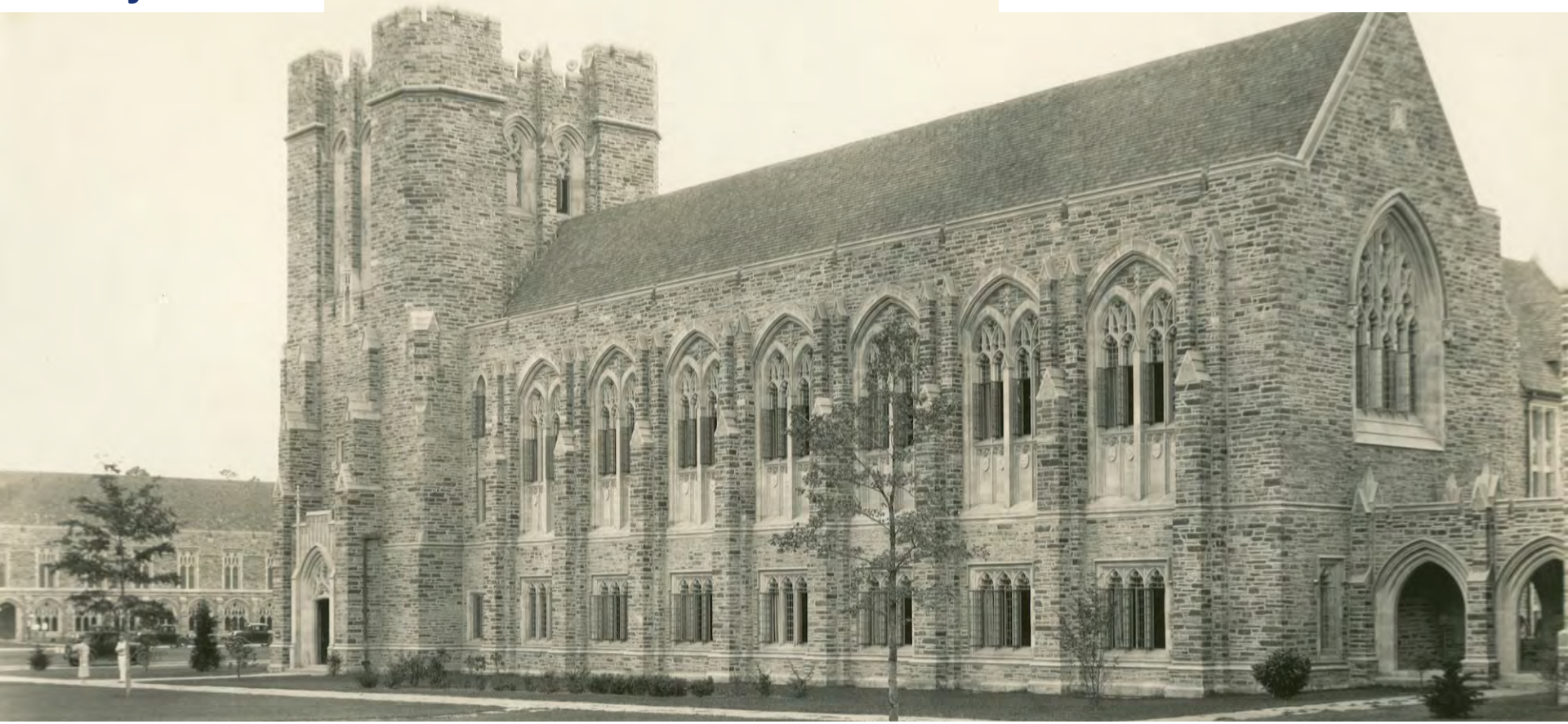
BitCurator at the Rubenstein

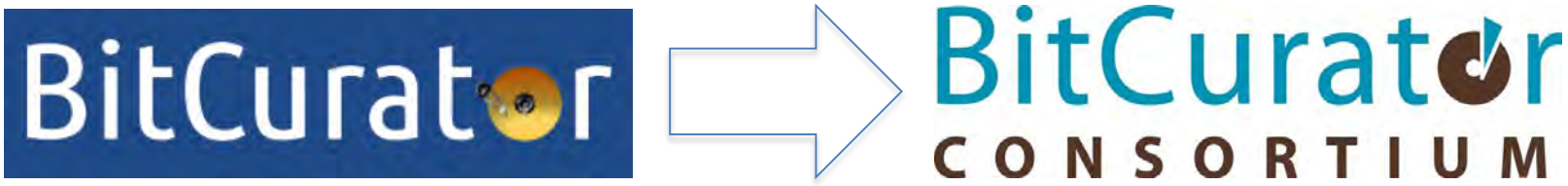
Matthew Farrell
16 July 2015

Association of Southeastern Research Libraries



Outline

- Institutional context
- Digital forensics: the (oh so) brief version
- BitCurator overview
- BitCurator at Duke

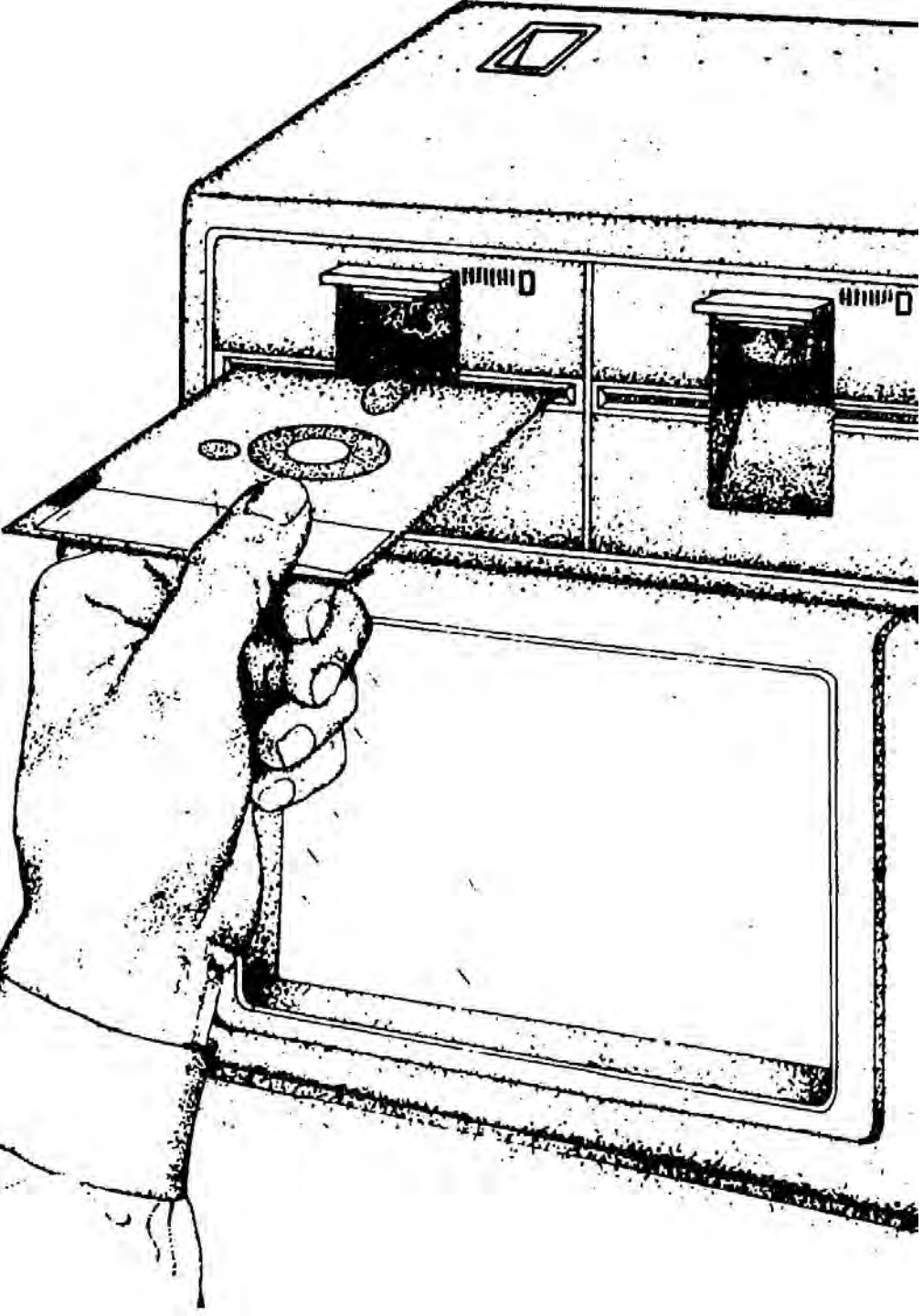

BitCurator
BitCurator
CONSORTIUM

Who we are and what we collect

Image: General Library Exterior, before 1948. Source: Duke University Archives. License: cc-by-nc-sa.

Digital Forensics, inadequately

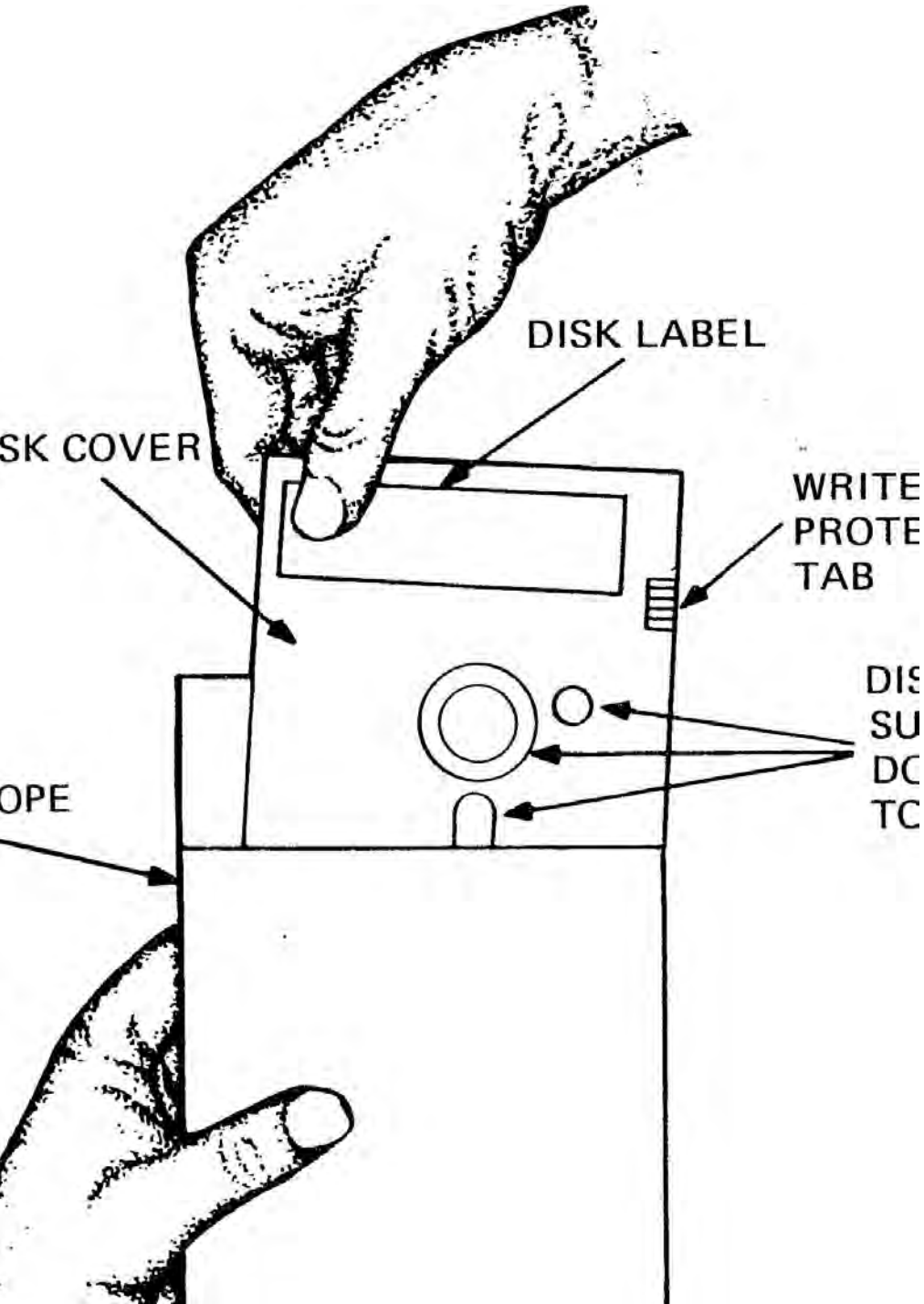
Image: Basic Forensic Imaging Kit. John Crel, photographer. Source: Flickr. License: cc-by-nd.



Boring, definition part I

- Academic definition of *digital forensics*:
  - "The use of scientifically derived methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence for the purpose of reconstructing events found to be criminal."
- Highlights
  - Criminal and legal audiences
  - Many digital preservation concepts

Image: Image from Page 41 of DEC Terminal Manual, 1982. Source: Internet Archive, Flickr Commons.



Boring definition part II

- Disk Image
- Filesystem v. Operating System
- Bit stream
- Hex code
- Fixity

Image: Image from Page 41 of DEC Terminal Manual, 1982. Source: Internet Archive, Flickr Commons.

