Identification and Redaction of Sensitive Information in Born-Digital Archival Materials: Research and Development Directions

Christopher (Cal) Lee
UNC School of Information and Library Science

Society of American Archivists Research Forum
Atlanta, GA
August 2, 2016

BitCurator Access

UNC
SCHOOL OF INFORMATION AND LIBRARY SCIENCE

The Andrew W. Mellon Foundation

Sensitive Stuff

- Archivists must often identify information within their collections (either in their existing holdings or as part of new acquisitions) that is sensitive and should thus be removed, closed, restricted or filtered from public view.
- Manually identifying such information can be time-consuming and prone to error.
- Luckily, there are many automated methods that archivists can use to support this work*

*Lee, Christopher A., and Kam Woods. "Automated Redaction of Private and Personal Data in Collections: Toward Responsible Stewardship of Digital Heritage." In *Proceedings of Memory of the World in the Digital Age: Digitization and Preservation: An International Conference on Permanent Access to Digital Documentary Heritage, 26-28 September 2012, Vancouver, British Columbia, Canada*, edited by Luciana Duranti and Elizabeth Shaffer, 298-313: United Nations Educational, Scientific and Cultural Organization, 2013.

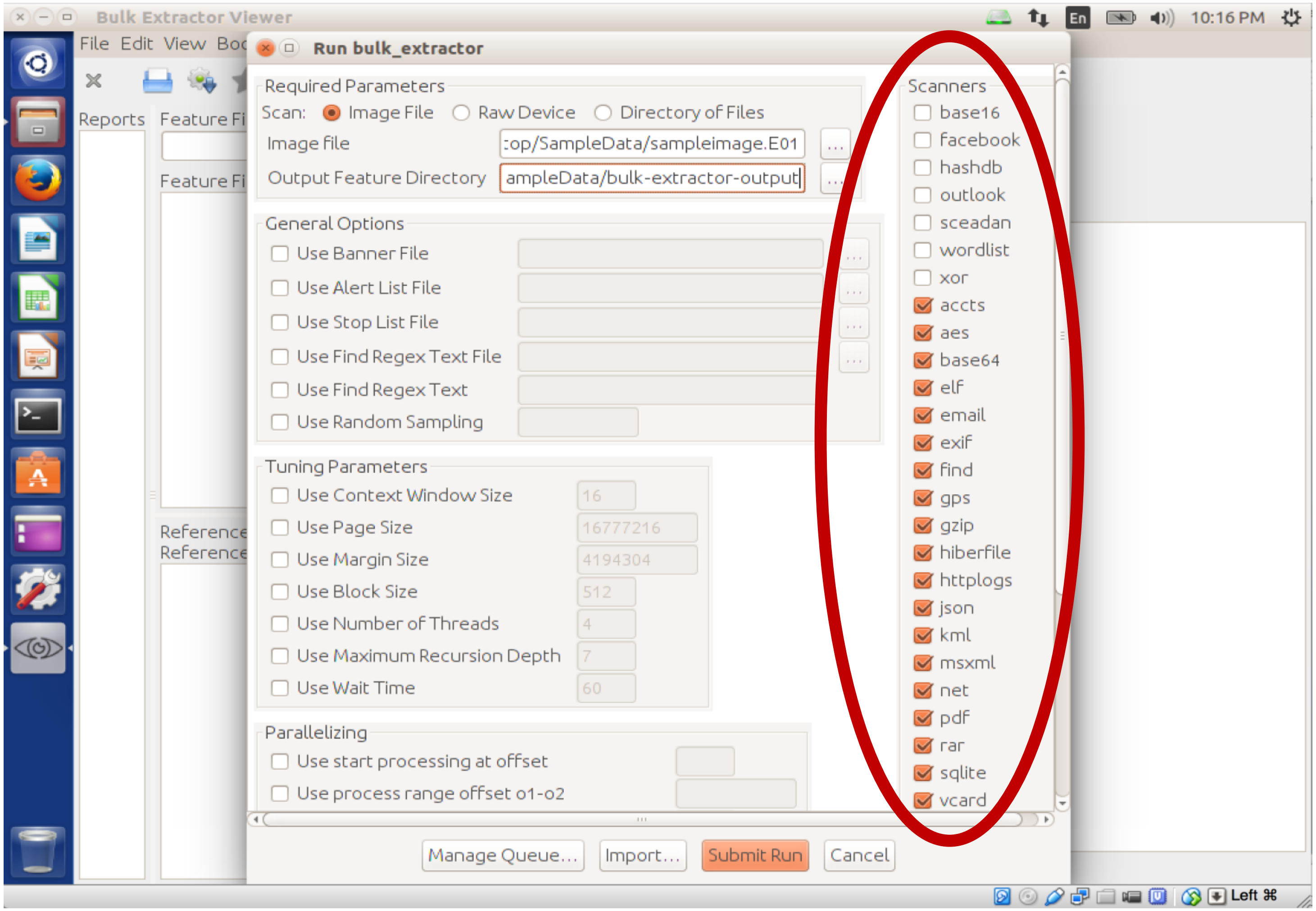
Computer-Assisted Redaction

- Still a relatively undeveloped area of archival practice
- Good news for us: numerous opportunities for further research and development

Examples of Potentially Private and Sensitive Information

- Personal identifiers (e.g. SSNs, DOBs, Drivers License #s, corp. and govt. IDs)
- Financial information (e.g. credit card numbers)
- Geolocation data
- Email messages, email addresses, attachments
- Traces of online activity (e.g. search histories, web caches, domain names, IP addresses)
- Recoverable data from deleted files
- Partially overwritten data

We know how to look for these things…



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
- □ 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
- □ Use process range offset o1-o2

Scanners

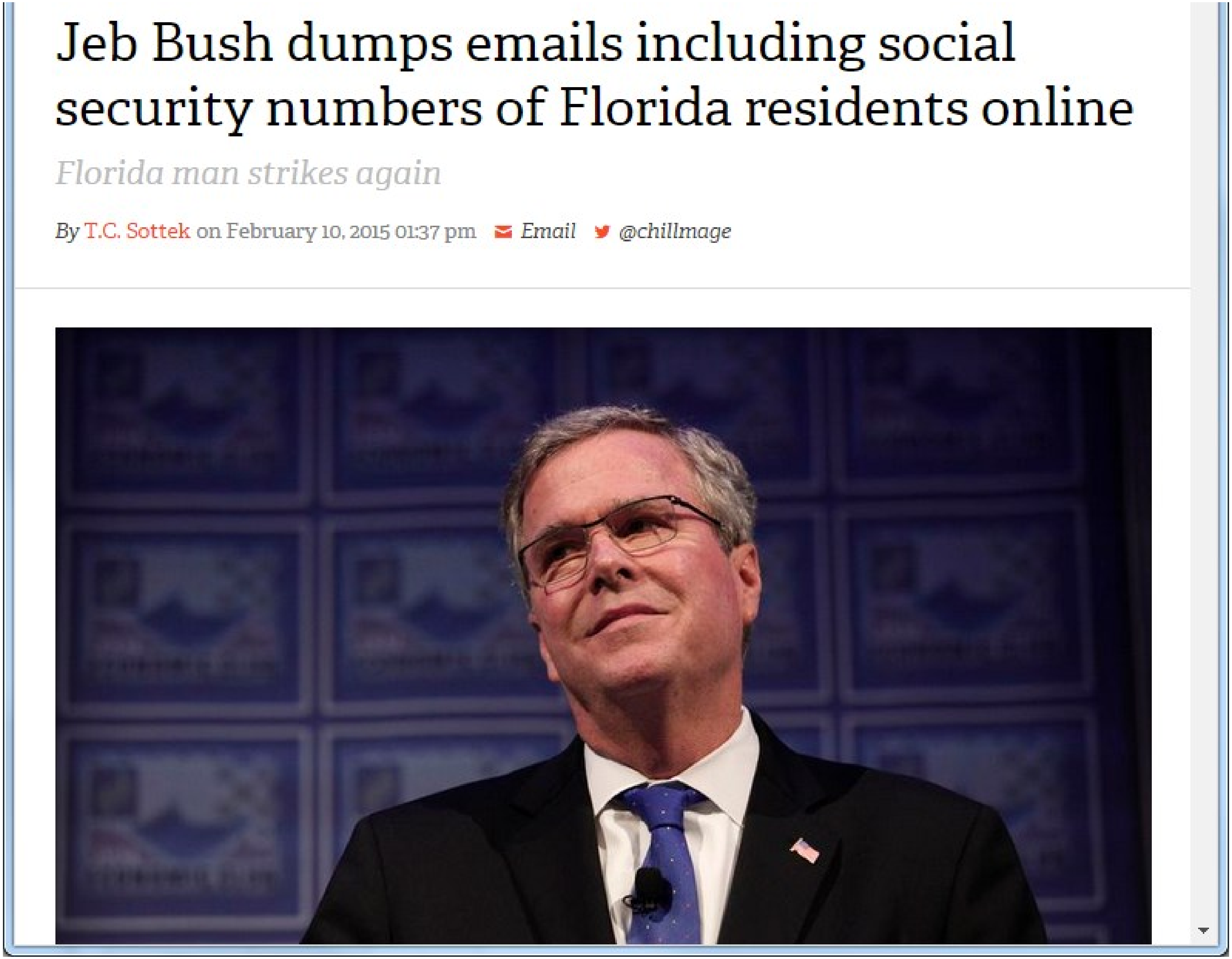
- □ base16
- □ facebook
- □ hashdb
- □ outlook
- □ sceadan
- □ wordlist
- □ xor
- ☑ accts
- ☑ aes
- ☑ base64
- ☑ elf
- ☑ email
- ☑ exif
- ☑ find
- ☑ gps
- ☑ gzip
- ☑ hiberfile
- ☑ httplogs
- ☑ json
- ☑ kml
- ☑ msxml
- ☑ net
- ☑ pdf
- ☑ rar
- ☑ sqlite
- ☑ vcard

Manage Queue… | Import… | Submit Run | Cancel

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

*bulk_extractor, developed by Simson Garfinkel. Available to run directly or within the BitCurator environment
http://wiki.bitcurator.net

