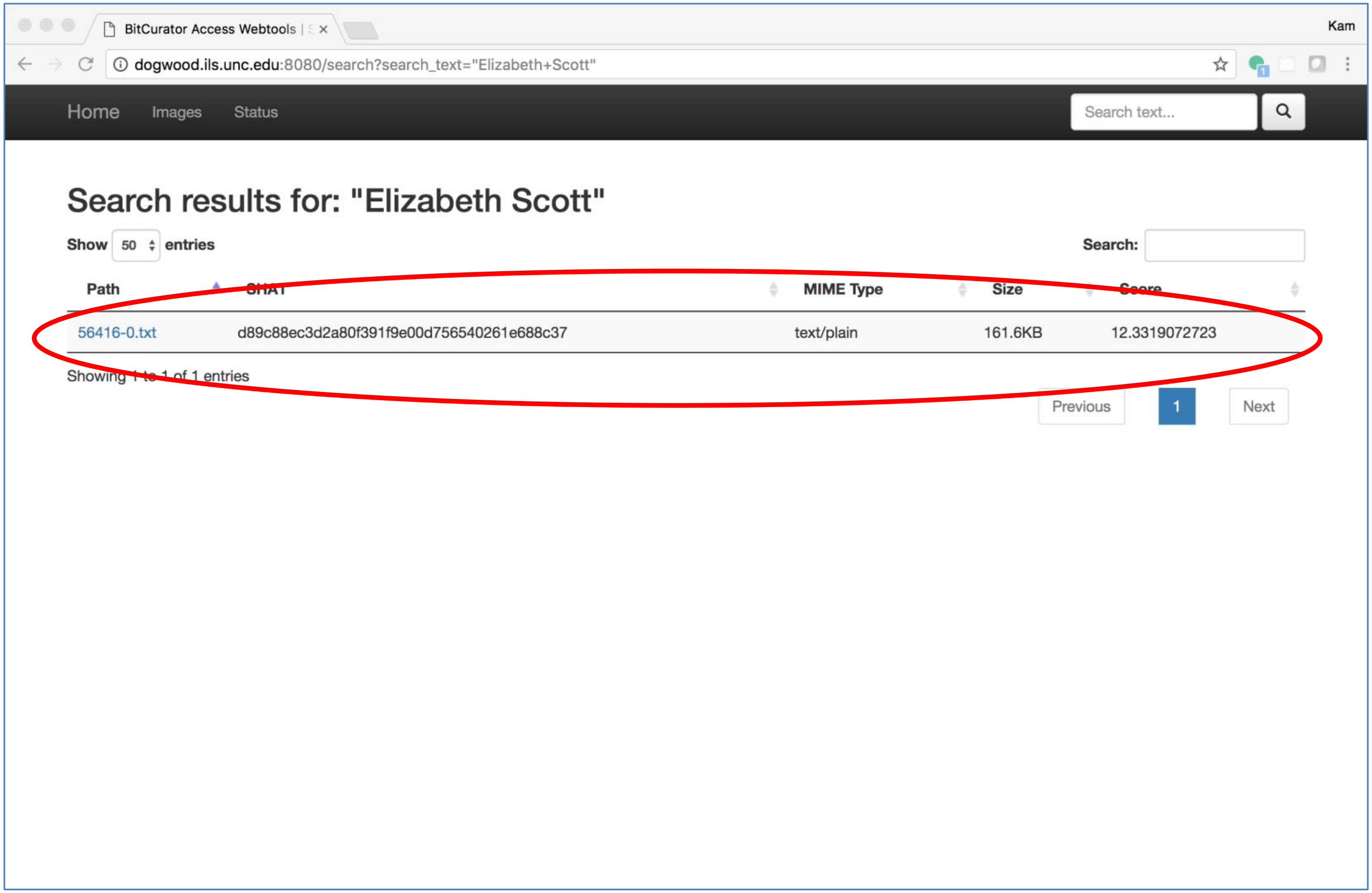
Search results for: Elizabeth Scott

Show 50 entries

Search:

Path	SHA1	MIME Type	Size	Score
025496.pdf	a330b688f217eb29f56a8da38a177c5e6080bc45	application/pdf	1003.1KB	6.6911740303
025501.doc	39a566a5764cbfb4062c168aeb5e33237a6f9edd	application/msword	87.0KB	6.75380468369
025510.doc	c0d675a7375ea126f53a0ca8ee521fafcddc9697	application/msword	228.0KB	3.83990597725
025511.doc	f2c63b5372aedce3a09d78e399af5d596120b805	application/msword	151.0KB	3.09244704247
025520.doc	11726be96a4a8b6ee6520614dc18cda810bff7f3	application/msword	164.0KB	3.09244704247
025523.doc	93a90c1d4eee2165ade141246e80a7df0385c7d0	application/msword	450.5KB	2.24780750275
1342-0.txt	ee19094b7f64279891b6051339e17d5ae6b6e92a	text/plain	709.2KB	7.69215250015
56411-0.txt	efe9098d3ae407959a486a27bd7b003d135a725d	text/plain	184.6KB	3.34649944305
56416-0.txt	d89c88ec3d2a80f391f9e00d756540261e688c37	text/plain	161.6KB	13.9991436005
Charlie_2009-11-20_0957_Received_98521.WANs.Greg+Hillier.pdf	e96efeaa874def987c8807287417d6c885e11deb	application/pdf	97.2KB	4.99391841888
urlscopyright.txt	a3ea2cc1097e9e9bcd8198a0c7ec4bffe981fa23	text/plain	367.9KB	1.21144580841
urlstime_machine.txt	acc21fad7bafc34d703d792e280a153d13e34507	text/plain	1.5MB	0.98449409008