What I talk about when I talk about digital forensics

Mapping forensic methods to archival practice

Acquisition: physical control

- Capture materials at lowest possible level
- Original order of files on media
- Establish fingerprint, monitor over time
- Document physical context
- Create access and preservation copies

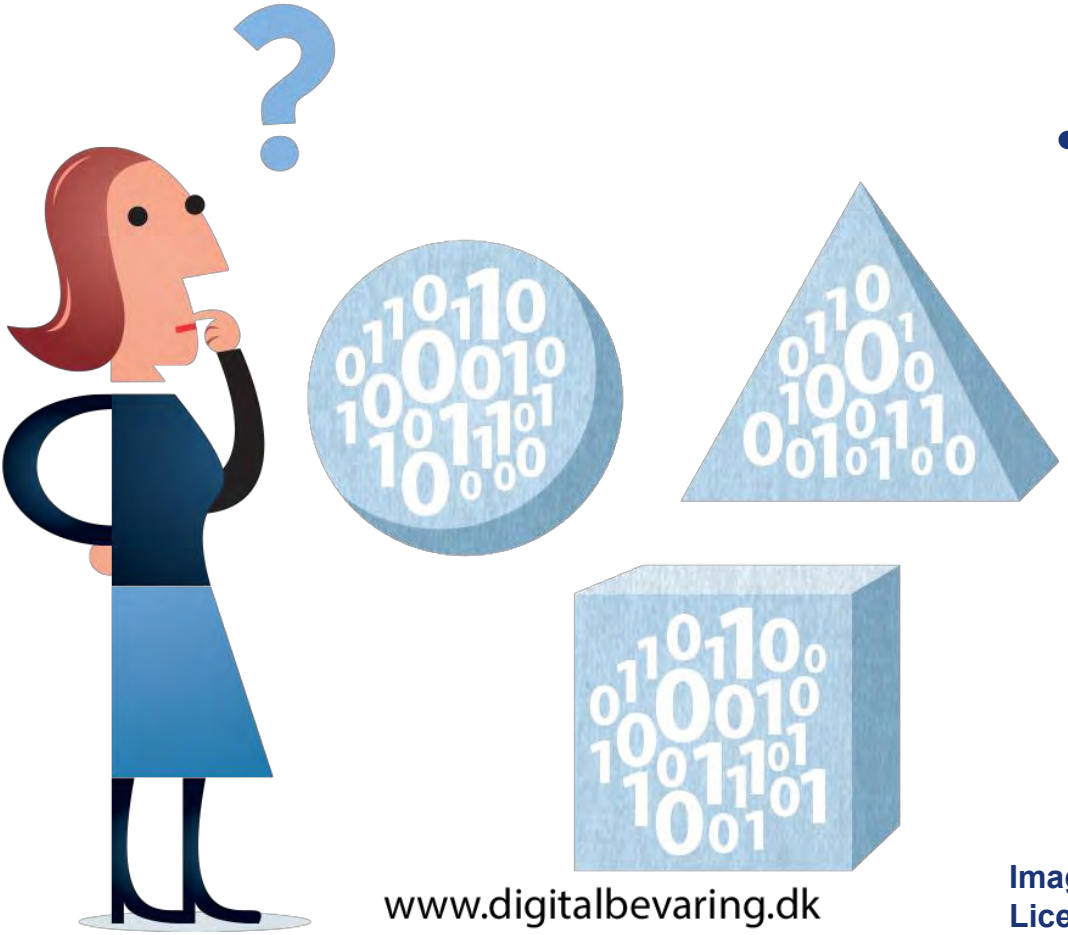
Acquisition: intellectual control

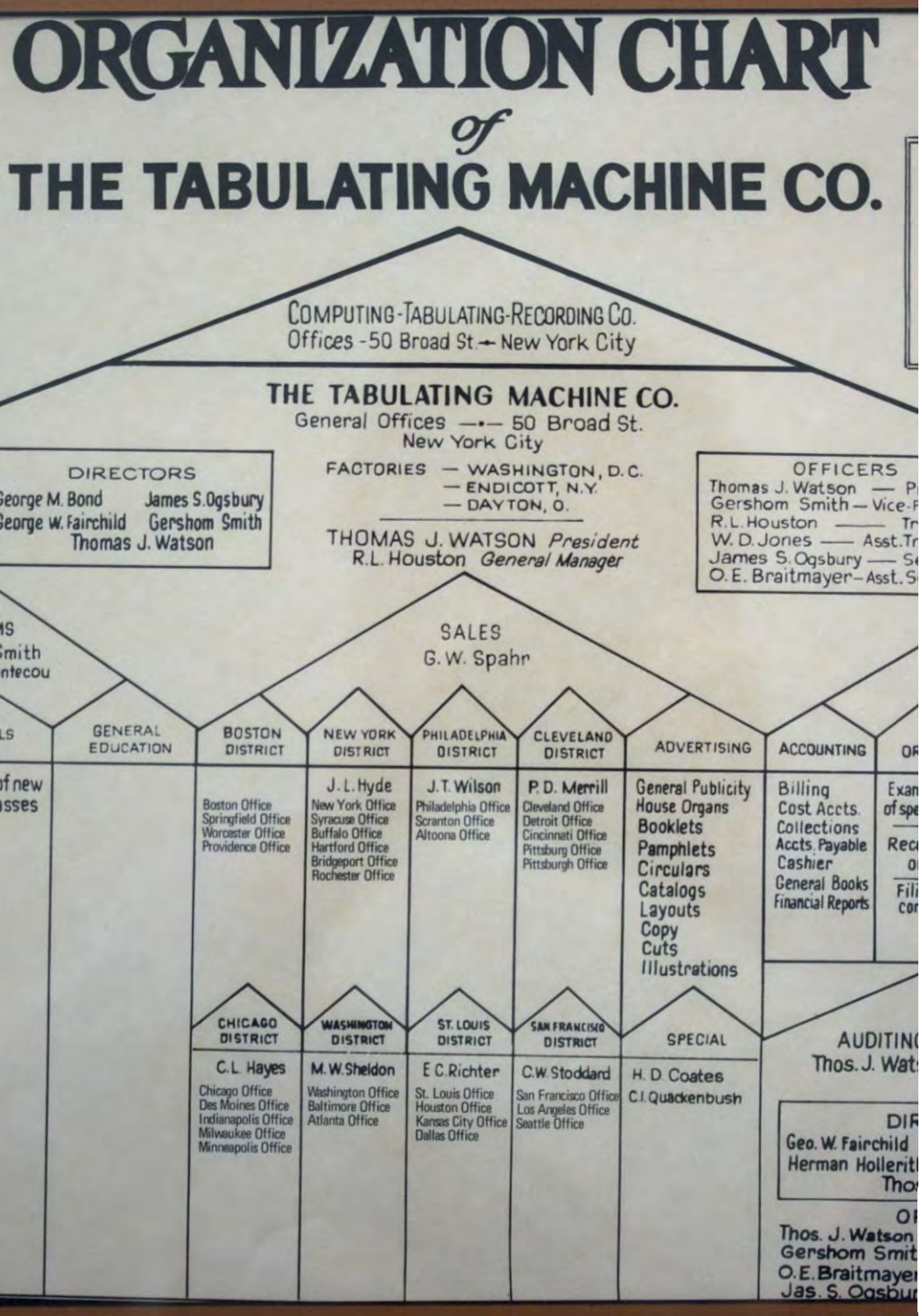
- Be aware of time bombs: scan for SEI, PII, viruses at scale
- Generate lists of contents, file extensions
- Extract filesystem-level metadata
- Document internal context

www.digitalbevaring.dk

Appraisal

- Mount, view contents from obsolete media, systems
- Fuzzy hashing techniques to help address duplication





Arrangement, Description

- Abstract groups of digital files for description
- Supports MPLP-like arrangement
- Export contents for more complex arrangements
- Granular filesystem metadata for creation, modified dates

Access

- Mounting disk images in the reading room
- Document creator's computer environment
- Experiments with browsing disk image via browser
- Facilitates experiments with emulation

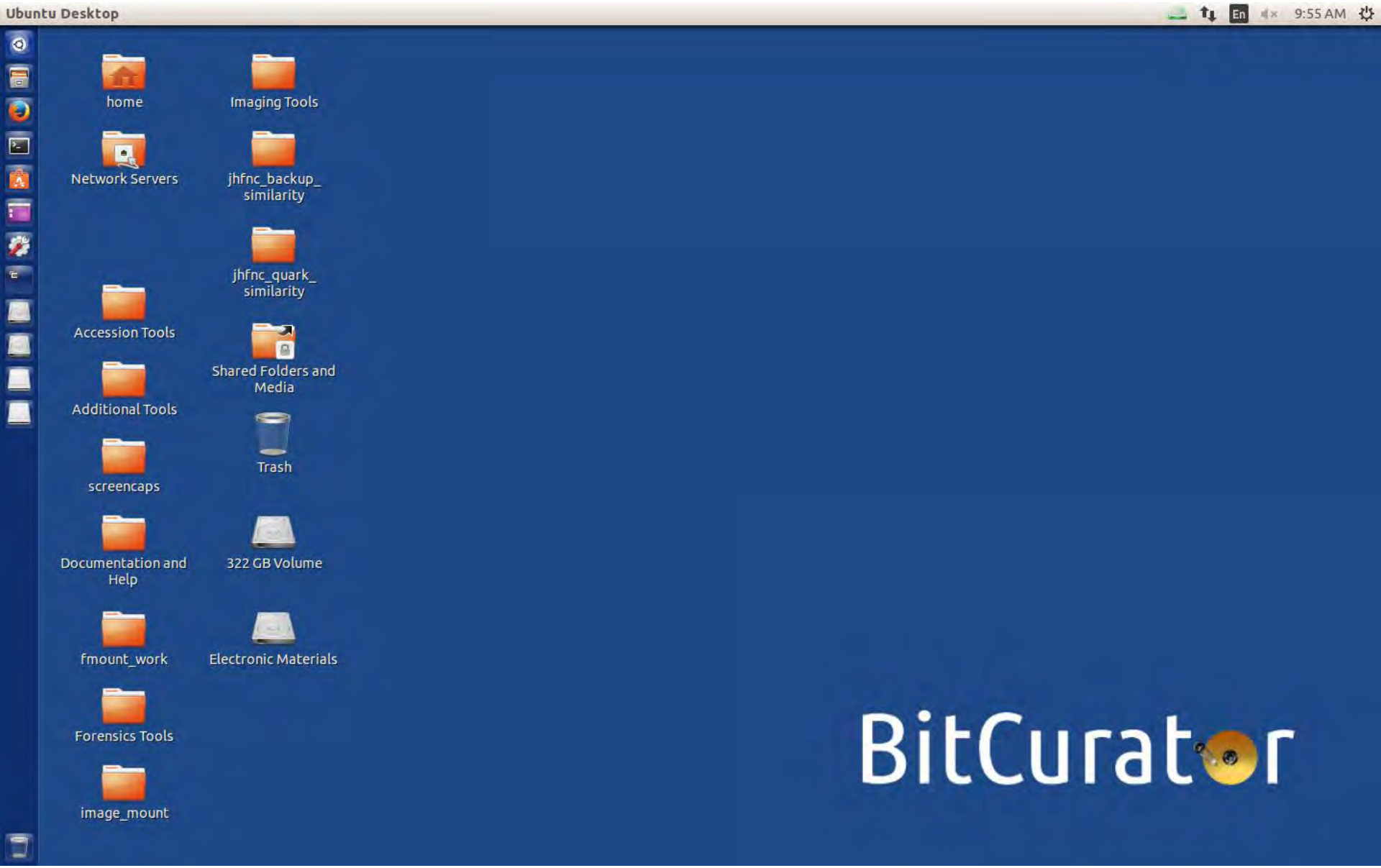
Where to learn more about DF

- SAA Digital Forensics workshops
  - Fundamentals
  - Advanced
- Gengenbach, Martin J. (2012) The way we do it here: Mapping digital forensics workflows in collecting institutions. Masters paper, UNC-Chapel Hill. Link
- CLIR report
  - Digital Forensics and Born-Digital Content in Cultural Heritage Repositories (2010)
  - Born-Digital: Guidelines for Donors, Dealers, and Archival Repositories (2013)
- BitCurator white papers
  - From Bitstreams to Heritage (2013)
  - From Code to Community (2014)