That news story about information leaked through email…



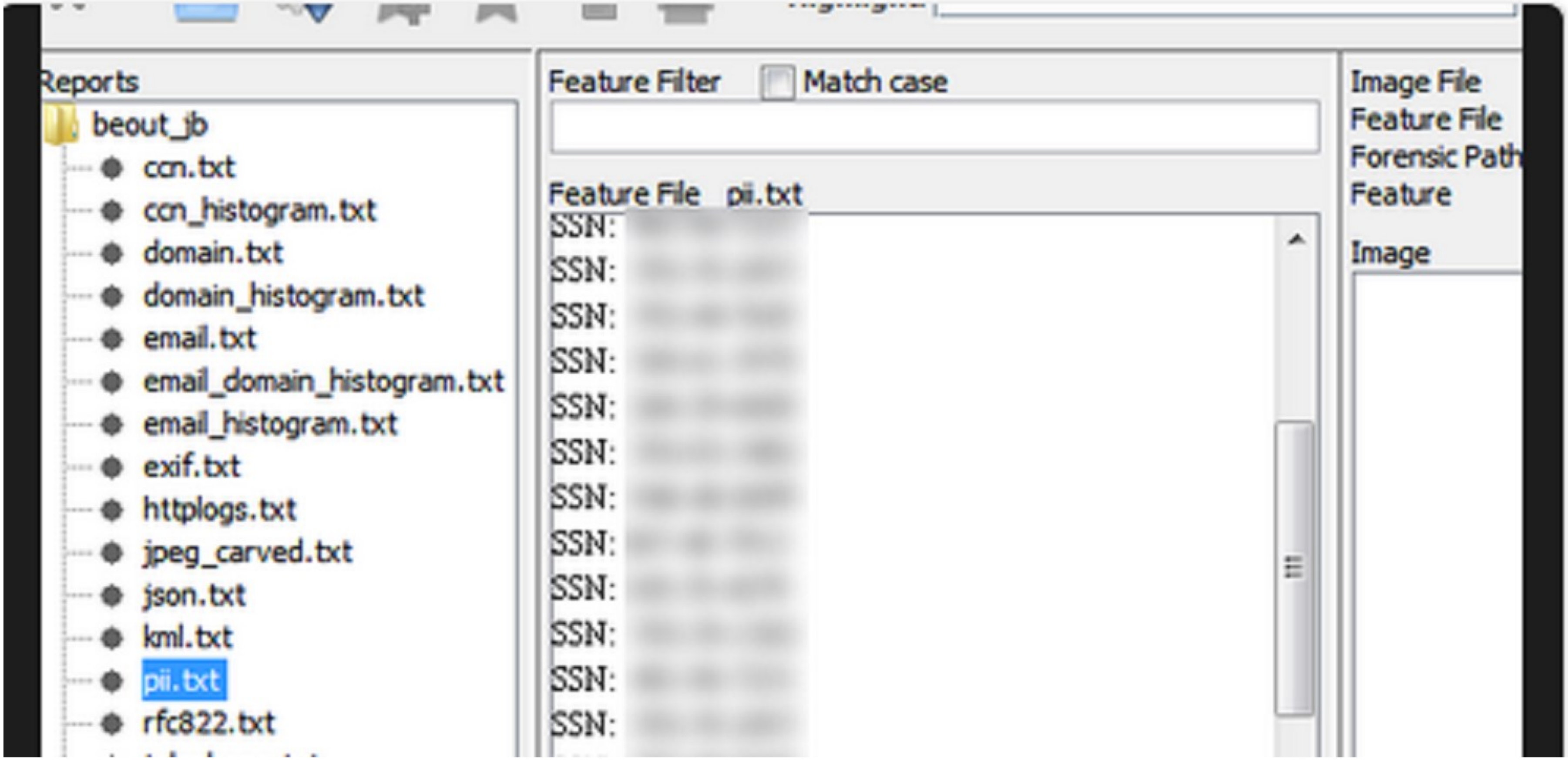
Jeb Bush dumps emails including social security numbers of Florida residents online

Florida man strikes again

By T.C. Sottek on February 10, 2015 01:37 pm *Email* *@chillmage*

https://www.theverge.com/2015/2/10/8013531/jeb-bush-florida-email-dump-privacy

A Real World Example of
Forensic Feature Extraction:



Reports

- beout_jb
  - ccn.txt
  - ccn_histogram.txt
  - domain.txt
  - domain_histogram.txt
  - email.txt
  - email_domain_histogram.txt
  - email_histogram.txt
  - exif.txt
  - httplogs.txt
  - jpeg_carved.txt
  - json.txt
  - kml.txt
  - pii.txt
  - rfc822.txt

Feature Filter ☐ Match case

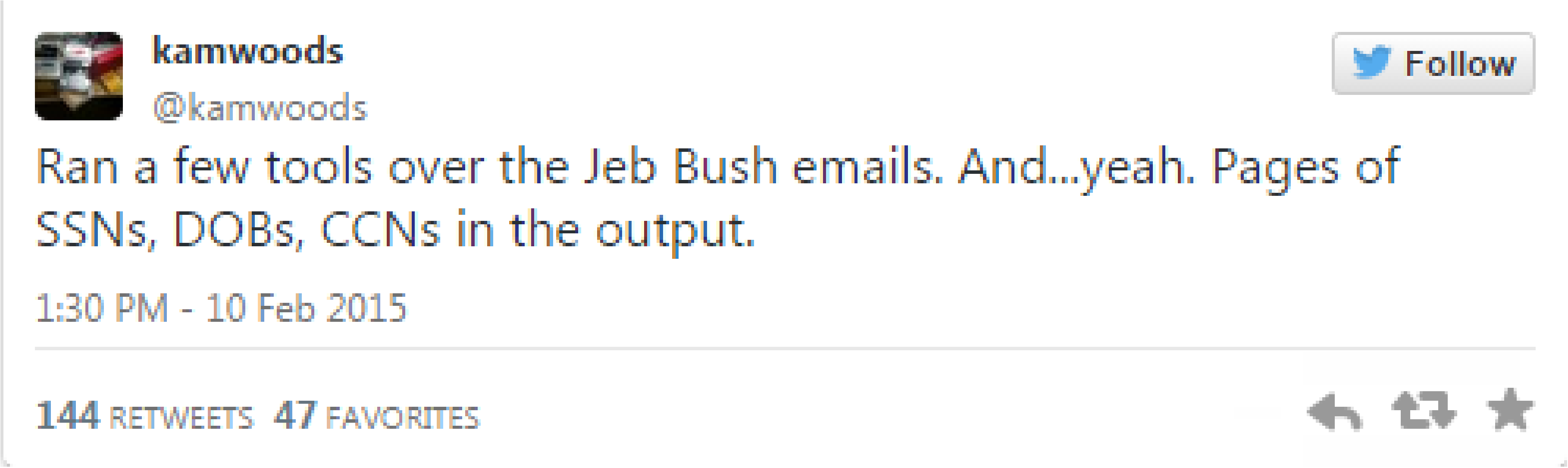
Feature File pii.txt

SSN: [redacted]
SSN: [redacted]
SSN: [redacted]
SSN: [redacted]
SSN: [redacted]
SSN: [redacted]
SSN: [redacted]
SSN: [redacted]
SSN: [redacted]
SSN: [redacted]
SSN: [redacted]
SSN: [redacted]

Image File
Feature File
Forensic Path
Feature

Image

A Real World Example of Forensic Feature Extraction:



kamwoods

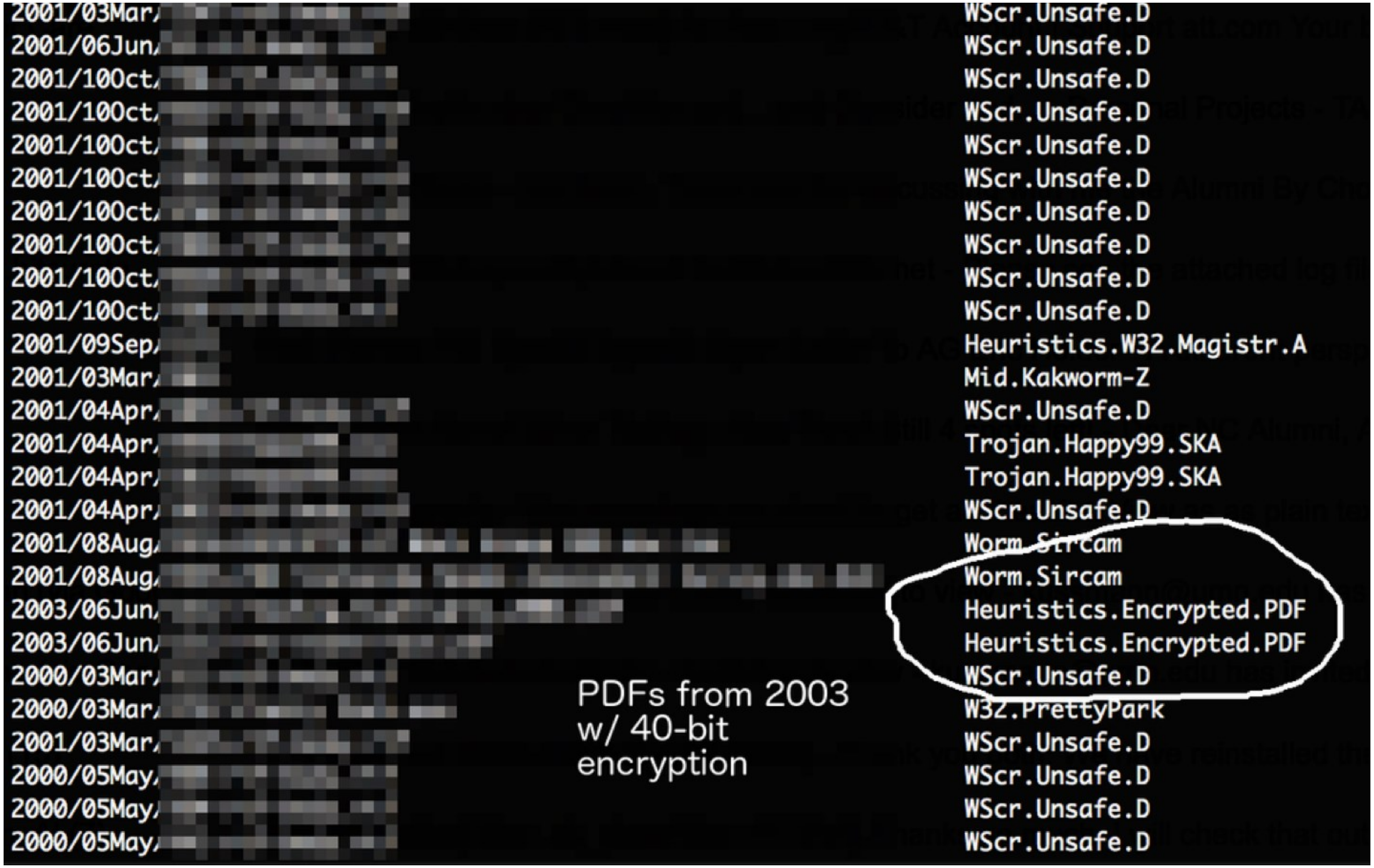
@kamwoods

Ran a few tools over the Jeb Bush emails. And...yeah. Pages of SSNs, DOBs, CCNs in the output.