Showing 1 to 12 of 12 entries



dogwood.ils.unc.edu:8080/search?search_text="Elizabeth+Scott"

Search results for: "Elizabeth Scott"

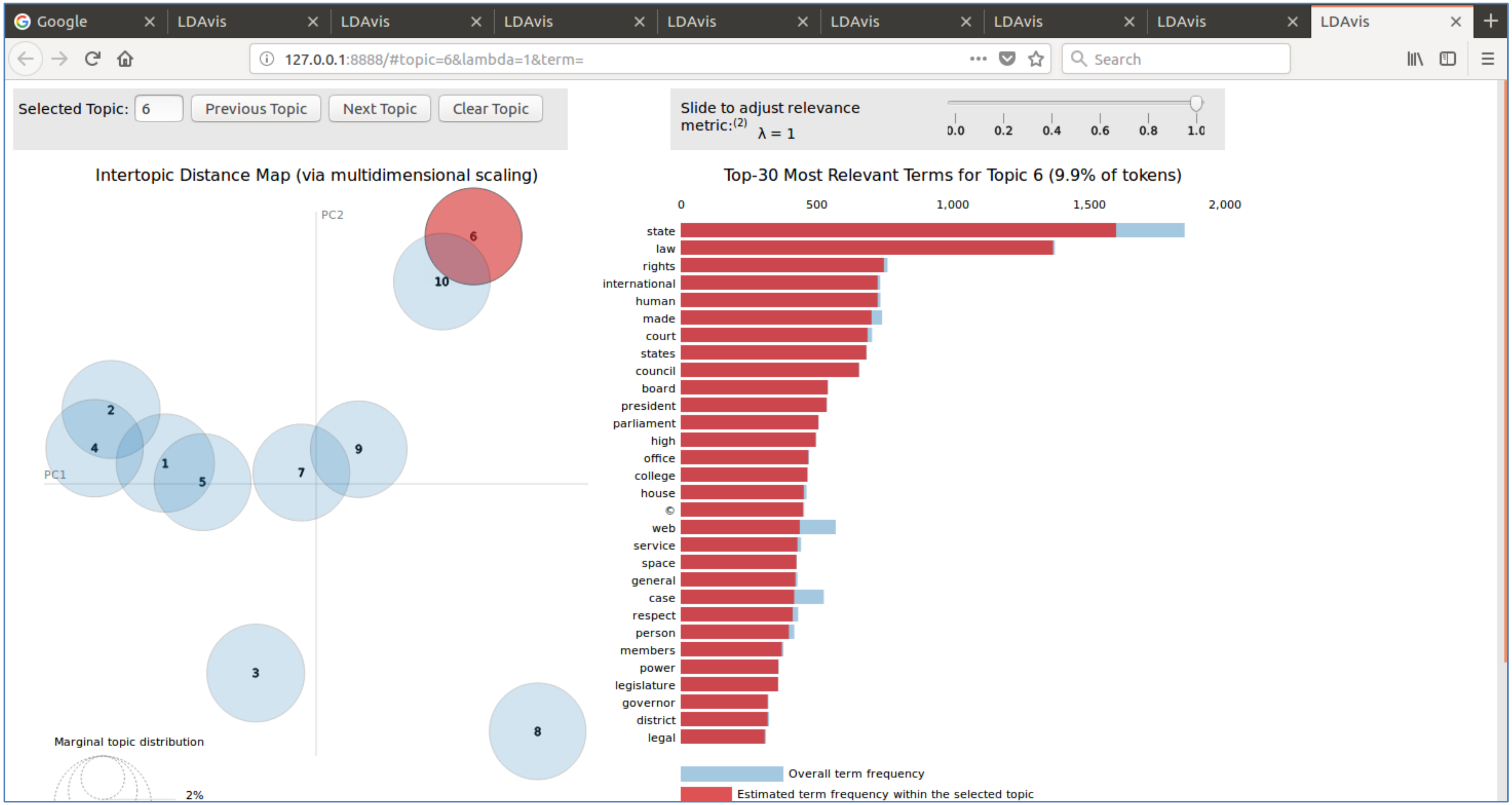
Show 50 entries

Search:

Path	SHA1	MIME Type	Size	Score
56416-0.txt	d89c88ec3d2a80f391f9e00d756540261e688c37	text/plain	161.6KB	12.3319072723

Showing 1 to 1 of 1 entries

BitCurator NLP / Topic Modeling Tools



127.0.0.1:8888/#topic=6&lambda=1&term=

Selected Topic: 6

Slide to adjust relevance metric:(2) λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 6 (9.9% of tokens)

Sources and Development Info

BitCurator in-development and past software projects on GitHub:

https://bitcurator.github.io

https://github.com/bitcurator/bitcurator-access-webtools

https://github.com/bitcurator/bitcurator-nlp-gentm

https://github.com/bitcurator/bitcurator-nlp-entspan