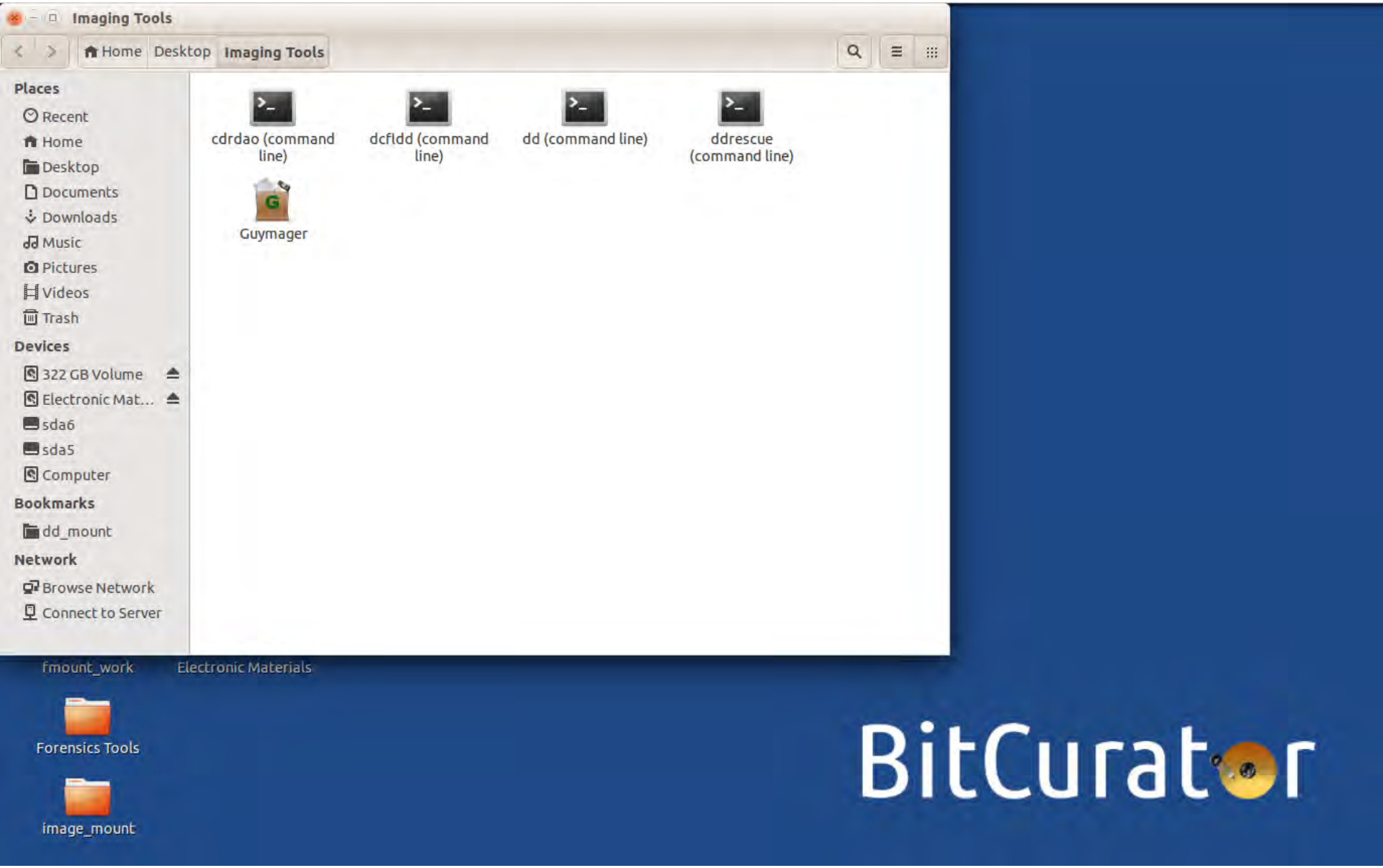
BitCurator Overview

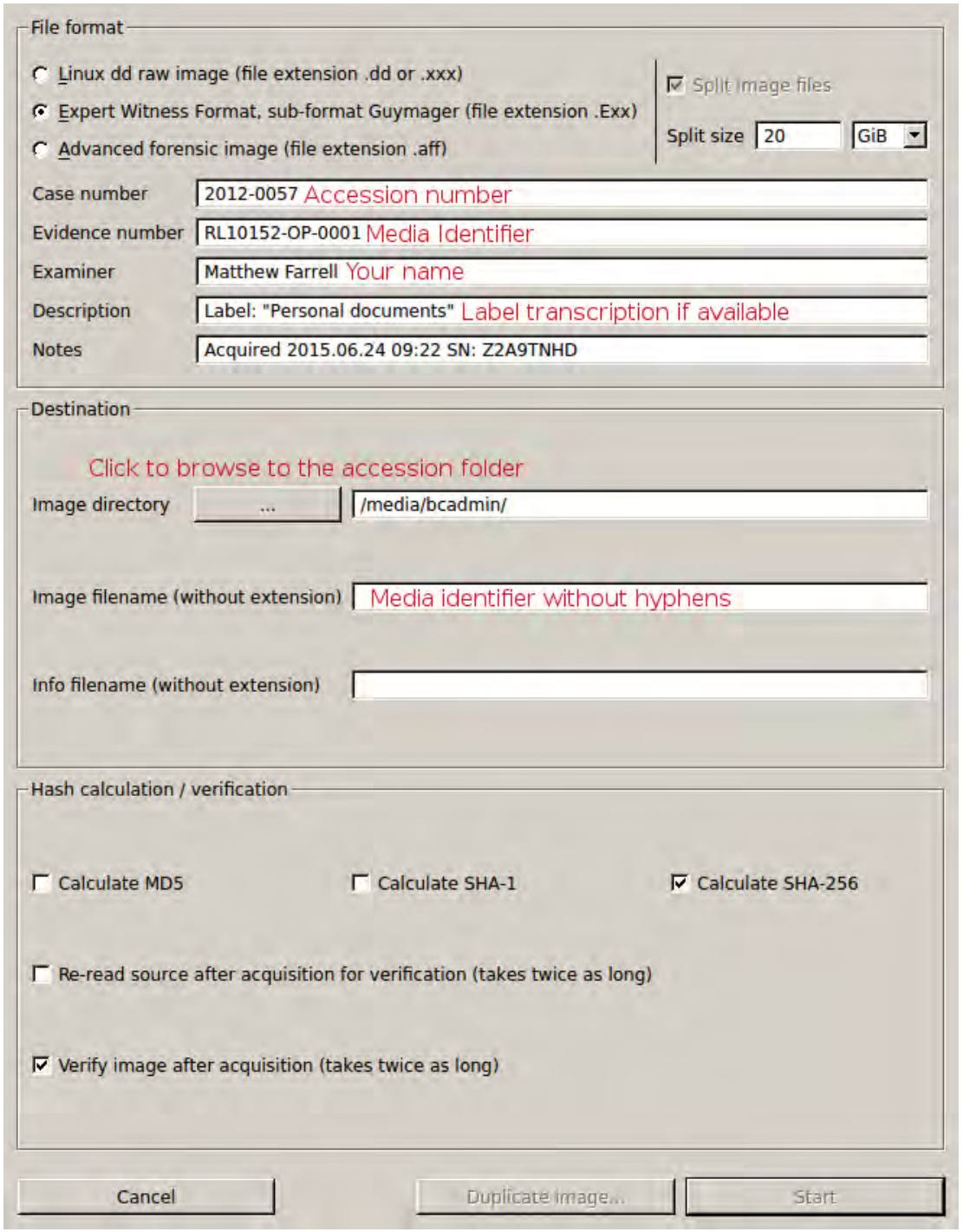
Image: General Library Exterior, before 1948. Source: Duke University Archives. License: cc-by-nc-sa.

Desktop Overview



Imaging Tools





Guymager example I

- Create image dialogue

Guymager Example II

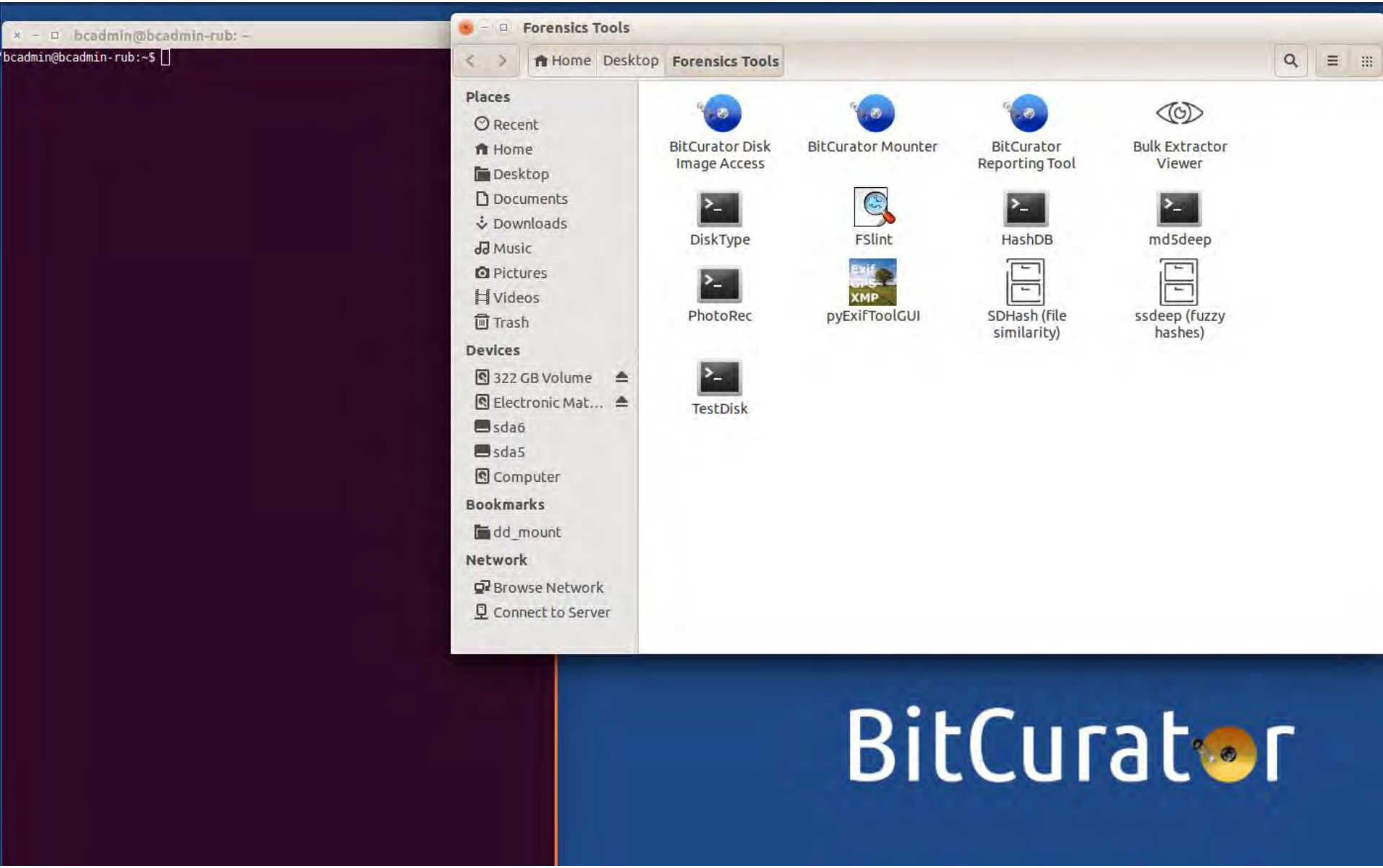
Serial nr.	Linux device	Model	State	Size	Hidden areas	Bad sectors	Progress	Average speed [MB/s]	Time remaining	FIFO qu usag [%]
CN085KRY7363915M53TC	/dev/sr0	PLDS PLDS DVD+/-RW DH-16ABS	Finished	27.5MB	unknown	0	100%	1.38		
Z2A9TNHD	/dev/sda	ATA ST3500413AS	Idle	500.1GB	none					
	/dev/loop0	Linux Loop: RL01262_strickla	Idle	31.4MB	unknown					
	/dev/loop1	Linux Loop: RL01262_strickla	Idle	31.4MB	unknown					
	/dev/loop2	Linux Loop: sda1	Idle	322.1GB	unknown					
	/dev/loop3	Linux Loop: sr0	Idle	27.5MB	unknown					
WD-WCATR7940902	/dev/sdb	ATA WDC WD10EALX-759BA1	Idle	1.0TB	none					

Size	27,498,496 bytes (26.2MiB / 27.5MB)
Sector size	2,048
Image file	/media/bcadmin/Electronic Materials/processing/RL01262_strickland/writings_series/projects/errand_upon_which_we_came/RL01262OP0001.dd
Info file	/media/bcadmin/Electronic Materials/processing/RL01262_strickland/writings_series/projects/errand_upon_which_we_came/RL01262OP0001.info
Current speed	
Started	15. May 13:56:51 (00:00:19)
Hash calculation	SHA-256
Source verification	off
Image verification	off