1:30 PM - 10 Feb 2015

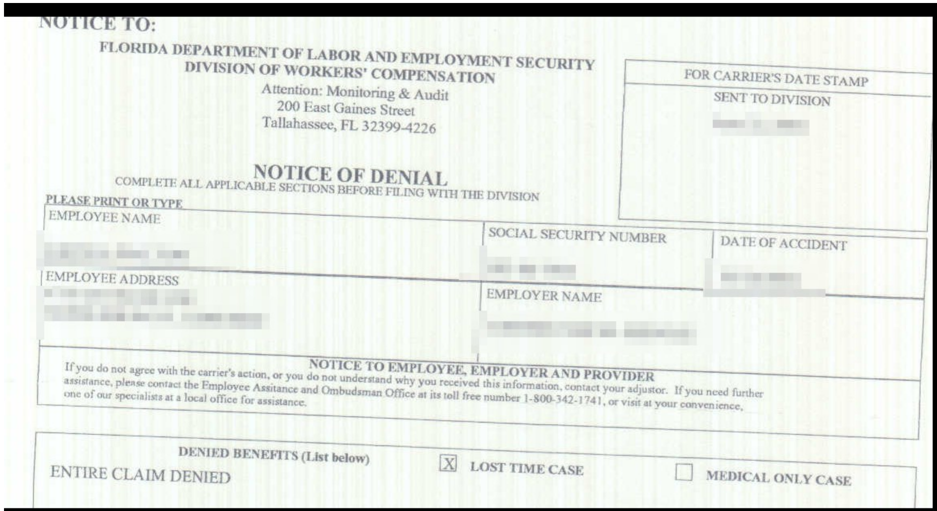
144 RETWEETS 47 FAVORITES

Encryption may be a marker for sensitivity



```
2001/03Mar/[redacted]    WScr.Unsafe.D
2001/06Jun/[redacted]    WScr.Unsafe.D
2001/10Oct/[redacted]    WScr.Unsafe.D
2001/10Oct/[redacted]    WScr.Unsafe.D
2001/10Oct/[redacted]    WScr.Unsafe.D
2001/10Oct/[redacted]    WScr.Unsafe.D
2001/10Oct/[redacted]    WScr.Unsafe.D
2001/10Oct/[redacted]    WScr.Unsafe.D
2001/10Oct/[redacted]    WScr.Unsafe.D
2001/10Oct/[redacted]    WScr.Unsafe.D
2001/09Sep/[redacted]    Heuristics.W32.Magistr.A
2001/03Mar/[redacted]    Mid.Kakworm-Z
2001/04Apr/[redacted]    WScr.Unsafe.D
2001/04Apr/[redacted]    Trojan.Happy99.SKA
2001/04Apr/[redacted]    Trojan.Happy99.SKA
2001/04Apr/[redacted]    WScr.Unsafe.D
2001/08Aug/[redacted]    Worm.Sircam
2001/08Aug/[redacted]    Worm.Sircam
2003/06Jun/[redacted]    Heuristics.Encrypted.PDF
2003/06Jun/[redacted]    Heuristics.Encrypted.PDF
2000/03Mar/[redacted]    WScr.Unsafe.D
2000/03Mar/[redacted]    W32.PrettyPark
2001/03Mar/[redacted]    WScr.Unsafe.D
2000/05May/[redacted]    WScr.Unsafe.D
2000/05May/[redacted]    WScr.Unsafe.D
2000/05May/[redacted]    WScr.Unsafe.D
```

PDFs from 2003
w/ 40-bit
encryption



NOTICE TO:

FLORIDA DEPARTMENT OF LABOR AND EMPLOYMENT SECURITY
DIVISION OF WORKERS' COMPENSATION
Attention: Monitoring & Audit
200 East Gaines Street
Tallahassee, FL 32399-4226

FOR CARRIER'S DATE STAMP
SENT TO DIVISION
[redacted]

NOTICE OF DENIAL

COMPLETE ALL APPLICABLE SECTIONS BEFORE FILING WITH THE DIVISION

PLEASE PRINT OR TYPE

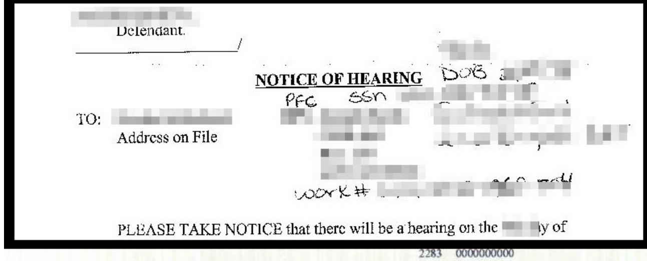
EMPLOYEE NAME [redacted]	SOCIAL SECURITY NUMBER [redacted]	DATE OF ACCIDENT [redacted]
EMPLOYEE ADDRESS [redacted]	EMPLOYER NAME [redacted]	

NOTICE TO EMPLOYEE, EMPLOYER AND PROVIDER

If you do not agree with the carrier's action, or you do not understand why you received this information, contact your adjustor. If you need further assistance, please contact the Employee Assistance and Ombudsman Office at its toll free number 1-800-342-1741, or visit at your convenience, one of our specialists at a local office for assistance.

DENIED BENEFITS (List below) ☒ LOST TIME CASE □ MEDICAL ONLY CASE

ENTIRE CLAIM DENIED



[redacted]
Defendant.

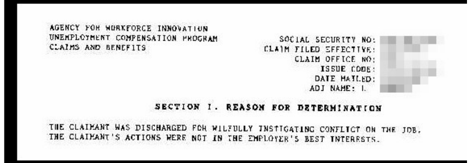
NOTICE OF HEARING DOB [redacted]
PFC SSN [redacted]

TO: [redacted]
Address on File

[redacted]

work # [redacted]

PLEASE TAKE NOTICE that there will be a hearing on the [redacted]y of

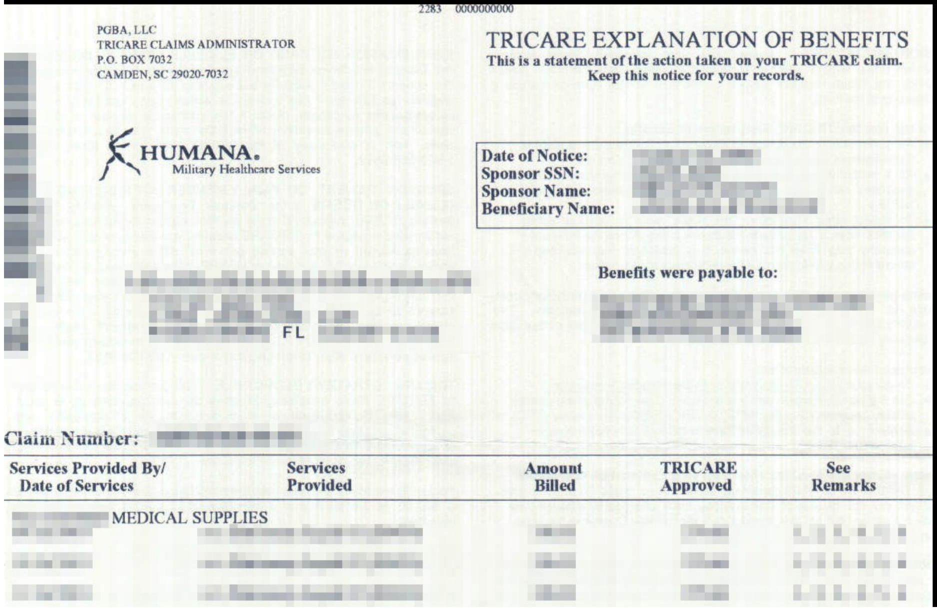


AGENCY FOR WORKFORCE INNOVATION
UNEMPLOYMENT COMPENSATION PROGRAM
CLAIMS AND BENEFITS

SOCIAL SECURITY NO: [redacted]
CLAIM FILED EFFECTIVE:
CLAIM OFFICE NO:
ISSUE CODE:
DATE MAILED:
ADJ NAME: I.

SECTION I. REASON FOR DETERMINATION

THE CLAIMANT WAS DISCHARGED FOR WILFULLY INSTIGATING CONFLICT ON THE JOB.
THE CLAIMANT'S ACTIONS WERE NOT IN THE EMPLOYER'S BEST INTERESTS.



2283 0000000000

PGBA, LLC
TRICARE CLAIMS ADMINISTRATOR
P.O. BOX 7032
CAMDEN, SC 29020-7032

TRICARE EXPLANATION OF BENEFITS

This is a statement of the action taken on your TRICARE claim.
Keep this notice for your records.

HUMANA.
Military Healthcare Services

[redacted]

Date of Notice: [redacted]
Sponsor SSN:
Sponsor Name:
Beneficiary Name:

[redacted]
FL [redacted]

Benefits were payable to:
[redacted]

Claim Number: [redacted]

Services Provided By/ Date of Services	Services Provided	Amount Billed	TRICARE Approved	See Remarks
[redacted] MEDICAL SUPPLIES				
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Other PII and sensitive data may be harder to find, requiring e.g.:

- OCR
- named entity recognition
- partial file reconstruction
- format-specific tools
- visual inspection

Example of EXIF Metadata from a JPEG File (Generated Using exiftool*)

```
---- ExifTool ----
ExifTool Version Number       : 9.38
---- System ----
File Name                     : IMG_20130823_151811.jpg
Directory                     : C:/Users/callee/Documents/images/digital-
forensics-lab
File Size                     : 1785 kB
File Modification Date/Time   : 2013:08:23 16:36:44-04:00
File Access Date/Time         : 2013:10:14 17:13:02-04:00
File Creation Date/Time       : 2013:08:23 16:36:44-04:00
File Permissions              : rw-rw-rw-
---- File ----
File Type                     : JPEG
MIME Type                     : image/jpeg
Exif Byte Order               : Big-endian (Motorola, MM)
Image Width                   : 2592
Image Height                  : 1944
Encoding Process              : Baseline DCT, Huffman coding
Bits Per Sample               : 8
Color Components              : 3
Y Cb Cr Sub Sampling          : YCbCr4:2:0 (2 2)
---- GPS ----
GPS Img Direction             : 83
GPS Img Direction Ref         : Magnetic North
GPS Latitude Ref              : North
GPS Latitude                  : 35 deg 55' 2.24"
GPS Longitude Ref             : West
GPS Longitude                 : 79 deg 2' 57.55"
GPS Altitude Ref              : Above Sea Level
GPS Altitude                  : 0 m
GPS Time Stamp                : 19:18:06
GPS Processing Method         : NETWORK
GPS Date Stamp                : 2013:08:23
---- IFD0 ----
Orientation                   : Unknown (0)
Camera Model Name             : Galaxy Nexus
Modify Date                   : 2013:08:23 15:18:11
Y Cb Cr Positioning           : Centered
Y Resolution                  : 72
Resolution Unit               : inches
X Resolution                  : 72
Make                          : Samsung
---- ExifIFD ----
Create Date                   : 2013:08:23 15:18:11
Date/Time Original            : 2013:08:23 15:18:11
Exif Version                  : 0220
Flash Energy                  : 0
Image Unique ID               : OAEL01
Exposure Time                 : 1/17
ISO                           : 125, 0, 0
```

Geolocation data embedded in EXIF metadata from a smartphone photo

*http://www.sno.phy.queensu.ca/~phil/exiftool/ (Also available through the BitCurator environment)

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. Three 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) and emulated access to redacted images**

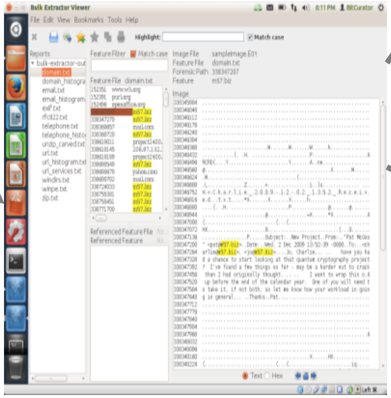
Redacting Born-Digital Materials

- Locating relevant items can be problematic
  - compressed files
  - proprietary formats
  - formatting variations
  - Encryption
  - …
- Digital forensics tools can help
  - Scan file systems and disk images block-by-block
  - Extract features when possible
  - Report on materials that resist analysis

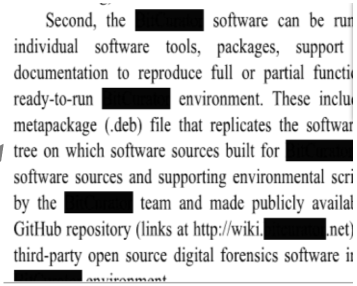
Redaction: Automating search for redaction candidates and redaction actions is desirable, and few tools exist to support this search across heterogeneous materials

Acquire disk image from original media

Identify items to redact

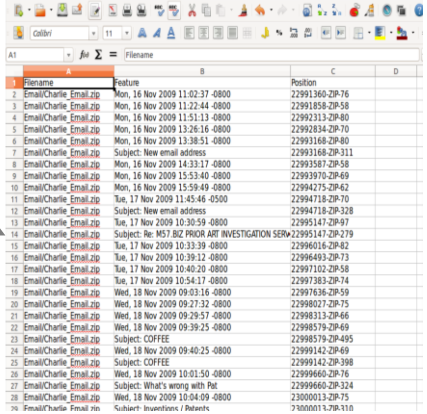


Generate redacted disk image and/or files



Second, the [redacted] software can be run
individual software tools, packages, support
documentation to reproduce full or partial functi
ready-to-run [redacted] environment. These inclu
metapackage (.deb) file that replicates the softwar
tree on which software sources built for [redacted]
software sources and supporting environmental scri
by the [redacted] team and made publicly availal
GitHub repository (links at http://wiki.[redacted].net)
third-party open source digital forensics software i
[redacted] environment.

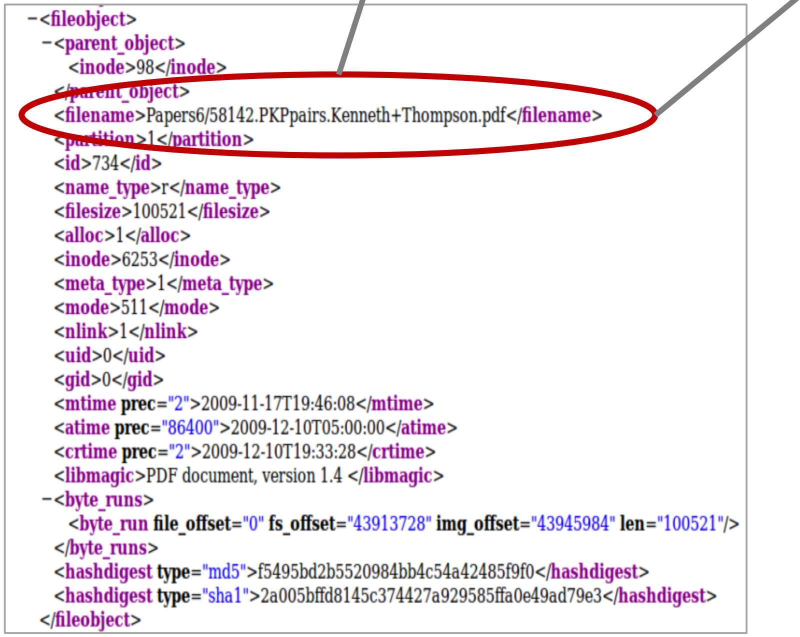
Report on redacted items for preservation record



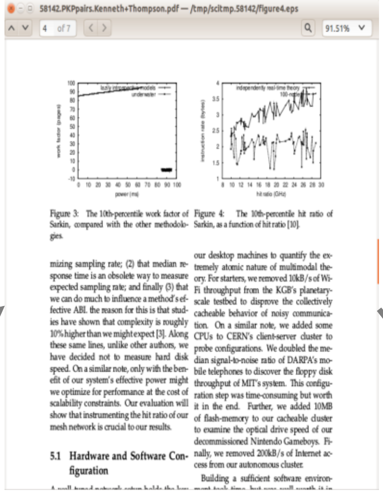
	A	B	C	D
1	Filename	Feature	Position	
2	Email/Charlie_Email.zip	Mon, 16 Nov 2009 11:02:37 -0800	22991360-ZIP-76	
3	Email/Charlie_Email.zip	Mon, 16 Nov 2009 11:22:44 -0800	22991858-ZIP-58	
4	Email/Charlie_Email.zip	Mon, 16 Nov 2009 11:51:13 -0800	22992313-ZIP-80	
5	Email/Charlie_Email.zip	Mon, 16 Nov 2009 13:26:16 -0800	22992834-ZIP-70	
6	Email/Charlie_Email.zip	Mon, 16 Nov 2009 13:38:51 -0800	22993168-ZIP-80	
7	Email/Charlie_Email.zip	Subject: New email address	22993168-ZIP-311	
8	Email/Charlie_Email.zip	Mon, 16 Nov 2009 14:33:17 -0800	22993587-ZIP-58	
9	Email/Charlie_Email.zip	Mon, 16 Nov 2009 15:53:40 -0800	22993970-ZIP-69	
10	Email/Charlie_Email.zip	Mon, 16 Nov 2009 15:59:49 -0800	22994275-ZIP-62	
11	Email/Charlie_Email.zip	Tue, 17 Nov 2009 11:45:46 -0500	22994718-ZIP-70	
12	Email/Charlie_Email.zip	Subject: New email address	22994718-ZIP-328	
13	Email/Charlie_Email.zip	Tue, 17 Nov 2009 10:30:59 -0800	22995147-ZIP-97	
14	Email/Charlie_Email.zip	Subject: Re: M57.BIZ PRIOR ART INVESTIGATION SERV	22995147-ZIP-279	
15	Email/Charlie_Email.zip	Tue, 17 Nov 2009 10:33:39 -0800	22996016-ZIP-82	
16	Email/Charlie_Email.zip	Tue, 17 Nov 2009 10:39:12 -0800	22996493-ZIP-73	
17	Email/Charlie_Email.zip	Tue, 17 Nov 2009 10:40:20 -0800	22997102-ZIP-58	
18	Email/Charlie_Email.zip	Tue, 17 Nov 2009 10:54:17 -0800	22997383-ZIP-74	
19	Email/Charlie_Email.zip	Wed, 18 Nov 2009 09:03:16 -0800	22997636-ZIP-59	
20	Email/Charlie_Email.zip	Wed, 18 Nov 2009 09:27:32 -0800	22998027-ZIP-75	
21	Email/Charlie_Email.zip	Wed, 18 Nov 2009 09:29:57 -0800	22998313-ZIP-66	
22	Email/Charlie_Email.zip	Wed, 18 Nov 2009 09:39:25 -0800	22998579-ZIP-69	
23	Email/Charlie_Email.zip	Subject: COFFEE	22998579-ZIP-495	
24	Email/Charlie_Email.zip	Wed, 18 Nov 2009 09:40:25 -0800	22999142-ZIP-69	
25	Email/Charlie_Email.zip	Subject: COFFEE	22999142-ZIP-398	
26	Email/Charlie_Email.zip	Wed, 18 Nov 2009 10:01:50 -0800	22999660-ZIP-76	
27	Email/Charlie_Email.zip	Subject: What's wrong with Pat	22999660-ZIP-324	
28	Email/Charlie_Email.zip	Wed, 18 Nov 2009 10:04:09 -0800	23000013-ZIP-75	
29	Email/Charlie_Email.zip	Subject: Inventions / Patents	23000013-ZIP-310	