View attached devices

Progress updates

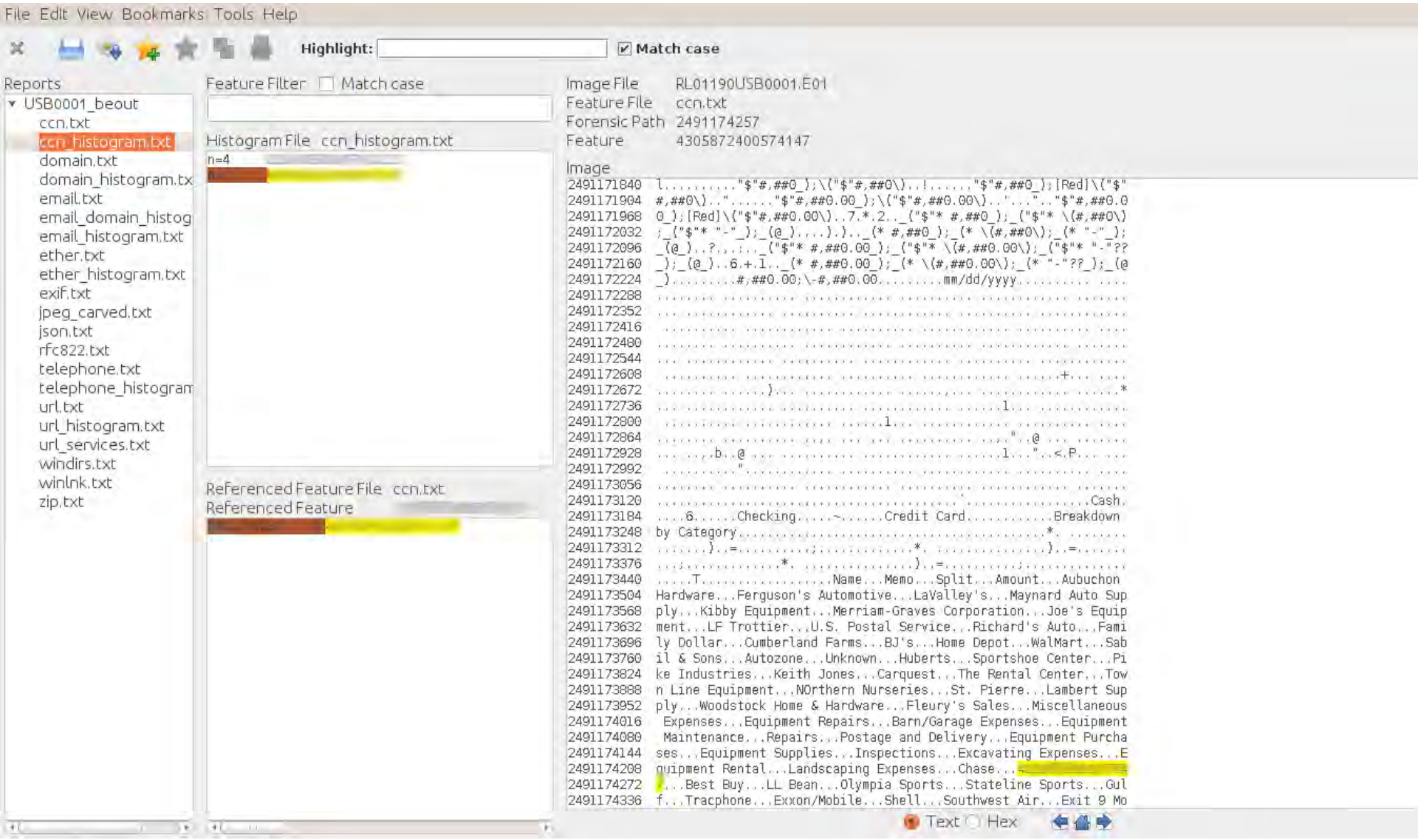
Forensics Tools

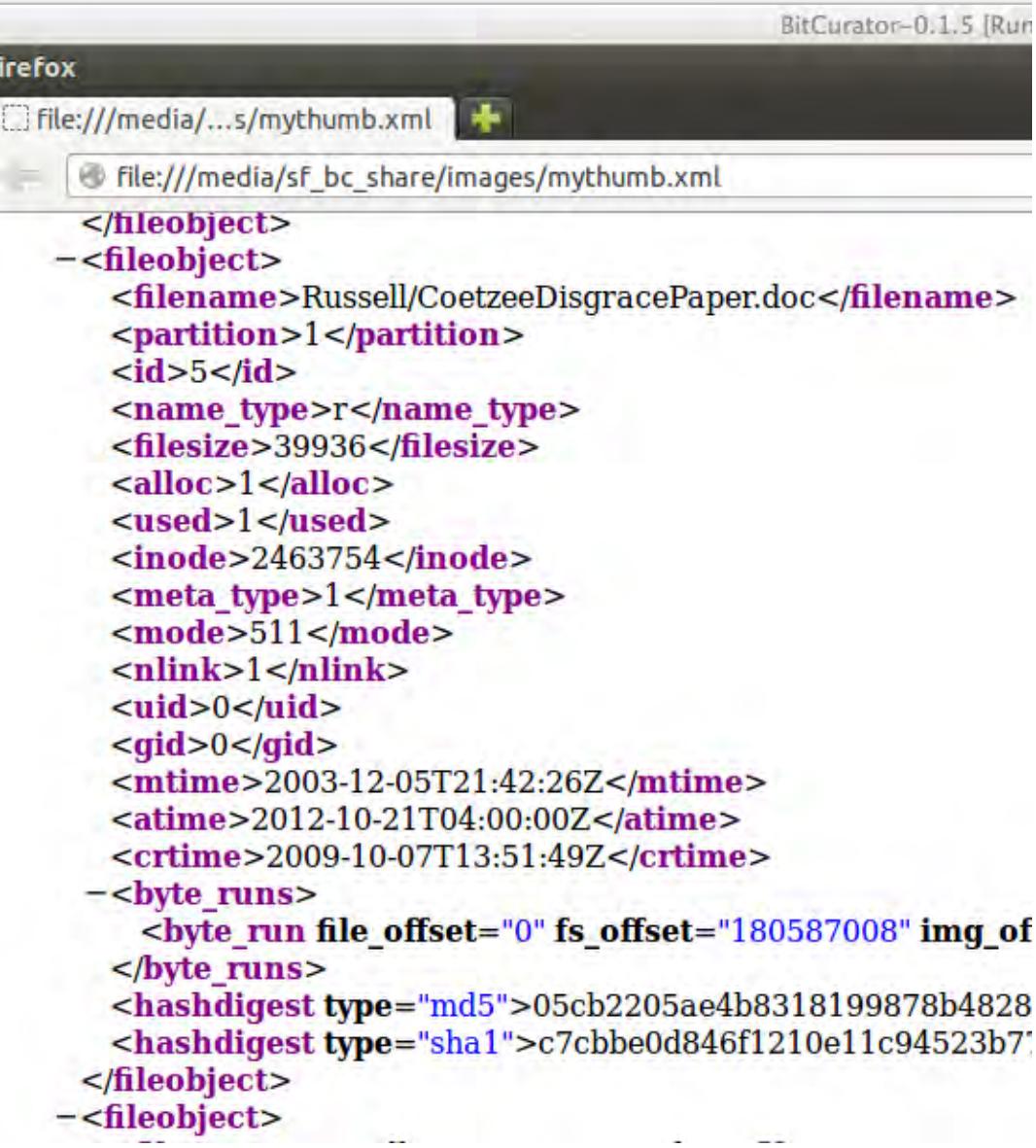


bulk_extractor and Viewer

caption

bulk_extractor's output

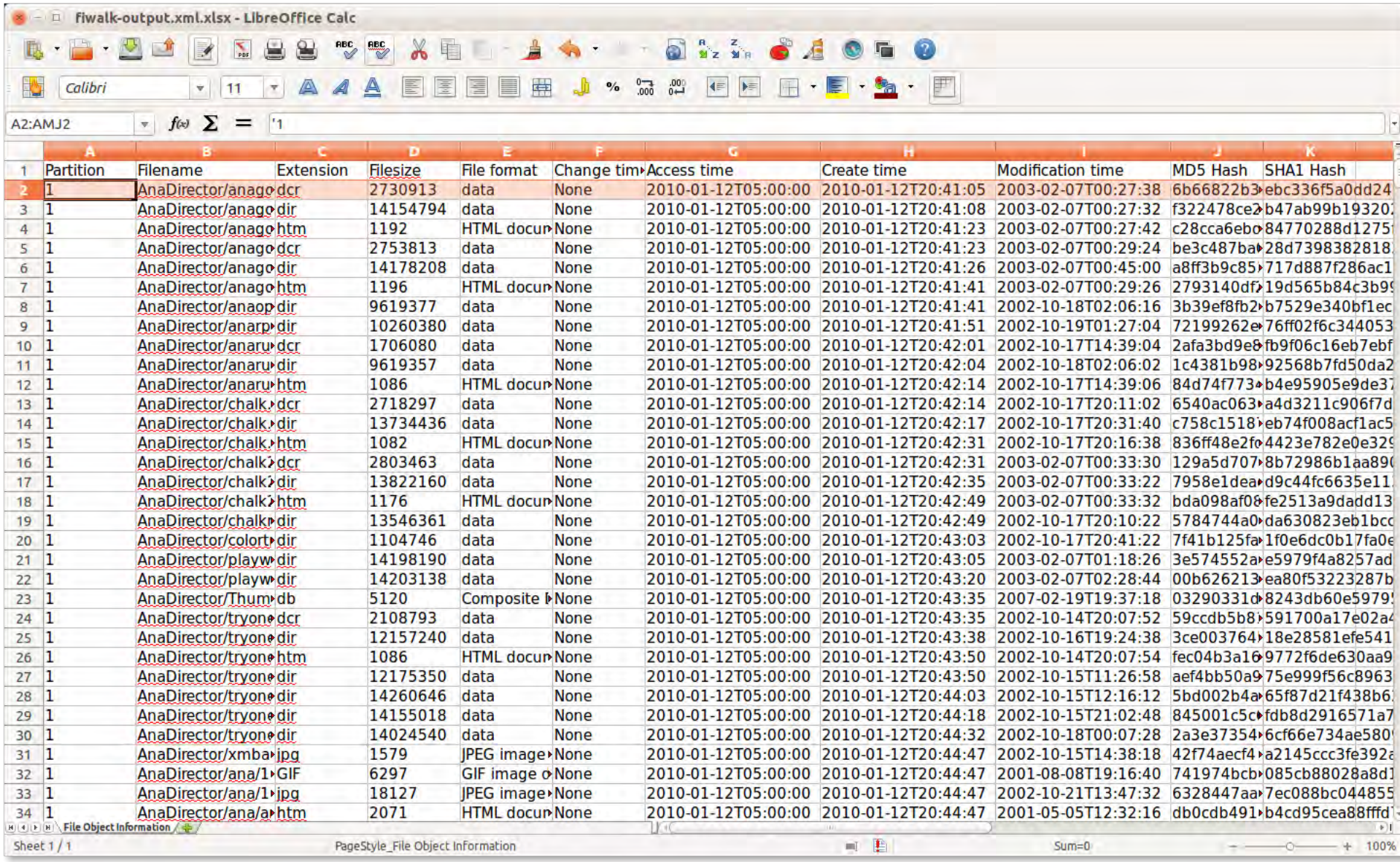




fiwalk

- Filesystem metadata
- Previously command-line only
- Default outputs to DFXML

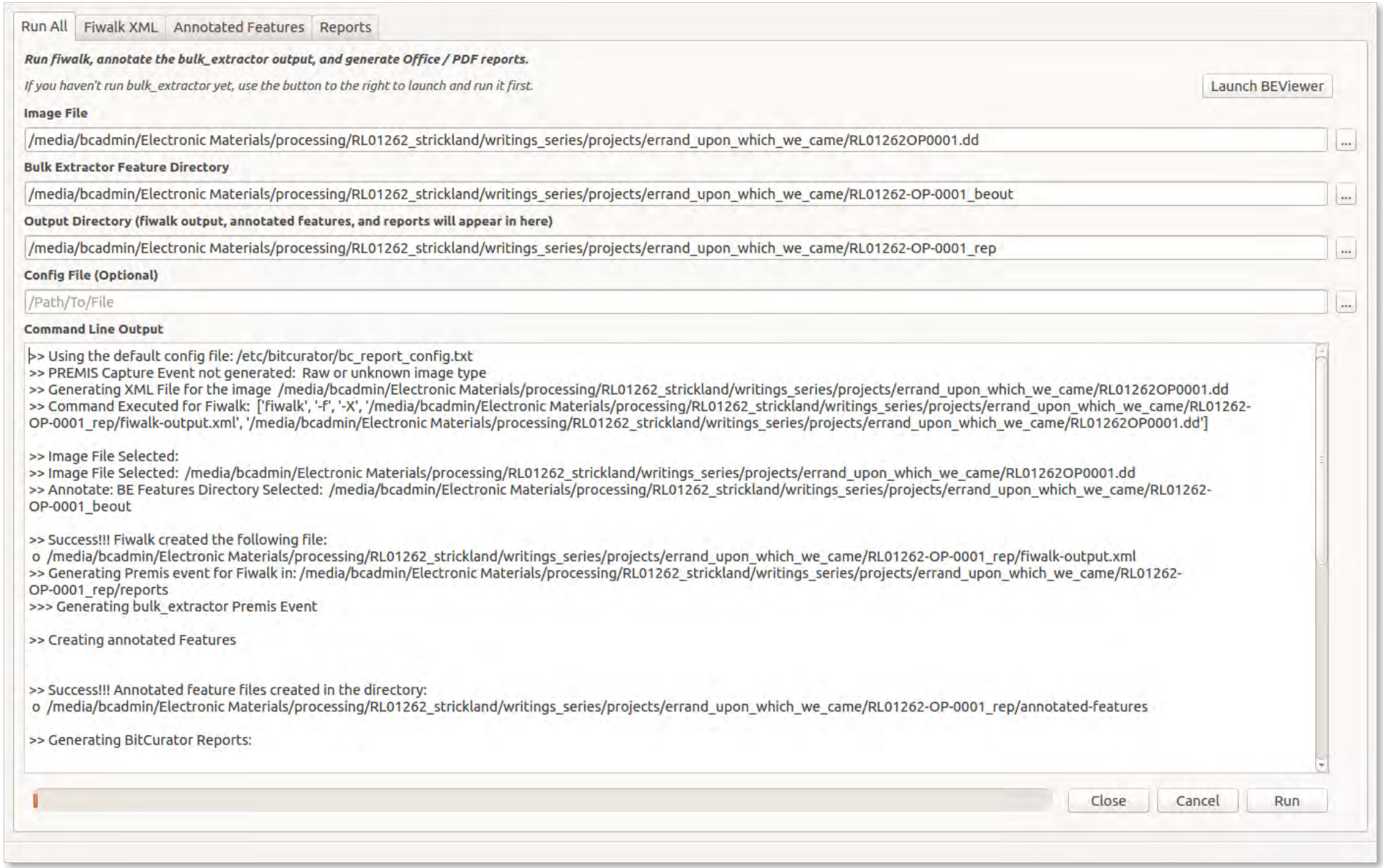
fiwalk converted



	A	B	C	D	E	F	G	H	I	J	K
1	Partition	Filename	Extension	Filesize	File format	Change tim	Access time	Create time	Modification time	MD5 Hash	SHA1 Hash
2	1	AnaDirector/anago	dcr	2730913	data	None	2010-01-12T05:00:00	2010-01-12T20:41:05	2003-02-07T00:27:38	6b66822b3	ebc336f5a0dd24
3	1	AnaDirector/anago	dir	14154794	data	None	2010-01-12T05:00:00	2010-01-12T20:41:08	2003-02-07T00:27:32	f322478ce2	b47ab99b19320
4	1	AnaDirector/anago	htm	1192	HTML docun	None	2010-01-12T05:00:00	2010-01-12T20:41:23	2003-02-07T00:27:42	c28cca6ebc	84770288d1275
5	1	AnaDirector/anago	dcr	2753813	data	None	2010-01-12T05:00:00	2010-01-12T20:41:23	2003-02-07T00:29:24	be3c487ba	28d7398382818
6	1	AnaDirector/anago	dir	14178208	data	None	2010-01-12T05:00:00	2010-01-12T20:41:26	2003-02-07T00:45:00	a8ff3b9c85	717d887f286ac1
7	1	AnaDirector/anago	htm	1196	HTML docun	None	2010-01-12T05:00:00	2010-01-12T20:41:41	2003-02-07T00:29:26	2793140df	19d565b84c3b9
8	1	AnaDirector/anaop	dir	9619377	data	None	2010-01-12T05:00:00	2010-01-12T20:41:41	2002-10-18T02:06:16	3b39ef8fb2	b7529e340bf1ec
9	1	AnaDirector/anarp	dir	10260380	data	None	2010-01-12T05:00:00	2010-01-12T20:41:51	2002-10-19T01:27:04	72199262e	76ff02f6c344053
10	1	AnaDirector/anaru	dcr	1706080	data	None	2010-01-12T05:00:00	2010-01-12T20:42:01	2002-10-17T14:39:04	2afa3bd9e8	fb9f06c16eb7ebf
11	1	AnaDirector/anaru	dir	9619357	data	None	2010-01-12T05:00:00	2010-01-12T20:42:04	2002-10-18T02:06:02	1c4381b98	92568b7fd50da2
12	1	AnaDirector/anaru	htm	1086	HTML docun	None	2010-01-12T05:00:00	2010-01-12T20:42:14	2002-10-17T14:39:06	84d74f773	b4e95905e9de37
13	1	AnaDirector/chalk	dcr	2718297	data	None	2010-01-12T05:00:00	2010-01-12T20:42:14	2002-10-17T20:11:02	6540ac063	a4d3211c906f7d
14	1	AnaDirector/chalk	dir	13734436	data	None	2010-01-12T05:00:00	2010-01-12T20:42:17	2002-10-17T20:31:40	c758c1518	eb74f008acf1ac5
15	1	AnaDirector/chalk	htm	1082	HTML docun	None	2010-01-12T05:00:00	2010-01-12T20:42:31	2002-10-17T20:16:38	836ff48e2fc	4423e782e0e329
16	1	AnaDirector/chalk	dcr	2803463	data	None	2010-01-12T05:00:00	2010-01-12T20:42:31	2003-02-07T00:33:30	129a5d707	8b72986b1aa890
17	1	AnaDirector/chalk	dir	13822160	data	None	2010-01-12T05:00:00	2010-01-12T20:42:35	2003-02-07T00:33:22	7958e1dea	d9c44fc6635e11
18	1	AnaDirector/chalk	htm	1176	HTML docun	None	2010-01-12T05:00:00	2010-01-12T20:42:49	2003-02-07T00:33:32	bda098af08	fe2513a9dadd13
19	1	AnaDirector/chalk	dir	13546361	data	None	2010-01-12T05:00:00	2010-01-12T20:42:49	2002-10-17T20:10:22	5784744a0	da630823eb1bcc
20	1	AnaDirector/colort	dir	1104746	data	None	2010-01-12T05:00:00	2010-01-12T20:43:03	2002-10-17T20:41:22	7f41b125fa	1f0e6dc0b17fa0e
21	1	AnaDirector/playw	dir	14198190	data	None	2010-01-12T05:00:00	2010-01-12T20:43:05	2003-02-07T01:18:26	3e574552a	e5979f4a8257ad
22	1	AnaDirector/playw	dir	14203138	data	None	2010-01-12T05:00:00	2010-01-12T20:43:20	2003-02-07T02:28:44	00b626213	ea80f53223287b
23	1	AnaDirector/Thum	db	5120	Composite I	None	2010-01-12T05:00:00	2010-01-12T20:43:35	2007-02-19T19:37:18	03290331d	8243db60e5979
24	1	AnaDirector/tryon	dcr	2108793	data	None	2010-01-12T05:00:00	2010-01-12T20:43:35	2002-10-14T20:07:52	59ccdb5b8	591700a17e02a4
25	1	AnaDirector/tryon	dir	12157240	data	None	2010-01-12T05:00:00	2010-01-12T20:43:38	2002-10-16T19:24:38	3ce003764	18e28581efe541
26	1	AnaDirector/tryon	htm	1086	HTML docun	None	2010-01-12T05:00:00	2010-01-12T20:43:50	2002-10-14T20:07:54	fec04b3a16	9772f6de630aa9
27	1	AnaDirector/tryon	dir	12175350	data	None	2010-01-12T05:00:00	2010-01-12T20:43:50	2002-10-15T11:26:58	aef4bb50a9	75e999f56c8963