Forensic disk imaging and metadata extraction provides clear provenance for redacted copies

File object identified in disk image and recorded in a forensic metadata format (DFXML)



```
-<fileobject>
  -<parent_object>
     <inode>98</inode>
   </parent_object>
   <filename>Papers6/58142.PKPpairs.Kenneth+Thompson.pdf</filename>
   <partition>1</partition>
   <id>734</id>
   <name_type>r</name_type>
   <filesize>100521</filesize>
   <alloc>1</alloc>
   <inode>6253</inode>
   <meta_type>1</meta_type>
   <mode>511</mode>
   <nlink>1</nlink>
   <uid>0</uid>
   <gid>0</gid>
   <mtime prec="2">2009-11-17T19:46:08</mtime>
   <atime prec="86400">2009-12-10T05:00:00</atime>
   <crtime prec="2">2009-12-10T19:33:28</crtime>
   <libmagic>PDF document, version 1.4 </libmagic>
  -<byte_runs>
     <byte_run file_offset="0" fs_offset="43913728" img_offset="43945984" len="100521"/>
   </byte_runs>
   <hashdigest type="md5">f5495bd2b5520984bb4c54a42485f9f0</hashdigest>
   <hashdigest type="sha1">2a005bffd8145c374427a929585ffa0e49ad79e3</hashdigest>
 </fileobject>
```



58142.PKPpairs.Kenneth+Thompson.pdf — /tmp/scitmp.58142/figure4.eps

Figure 3: The 10th-percentile work factor of Sarkin, compared with the other methodologies.

Figure 4: The 10th-percentile hit ratio of Sarkin, as a function of hit ratio [10].

mizing sampling rate; (2) that median response time is an obsolete way to measure expected sampling rate; and finally (3) that we can do much to influence a method's effective ABI. the reason for this is that studies have shown that complexity is roughly 10% higher than we might expect [3]. Along these same lines, unlike other authors, we have decided not to measure hard disk speed. On a similar note, only with the benefit of our system's effective power might we optimize for performance at the cost of scalability constraints. Our evaluation will show that instrumenting the hit ratio of our mesh network is crucial to our results.

5.1 Hardware and Software Configuration

our desktop machines to quantify the extremely atomic nature of multimodal theory. For starters, we removed 10kB/s of Wi-Fi throughput from the KGB's planetary-scale testbed to disprove the collectively cacheable behavior of noisy communication. On a similar note, we added some CPUs to CERN's client-server cluster to probe configurations. We doubled the median signal-to-noise ratio of DARPA's mobile telephones to discover the floppy disk throughput of MIT's system. This configuration step was time-consuming but worth it in the end. Further, we added 10MB of flash-memory to our cacheable cluster to examine the optical drive speed of our decommissioned Nintendo Gameboys. Finally, we removed 200kB/s of Internet access from our autonomous cluster.

Building a sufficient software environ-

Original file (unredacted in disk image)

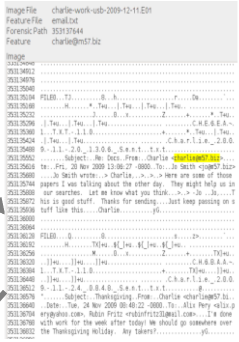
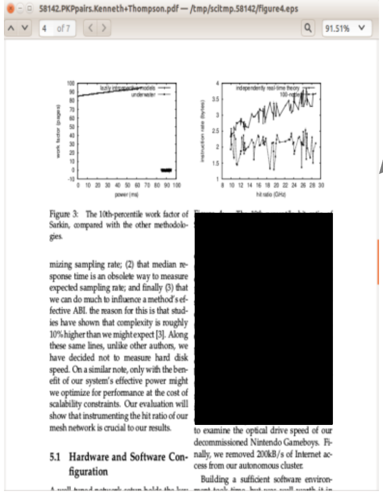


Image File charlie-work-usb-2009-12-11.E01
Feature File email.txt
Forensic Path 353137644
Feature charlie@m57.biz

Image

```
353134912 ..............................................................
353134976 ..............................................................
353135040 ..............................................................
353135104 FILE0...TJ..........8..h...................r.....Dm.........`...
353135168 ........H.......*..T+u...|.T+u...|.T+u...|.T+u.. ...............
353135232 .................J......0...x...........Z.......+.......*..T+u..
353135296 .|.T+u...|.T+u...|.T+u................. .........C.H.E.6.E.A.~.
353135360 1...T.X.T.-.1.1.0.....................+.......*..T+u...|.T+u..
353135424 .|.T+u...|.T+u................ ....... .C.h.a.r.l.i.e._.2.0.0.
353135488 9.-.1.1.-.2.0._.1.3.0.6._.S.e.n.t...t.x.t......................
353135552 ........Subject:..Re: Docs..From:..Charlie <charlie@m57.biz>....
353135616 te:..Fri, 20 Nov 2009 13:06:27 -0800..To:..Jo Smith <jo@m57.biz>
353135680 .....Jo Smith wrote:..> Charlie,..>..>..> Here are some of those
353135744 papers I was talking about the other day.  They might help us in
353135808  our searches.  Let me know what you think...>..> -Jo ..Jo,....T
353135872 his is good stuff.  Thanks for sending....Just keep passing on s
353135936 tuff like this.....Charlie...........yG.........................
353136000 ..............................................................
353136064 ..............................................................
353136128 FILE0....Q..........8......................s.....z>........`...
353136192 ........H........TX]+u..$[_]+u..$[_]+u..$[_]+u.. ...............
353136256 ...................M......0...x...........Z.......+........TX]+u..
353136320 ..]]+u....]]+u....]]+u................. .........C.H.6.B.A.A.~.
353136384 1...T.X.T.-.1.1.0.....................+........TX]+u....]]+u..
353136448 ..]]+u....]]+u................ ....... .C.h.a.r.l.i.e._.2.0.0.
353136512 9.-.1.1.-.2.4._.0.8.4.0._.S.e.n.t...t.x.t...........@..........
353136576 *.......Subject:..Thanksgiving..From:..Charlie <charlie@m57.bi..
353136640 ..Date:..Tue, 24 Nov 2009 08:40:22 -0800..To:..Alix Pery <alix.p
353136704 ery@yahoo.com>, Rubin Fritz <rubinfritz3l@mail.com>....I'm done
353136768 with work for the week after today! We should go somewhere over
353136832 the Thanksgiving Holiday.  Any takers?...............yG.........
```

PII identified at byte offsets



58142.PKPpairs.Kenneth+Thompson.pdf — /tmp/scitmp.58142/figure4.eps

Figure 3: The 10th-percentile work factor of Sarkin, compared with the other methodologies.

mizing sampling rate; (2) that median response time is an obsolete way to measure expected sampling rate; and finally (3) that we can do much to influence a method's effective ABI. the reason for this is that studies have shown that complexity is roughly 10% higher than we might expect [3]. Along these same lines, unlike other authors, we have decided not to measure hard disk speed. On a similar note, only with the benefit of our system's effective power might we optimize for performance at the cost of scalability constraints. Our evaluation will show that instrumenting the hit ratio of our mesh network is crucial to our results.

5.1 Hardware and Software Configuration

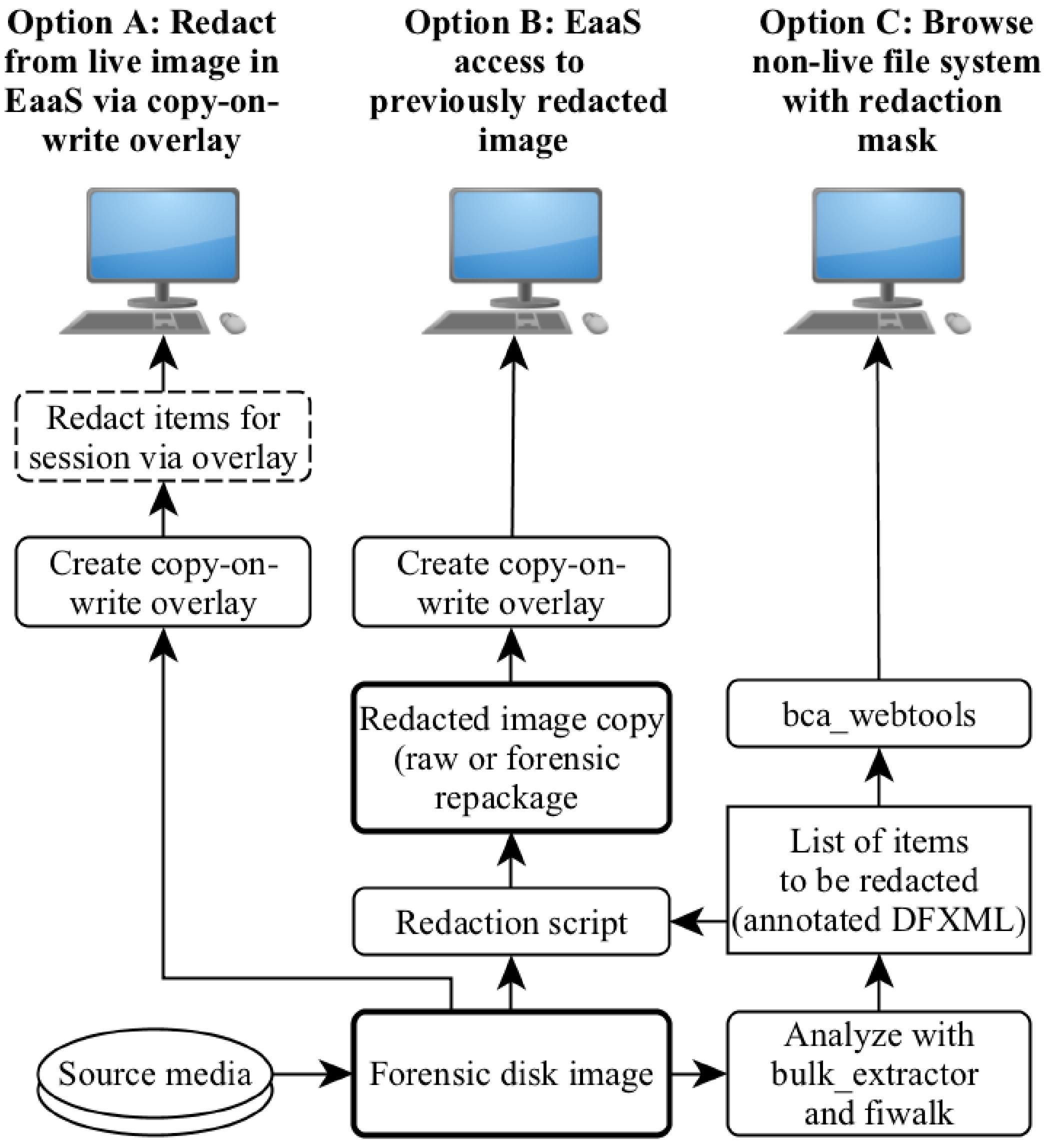
[redacted]

to examine the optical drive speed of our decommissioned Nintendo Gameboys. Finally, we removed 200kB/s of Internet access from our autonomous cluster.

Building a sufficient software environ-

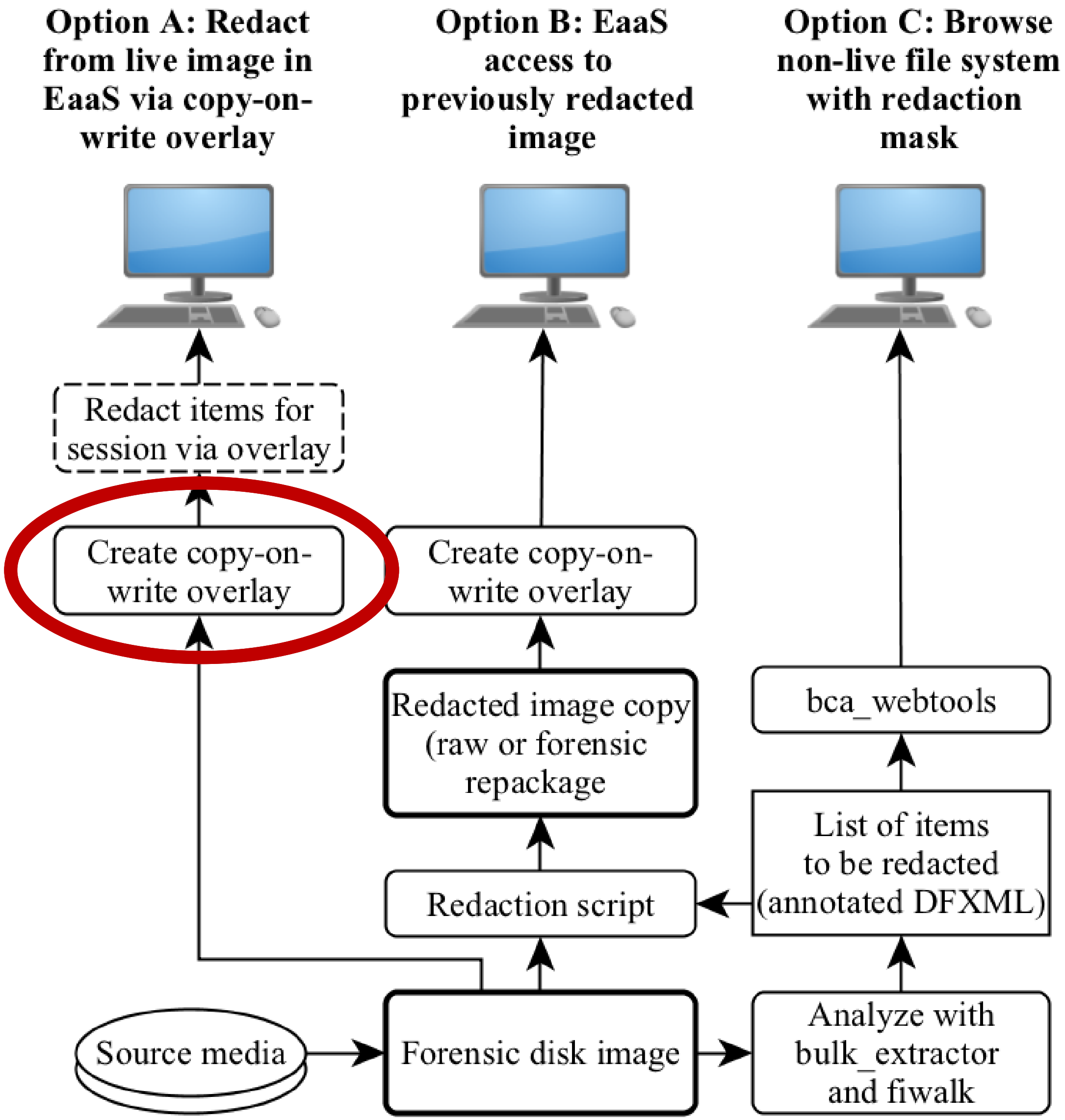
Redacted copy (alternate location)

Automated Redaction and Access Options


Option A: Redact from live image in EaaS via copy-on-write overlay
Option B: EaaS access to previously redacted image
Option C: Browse non-live file system with redaction mask
Redact items for session via overlay
Create copy-on-write overlay
Create copy-on-write overlay
Redacted image copy (raw or forensic repackage
bca_webtools
List of items to be redacted (annotated DFXML)
Redaction script
Source media
Forensic disk image
Analyze with bulk_extractor and fiwalk

EaaS = Emulation-as-a-Service. http://bw-fla.uni-freiburg.de/

Automated Redaction and Access Options



Option A: Redact from live image in EaaS via copy-on-write overlay

Option B: EaaS access to previously redacted image

Option C: Browse non-live file system with redaction mask

Redact items for session via overlay

Create copy-on-write overlay

Create copy-on-write overlay

Redacted image copy (raw or forensic repackage

bca_webtools

List of items to be redacted (annotated DFXML)