28	1	AnaDirector/tryon	dir	14260646	data	None	2010-01-12T05:00:00	2010-01-12T20:44:03	2002-10-15T12:16:12	5bd002b4a	65f87d21f438b6
29	1	AnaDirector/tryon	dir	14155018	data	None	2010-01-12T05:00:00	2010-01-12T20:44:18	2002-10-15T21:02:48	845001c5c	fdb8d2916571a7
30	1	AnaDirector/tryon	dir	14024540	data	None	2010-01-12T05:00:00	2010-01-12T20:44:32	2002-10-18T00:07:28	2a3e37354	6cf66e734ae580
31	1	AnaDirector/xmba	jpg	1579	JPEG image	None	2010-01-12T05:00:00	2010-01-12T20:44:47	2002-10-15T14:38:18	42f74aecf4	a2145ccc3fe392a
32	1	AnaDirector/ana/1	GIF	6297	GIF image d	None	2010-01-12T05:00:00	2010-01-12T20:44:47	2001-08-08T19:16:40	741974bcb	085cb88028a8d1
33	1	AnaDirector/ana/1	jpg	18127	JPEG image	None	2010-01-12T05:00:00	2010-01-12T20:44:47	2002-10-21T13:47:32	6328447aa	7ec088bc044855
34	1	AnaDirector/ana/a	htm	2071	HTML docun	None	2010-01-12T05:00:00	2010-01-12T20:44:47	2001-05-05T12:32:16	db0cdb491	b4cd95cea88fffd

Spreadsheet transformation

Reporting Tool and outputs I



Reporting Tool and outputs II

annotated_ccn.txt × | annotated_rfc822.txt ×

```
1 # Position      Feature Context Filename        MD5
2 # /home/bcadmin/Tools/bulk_extractor/python/identify_filenames.py --all --image_filename /media/bcadmin/Electronic Materials/processing/
  RL01190_simmons/RL01190USB0001.E01 /media/bcadmin/Electronic Materials/processing/RL01190_simmons/USB0001_beout /media/bcadmin/Electronic Materials/
  processing/RL01190_simmons/USB0001_reports/annotated-features
3 # BANNER FILE NOT PROVIDED (-b option)
4 # BULK_EXTRACTOR-Version: 1.5.5-dev ($Rev: 10844 $)
5 # Feature-Recorder: ccn
6 # Filename: /media/bcadmin/Electronic Materials/processing/RL01190_simmons/RL01190USB0001.E01
7 # Feature-File-Version: 1.1
8 1440854999      [redacted]      SOPQQMIE>;9:950/[redacted]'\x1D\x12\x0D\x0D\x0E\x0E\x0D\x0C\x0E\x0F\x0F\x12\x12\x10\x0D
9 1449268982      [redacted]      SOPQQMIE>;9:950/[redacted]'\x1D\x12\x0D\x0D\x0E\x0E\x0D\x0C\x0E\x0F\x0F\x12\x12\x10\x0D
10 2491174257     [redacted]      enses\x05\x00\x00Chase\x10\x00\x00[redacted],x08\x00\x00Best Buy\x07\x00\x00LL
11 2867065815     [redacted]      SOPQQMIE>;9:950/[redacted]'\x1D\x12\x0D\x0D\x0E\x0E\x0D\x0C\x0E\x0F\x0F\x12\x12\x10\x0D
12 2875479798     [redacted]      SOPQQMIE>;9:950/[redacted]'\x1D\x12\x0D\x0D\x0E\x0E\x0D\x0C\x0E\x0F\x0F\x12\x12\x10\x0D
13 # Total features input: 5
14 # Total features located to files: 0
15 # Total features in unallocated space: 5
16 # Total features in encoded regions: 0
17 # Total processing time: 0.00024 seconds
```

- Maps features ID'd by bulk_extractor to specific files, where possible
- Reports which features apply to files and which do not

BitCurator Disk Image Access tool

Reporting Tool and outputs III

Name	Size	Type
Chapter 1 Saraband for a Saint 1-26.doc	10.6 MB	Docum
Chapter 2 The Gypsy Condesa 27-42.doc	1.4 MB	Docum
Chapter 3 Meeting Isabel 43-60.doc	105.0 kB	Docum
Chapter 4 Charleston 61-76.doc	99.8 kB	Docum
Chapter 5 Saint Teresa 77-99.doc	135.7 kB	Docum
Chapter 6 Chronology 100-112.doc	247.8 kB	Docum
Chapter 7 The Sackvilles & Other Connections 113-152....	1.4 MB	Docum
Chapter 8 Orlando 153-182.doc	178.7 kB	Docum
Chapter 9 Ballad of the Sad Cafe Where it Came From ...	704.5 kB	Docum
Chapter 10 Ballad The Town The Time The People 219-2...	53.8 kB	Docum
Chapter 11 Ballad The Cast 228-243.doc	471.6 kB	Docum
Chapter 12 Ballad A Theory of Love 244-250.doc	47.6 kB	Docum
Chapter 13 Ballad The Tale 251-275.doc	429.1 kB	Docum
Chapter 15 Ballad Afterwards 276-282.doc	46.6 kB	Docum
Chapter 16 Ballad Sources 283-299.doc	86.0 kB	Docum
Chapter 17 Ballad The Film 300-321.doc	114.2 kB	Docum
Chapter 18 The Transsexual Phenomenon 322-378.doc	2.8 MB	Docum
Chapter 19 The Last Book The Last Mentor 379-416.doc	427.0 kB	Docum
Chapter 20 Me Papoose Sitter Revisited 417-443.doc	1.3 MB	Docum

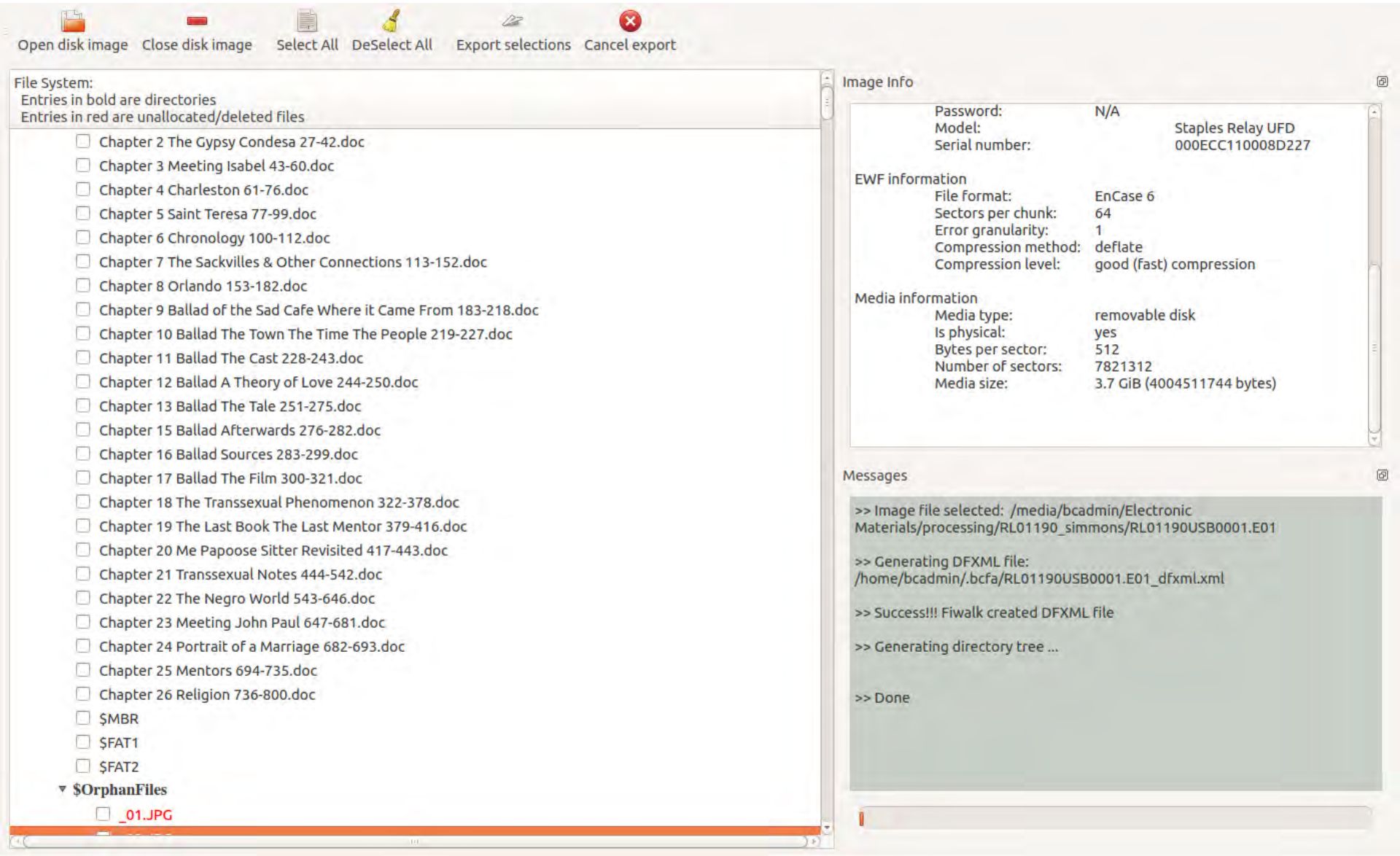
Report: File System Statistics and Files