Redaction script

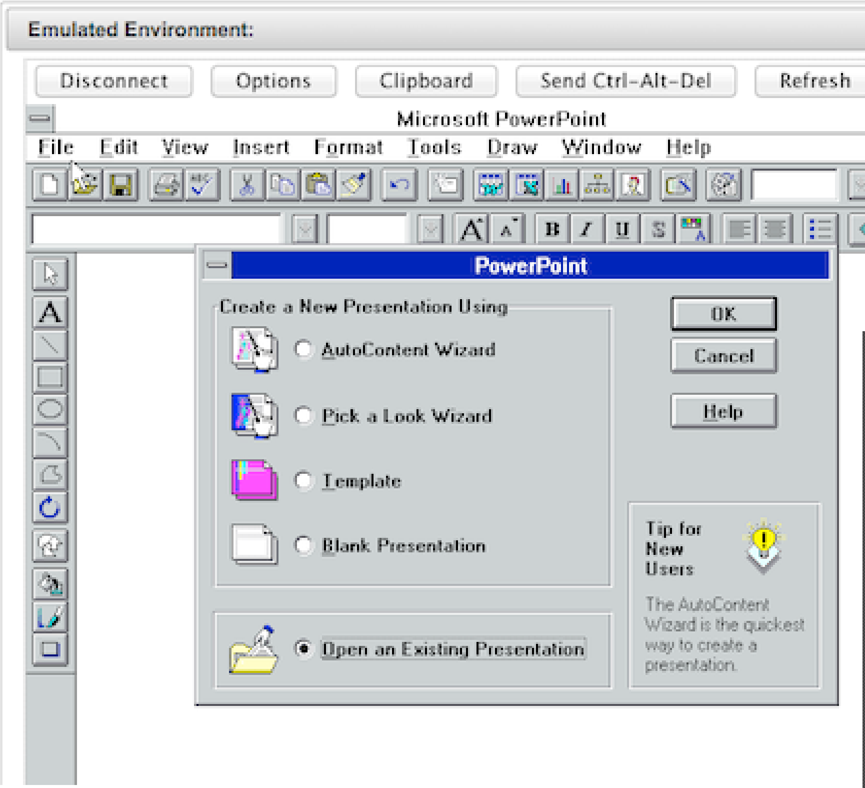
Source media

Forensic disk image

Analyze with bulk_extractor and fiwalk

EaaS = Emulation-as-a-Service. http://bw-fla.uni-freiburg.de/

Emulation as a Service



Create a New Presentation Using

○ AutoContent Wizard

○ Pick a Look Wizard

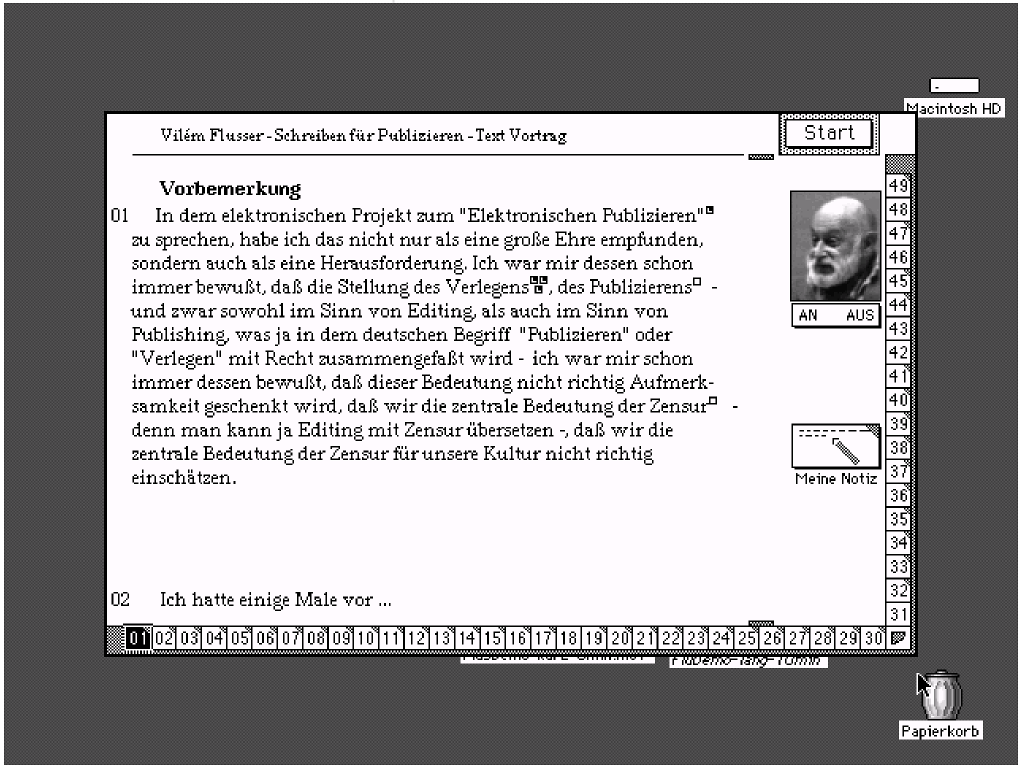
○ Template

○ 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

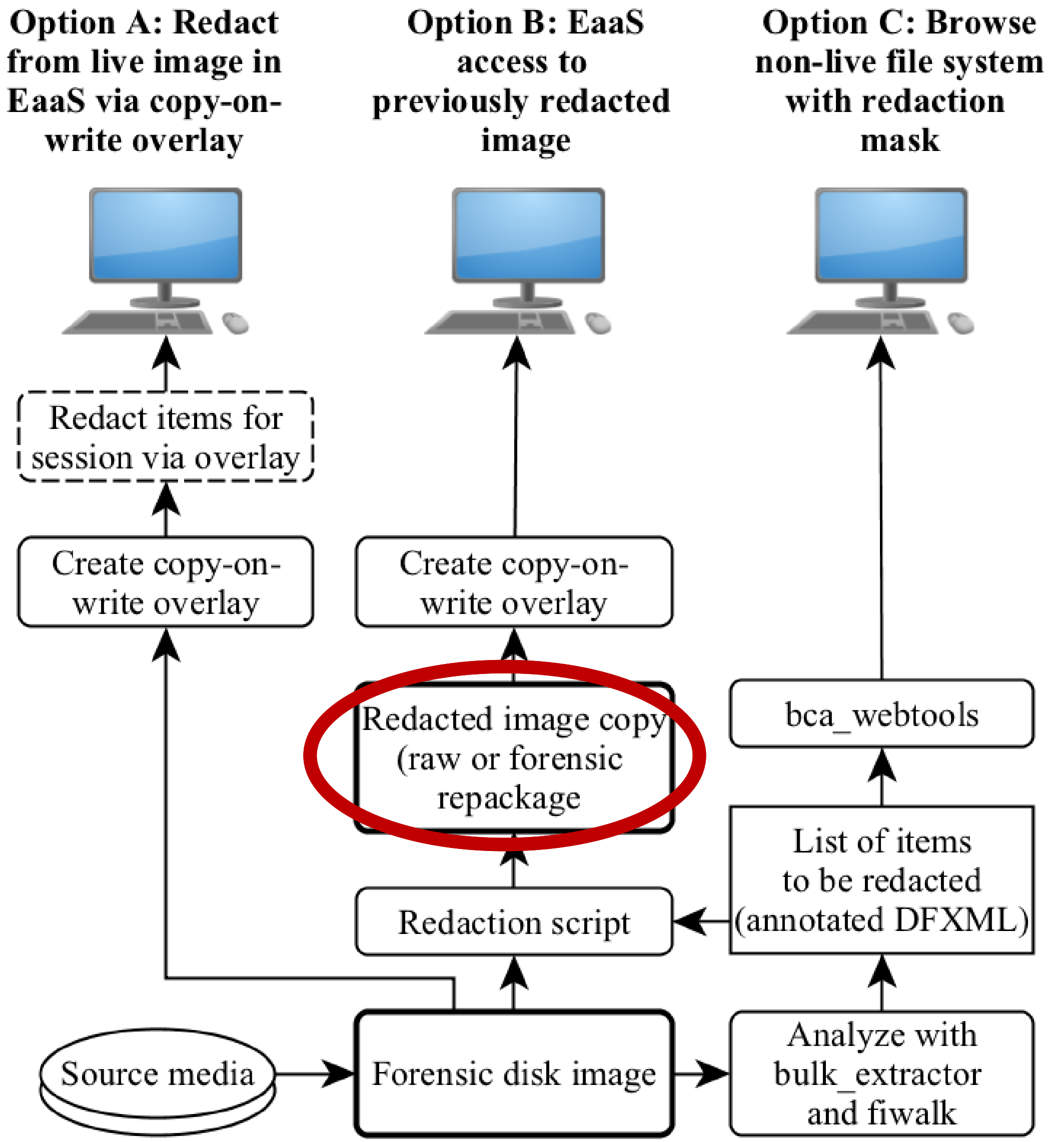
01 In dem elektronischen Projekt zum "Elektronischen Publizieren" zu sprechen, habe ich das nicht nur als eine große Ehre empfunden, sondern auch als eine Herausforderung. Ich war mir dessen schon immer bewußt, daß die Stellung des Verlegens, des Publizierens - und zwar sowohl im Sinn von Editing, als 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 ...

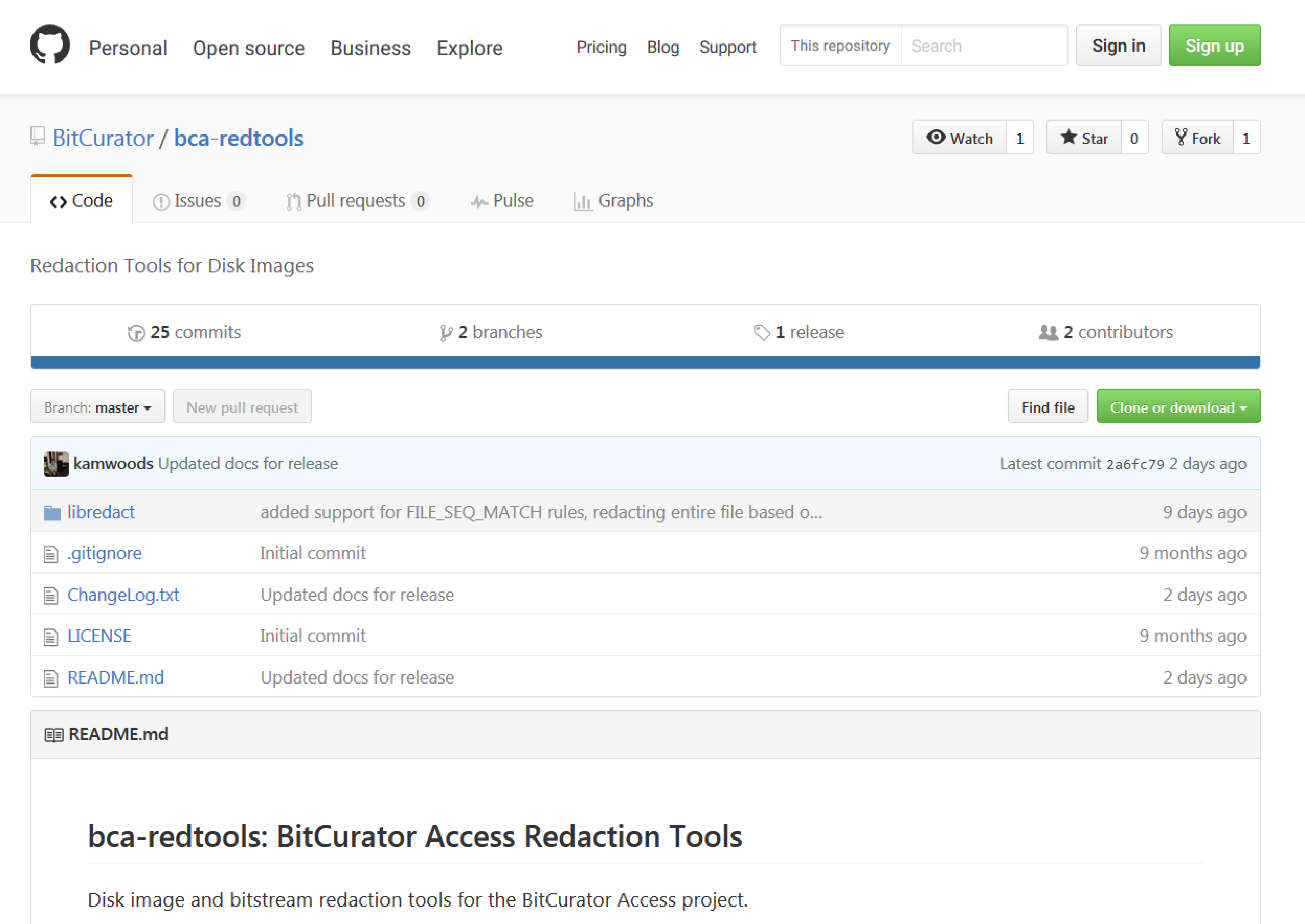
http://bw-fla.uni-freiburg.de/demos.html

See also: Woods, Kam, Christopher Lee, Oleg Stobbe, Thomas Liebetraut and Klaus Rechert. "Functional Access to Forensic Disk Images in a Web Service." In *Proceedings of the 12th International Conference on Digital Curation*, edited by Christopher A. Lee, Jonathan Crabtree, Leo Konstantelos, Nancy McGovern, Yukio Maeda, Maureen Pennock, Helen Tibbo, Kam Woods, and Eld Zierau, 191-195. Chapel Hill, NC: University of North Carolina, School of Information and Library Science, 2015.