Deleted Files

Disk Image: RL01190USB0001.E01

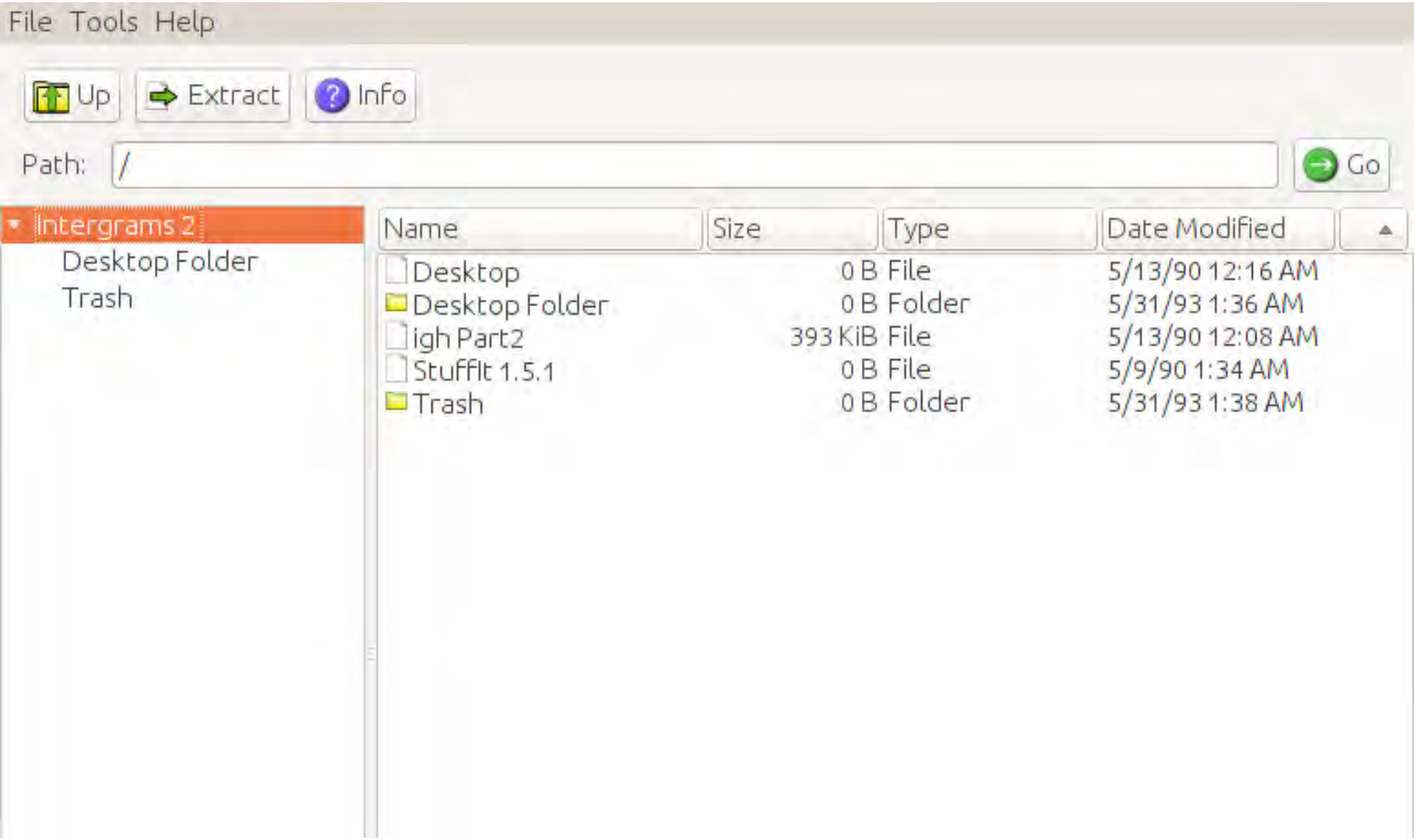
Partitio	Deleted File
1	$OrphanFiles
1	$OrphanFiles/_01.JPG
1	$OrphanFiles/_02.JPG
1	$OrphanFiles/_03.JPG
1	$OrphanFiles/_04.JPG
1	$OrphanFiles/_05.JPG
1	$OrphanFiles/_06.JPG
1	$OrphanFiles/_07.JPG
1	$OrphanFiles/_08.JPG
1	$OrphanFiles/_09.JPG
1	$OrphanFiles/_10.JPG
1	$OrphanFiles/_11.JPG
1	$OrphanFiles/_12.JPG
1	$OrphanFiles/_13.JPG
1	$OrphanFiles/_14.JPG
1	$OrphanFiles/_15.JPG
1	$OrphanFiles/_16.JPG
1	$OrphanFiles/_17.JPG
1	$OrphanFiles/_18.JPG
1	$OrphanFiles/_19.JPG
1	$OrphanFiles/_20.JPG

Disk Image Access tools I



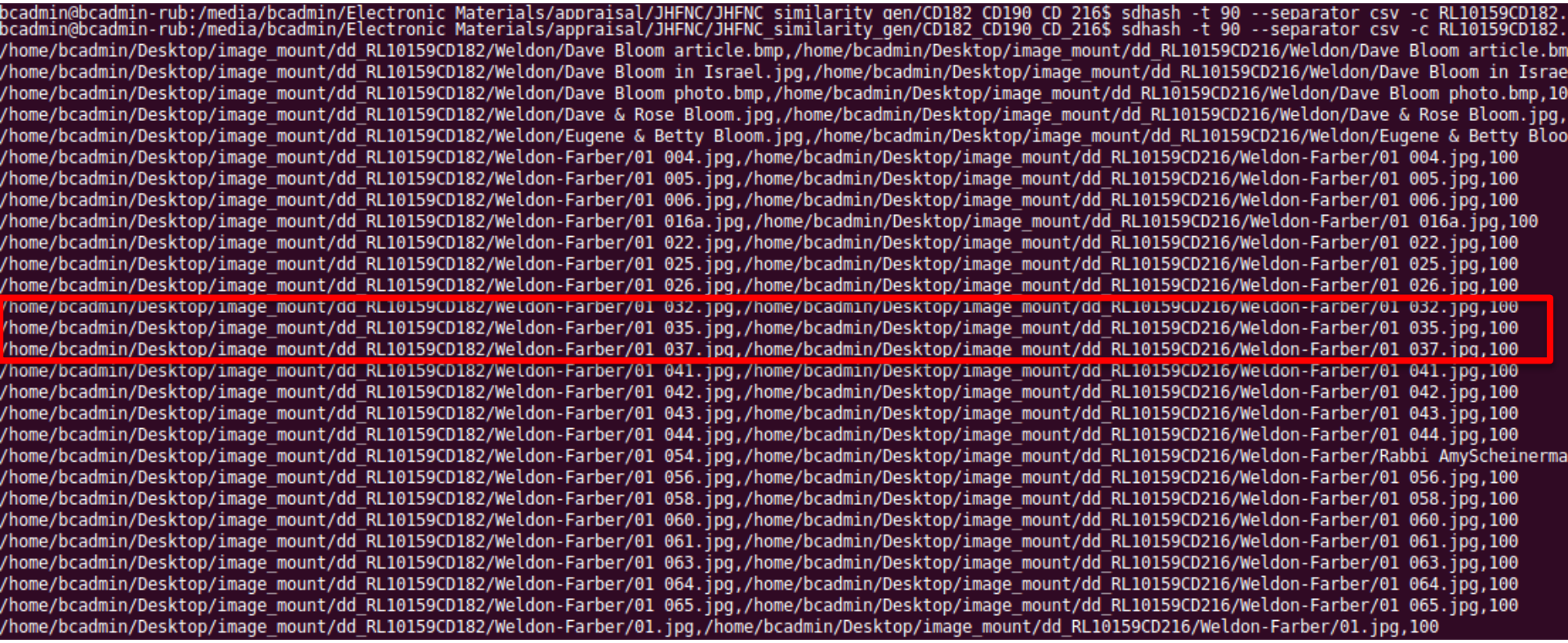
BitCurator Disk Image Access tool

Disk Image Access tools II



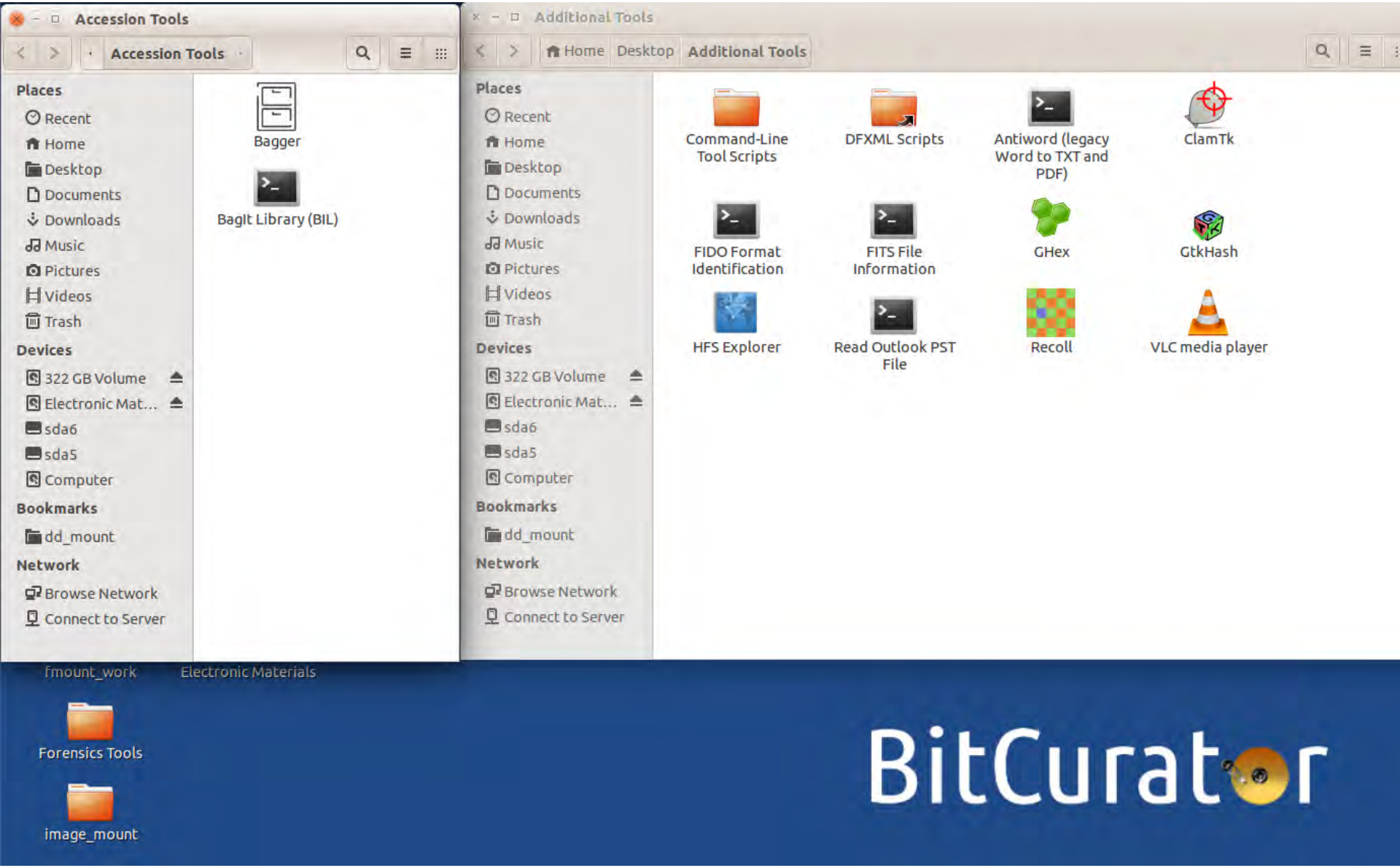
HFS Explorer

De-duplication, file similarity



- Hashing, fuzzy hashing techniques
- FSlint
- sdhash, ssdeep

Additional tools



Decision points: VM or Native OS?

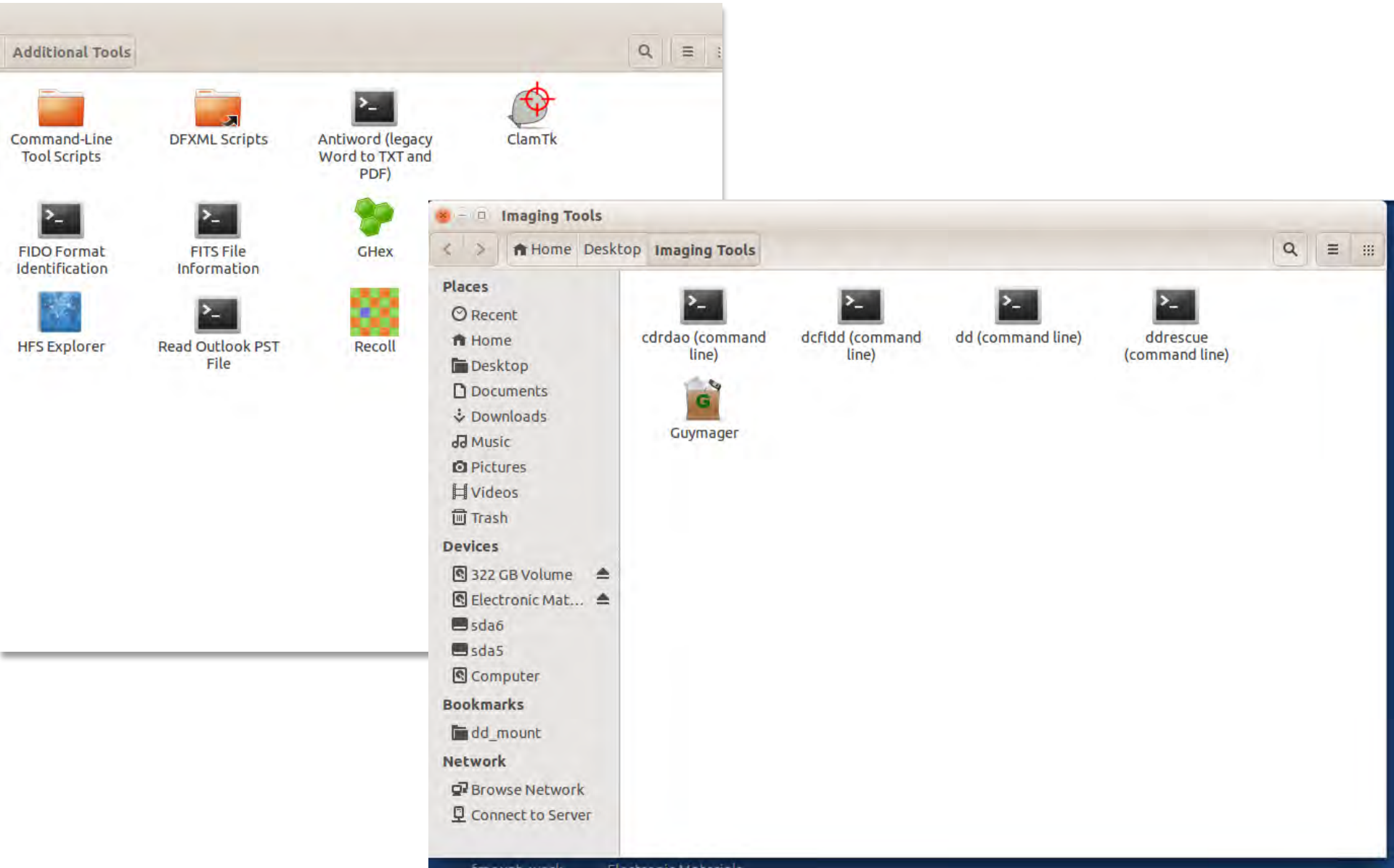
VM

- Pros
  - Update with each new release quickly
  - Distribute amongst staff
- Cons
  - VirtualBox, VMware support issues
  - Resource sharing
  - Virtualized storage

Native OS

- Pros
  - Run forensic processes more efficiently
  - Less virtualized storage
- Cons
  - Linux learning curve
  - Support at your institution
  - Likely that BC can't do *everything* you want

Decision points: Which tools to use?

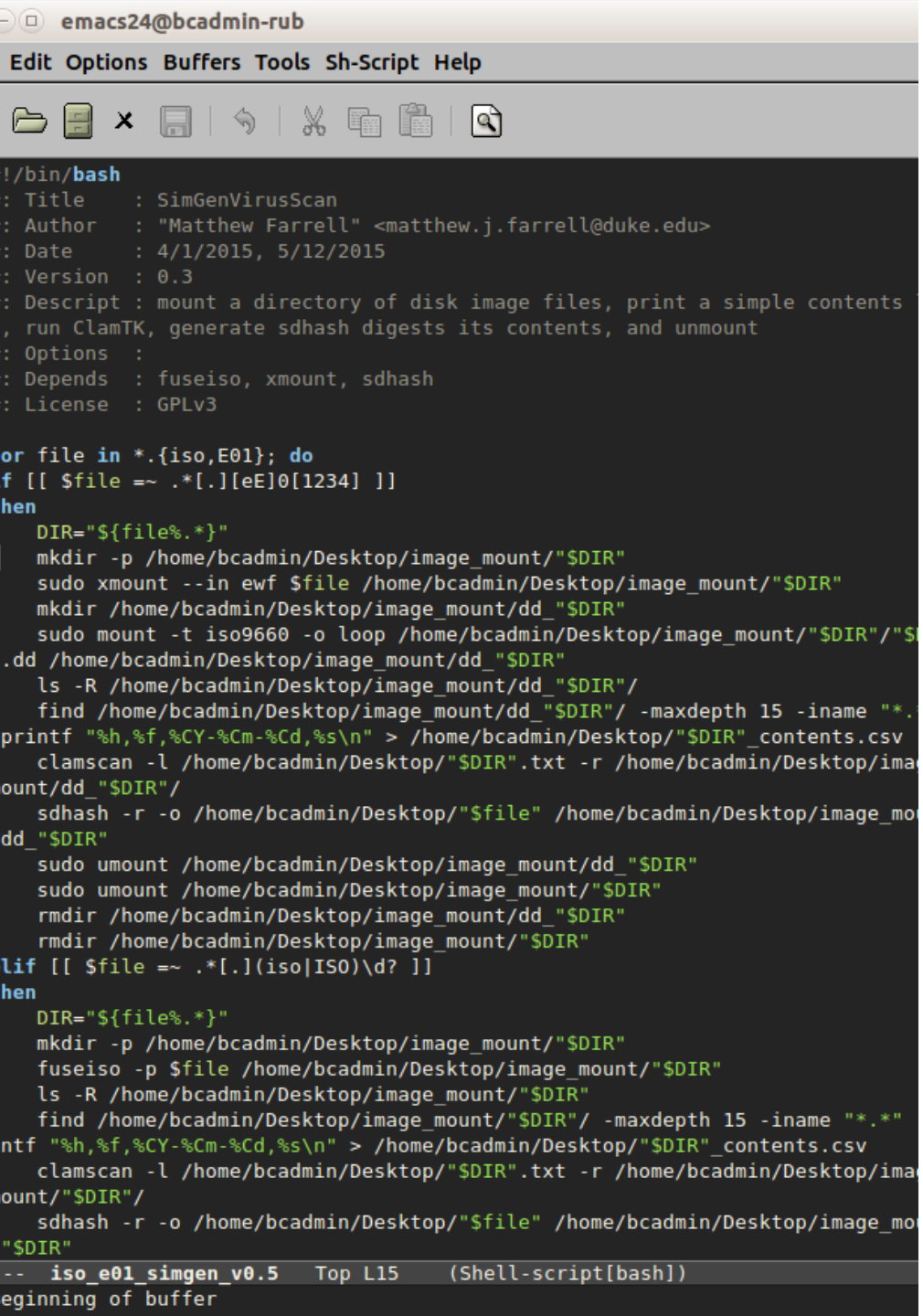


Specific Uses at Duke

Image: General Library Exterior, before 1948. Source: Duke University Archives. License: cc-by-nc-sa.

Acquisitions: Imaging

- Decide per collection, Windows or BC for imaging
- Question: who is doing the imaging?
- Floppies—Windows
- Most everything else—BC

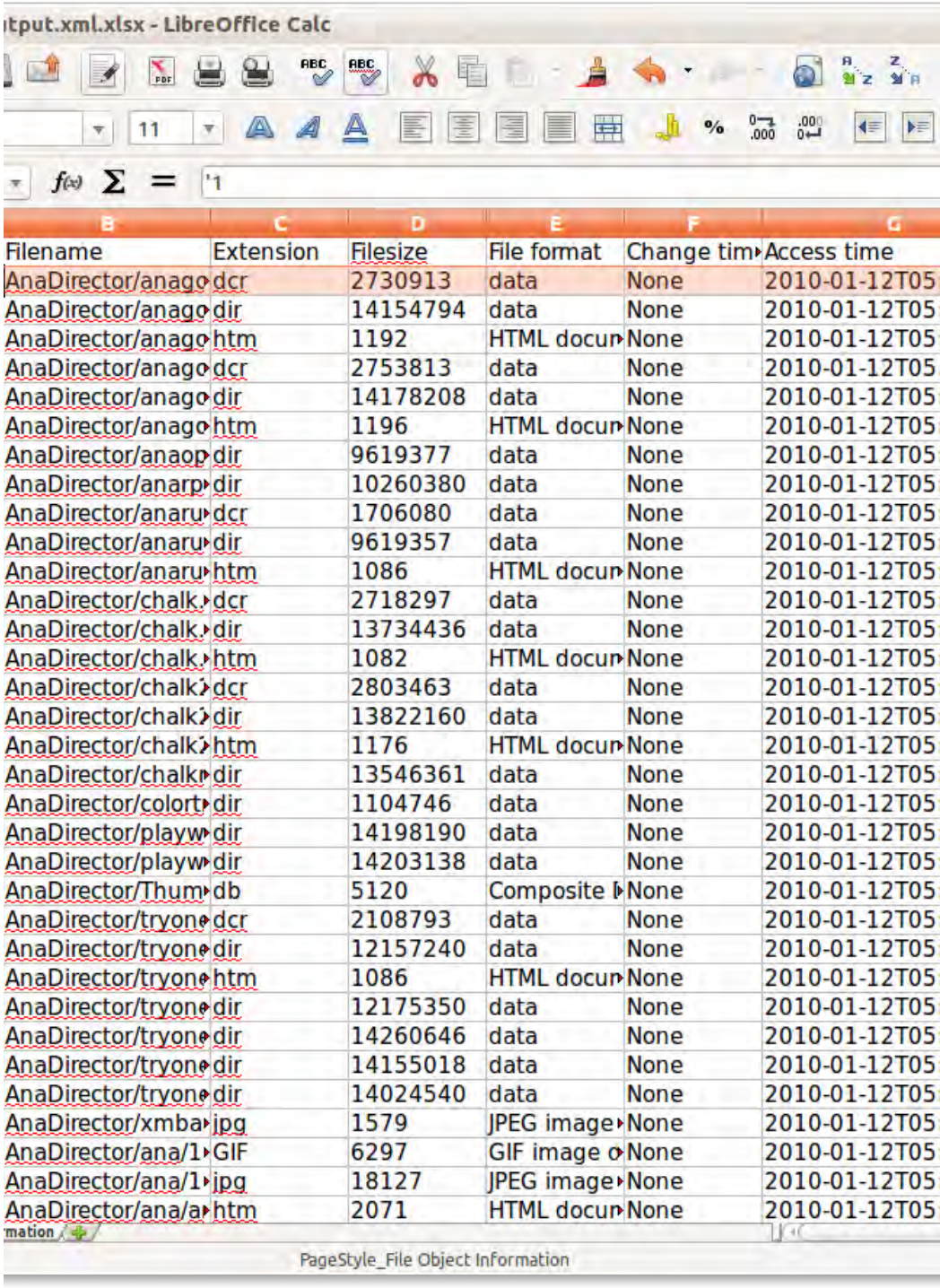


Acquisitions: Basic Reporting

- Script to mount disk image, create directory printout, run virus scan, create fuzzy hashes
- bulk_extractor for sensitive content

Analysis for Arrangement and Description

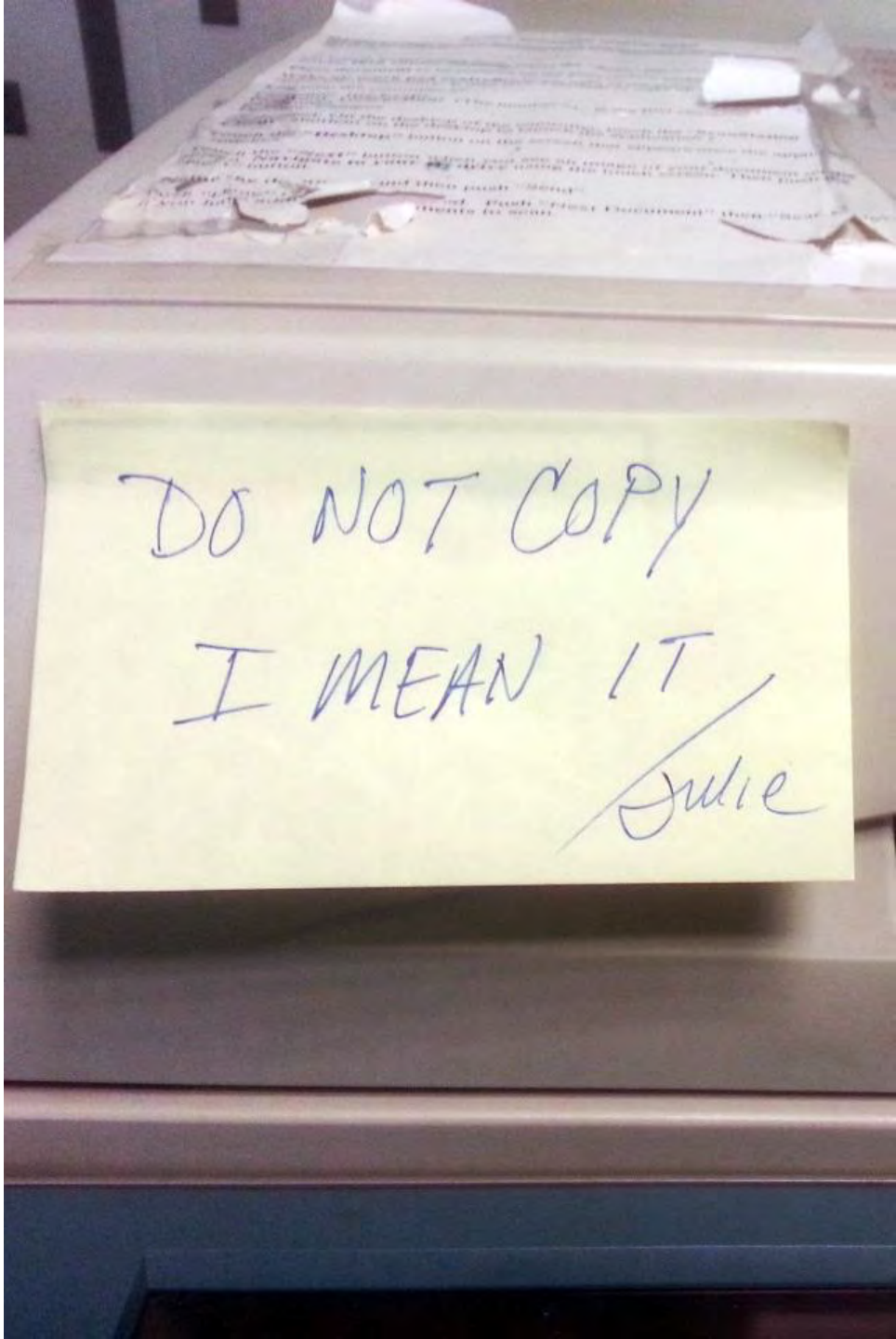
- Fiwalk reports
- FITS
- Disk image mounting

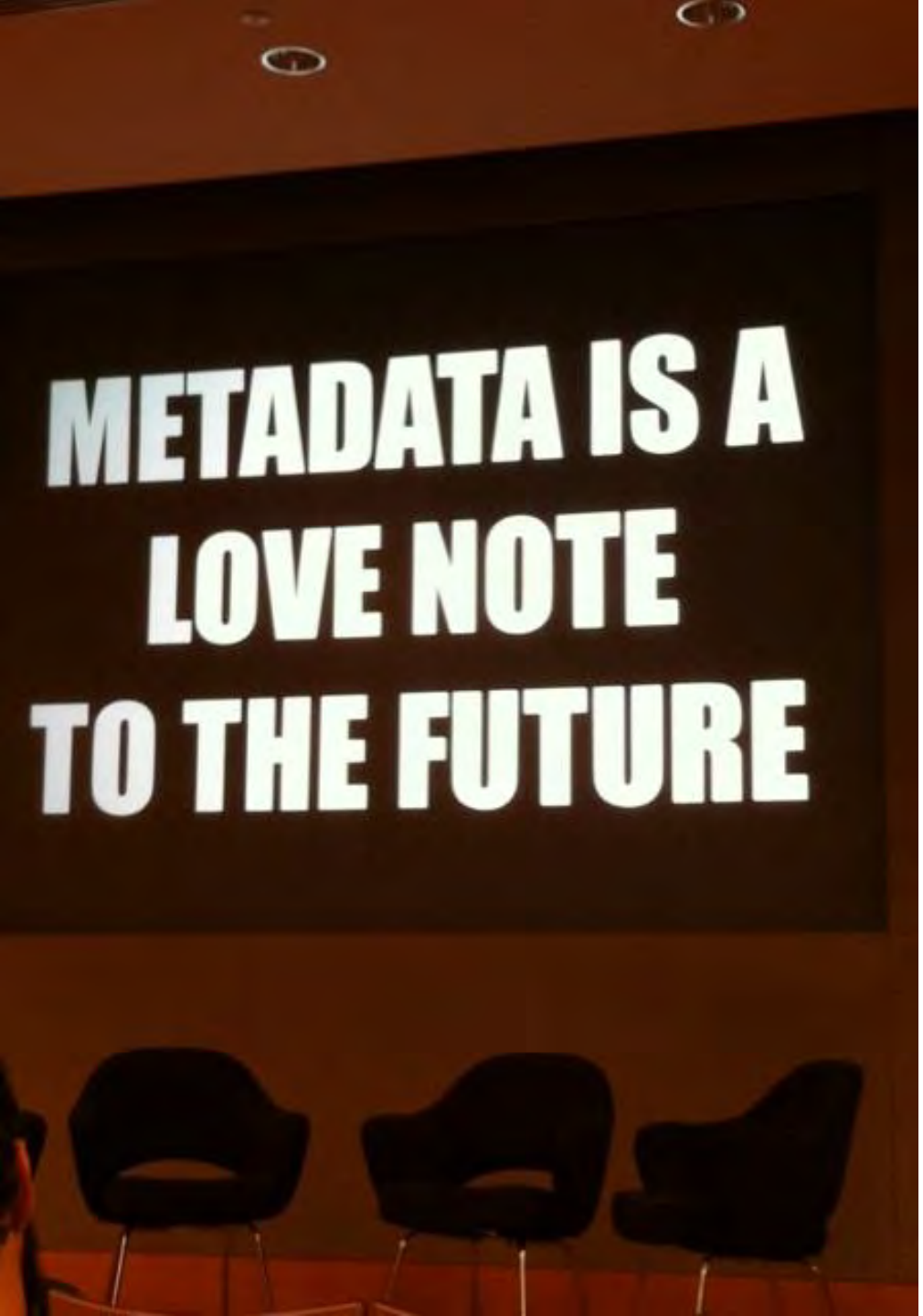


Filename	Extension	Filesize	File format	Change tim	Access time
AnaDirector/anago	dcr	2730913	data	None	2010-01-12T05
AnaDirector/anago	dir	14154794	data	None	2010-01-12T05
AnaDirector/anago	htm	1192	HTML docun	None	2010-01-12T05
AnaDirector/anago	dcr	2753813	data	None	2010-01-12T05
AnaDirector/anago	dir	14178208	data	None	2010-01-12T05
AnaDirector/anago	htm	1196	HTML docun	None	2010-01-12T05
AnaDirector/anaop	dir	9619377	data	None	2010-01-12T05
AnaDirector/anarp	dir	10260380	data	None	2010-01-12T05
AnaDirector/anaru	dcr	1706080	data	None	2010-01-12T05
AnaDirector/anaru	dir	9619357	data	None	2010-01-12T05
AnaDirector/anaru	htm	1086	HTML docun	None	2010-01-12T05
AnaDirector/chalk	dcr	2718297	data	None	2010-01-12T05
AnaDirector/chalk	dir	13734436	data	None	2010-01-12T05
AnaDirector/chalk	htm	1082	HTML docun	None	2010-01-12T05
AnaDirector/chalk	dcr	2803463	data	None	2010-01-12T05
AnaDirector/chalk	dir	13822160	data	None	2010-01-12T05
AnaDirector/chalk	htm	1176	HTML docun	None	2010-01-12T05
AnaDirector/chalk	dir	13546361	data	None	2010-01-12T05
AnaDirector/colort	dir	1104746	data	None	2010-01-12T05
AnaDirector/playw	dir	14198190	data	None	2010-01-12T05
AnaDirector/playw	dir	14203138	data	None	2010-01-12T05
AnaDirector/Thum	db	5120	Composite I	None	2010-01-12T05
AnaDirector/tryone	dcr	2108793	data	None	2010-01-12T05
AnaDirector/tryone	dir	12157240	data	None	2010-01-12T05
AnaDirector/tryone	htm	1086	HTML docun	None	2010-01-12T05
AnaDirector/tryone	dir	12175350	data	None	2010-01-12T05
AnaDirector/tryone	dir	14260646	data	None	2010-01-12T05
AnaDirector/tryone	dir	14155018	data	None	2010-01-12T05
AnaDirector/tryone	dir	14024540	data	None	2010-01-12T05
AnaDirector/xmba	jpg	1579	JPEG image	None	2010-01-12T05
AnaDirector/ana/1	GIF	6297	GIF image c	None	2010-01-12T05
AnaDirector/ana/1	jpg	18127	JPEG image	None	2010-01-12T05
AnaDirector/ana/a	htm	2071	HTML docun	None	2010-01-12T05

Creation of Copies

- Disk image access tools to create access sets
- Export subsets for more complex arrangements





Future Steps

- Further work on similarity automation
- Leverage DFXML for other systems
- Expand use to other members of TS

Image: Metadata is a love note to the future, 2012.
Source: Flickr. License: cc-by-2.0.

BitCurator CONSORTIUM

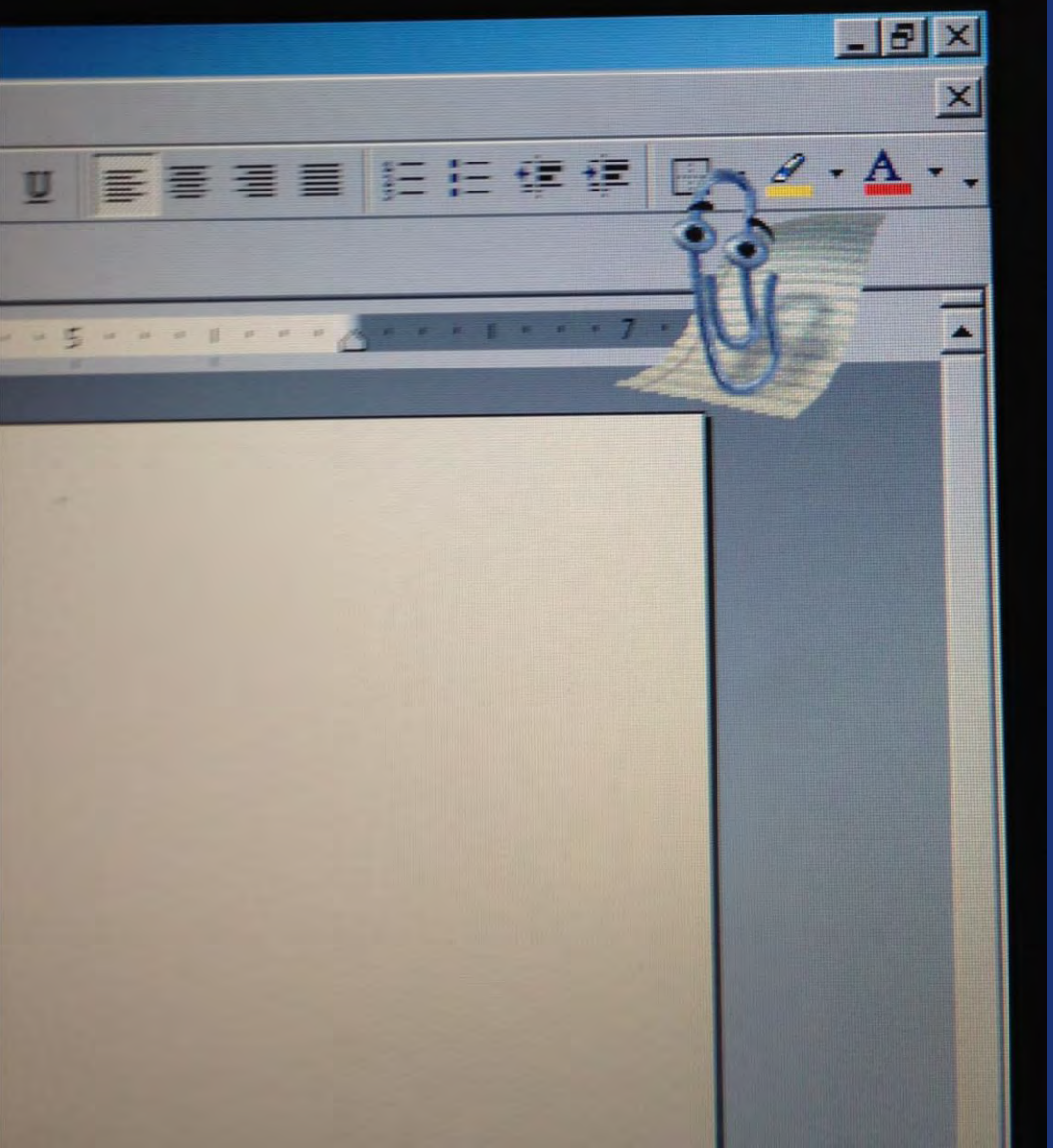
Current Membership

- British Library
- Duke University
- Emory University
- Harvard University
- Maryland Institute for Technology in the Humanities
- Massachusetts Institute for Technology
- New York University
- School of Information & Library Science, UNC-CH
- Princeton University
- Stanford University
- University of Maryland Libraries
- University of Virginia
- McMaster University
- Northwestern University
- Texas State Libraries and Archives Commission
- University of Colorado Boulder
- University of Manchester
- Yale University
- Penn State University

Recent News and Plans

- August 2015: Consortium Manager will join
- FY2016: Formalize the BCC Help Desk
- Fall 2015: launch BCC website
  - Workflow exchange
  - Script repository
- Fall-Winter 2015: compiling, documenting workflows
- January 2016: 2nd Annual BitCurator User Forum, UNC-Chapel Hill

Matthew Farrell
matthew.j.farrell@duke.edu
@laissezfarrell



Thanks!