Automated Redaction and Access Options


Option A: Redact from live image in EaaS via copy-on-write overlay
Option B: EaaS access to previously redacted image
Option C: Browse non-live file system with redaction mask
Redact items for session via overlay
Create copy-on-write overlay
Create copy-on-write overlay
Redacted image copy (raw or forensic repackage
bca_webtools
List of items to be redacted (annotated DFXML)
Redaction script
Source media
Forensic disk image
Analyze with bulk_extractor and fiwalk

EaaS = Emulation-as-a-Service. http://bw-fla.uni-freiburg.de/



BitCurator / **bca-redtools**

Redaction Tools for Disk Images

25 commits　**2** branches　**1** release　**2** contributors

kamwoods Updated docs for release　Latest commit `2a6fc79` 2 days ago

libredact	added support for FILE_SEQ_MATCH rules, redacting entire file based o...	9 days ago
.gitignore	Initial commit	9 months ago
ChangeLog.txt	Updated docs for release	2 days ago
LICENSE	Initial commit	9 months ago
README.md	Updated docs for release	2 days ago

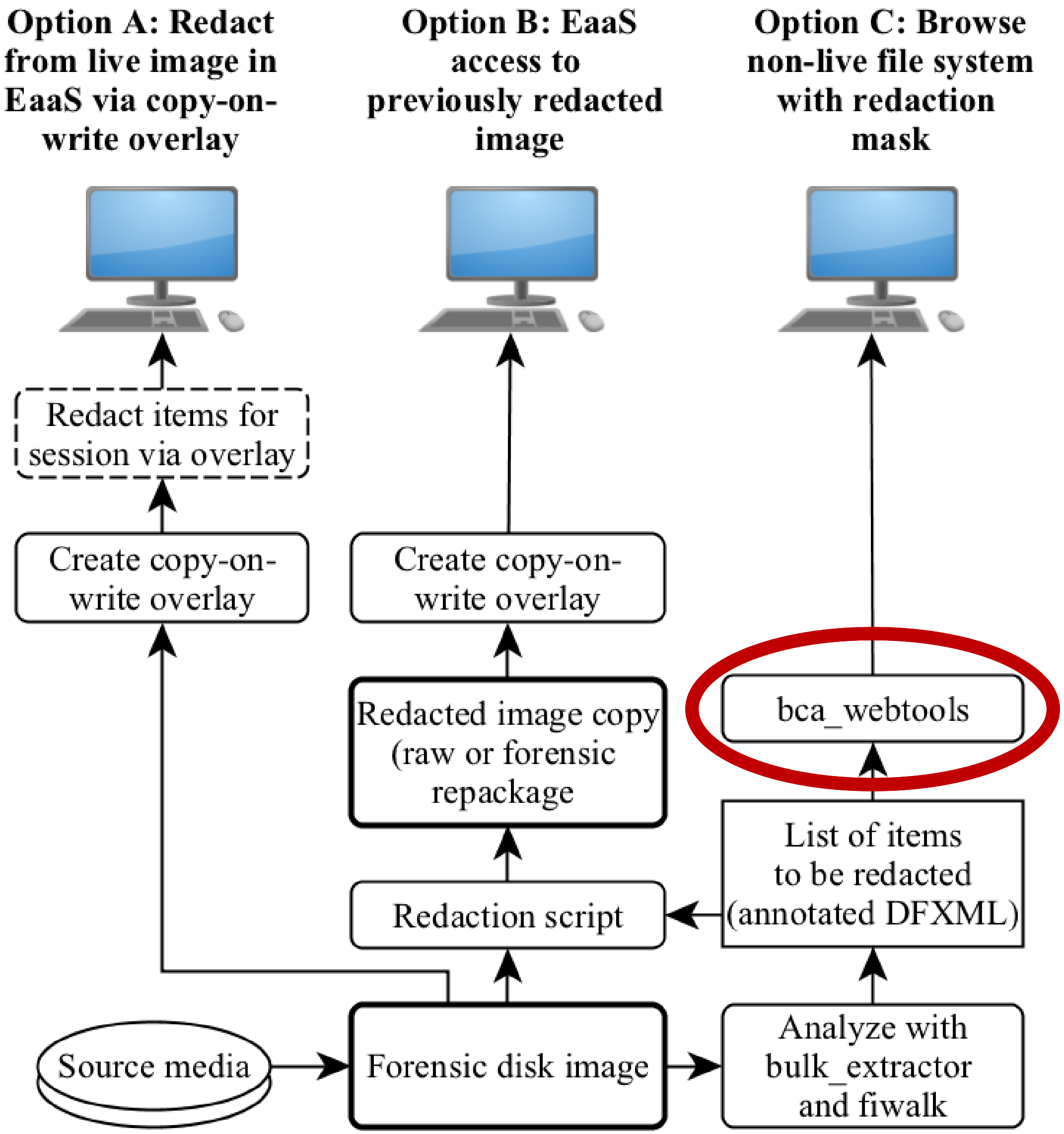
README.md

bca-redtools: BitCurator Access Redaction Tools

Disk image and bitstream redaction tools for the BitCurator Access project.

https://github.com/BitCurator/bca-redtools

Automated Redaction and Access Options


Option A: Redact from live image in EaaS via copy-on-write overlay
Option B: EaaS access to previously redacted image
Option C: Browse non-live file system with redaction mask
Redact items for session via overlay
Create copy-on-write overlay
Create copy-on-write overlay
Redacted image copy (raw or forensic repackage
bca_webtools
List of items to be redacted (annotated DFXML)
Redaction script
Source media
Forensic disk image
Analyze with bulk_extractor and fiwalk

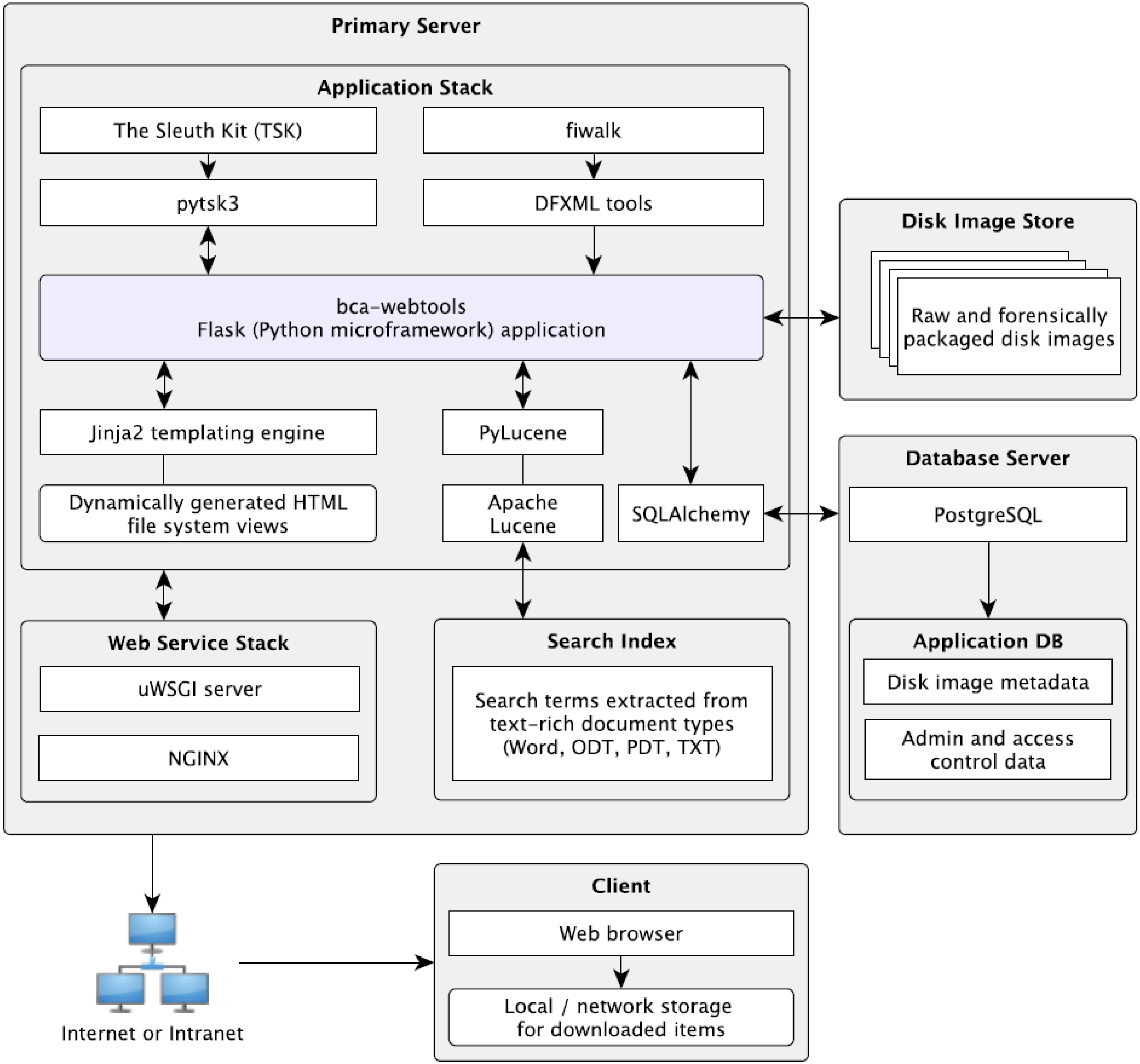
EaaS = Emulation-as-a-Service. http://bw-fla.uni-freiburg.de/

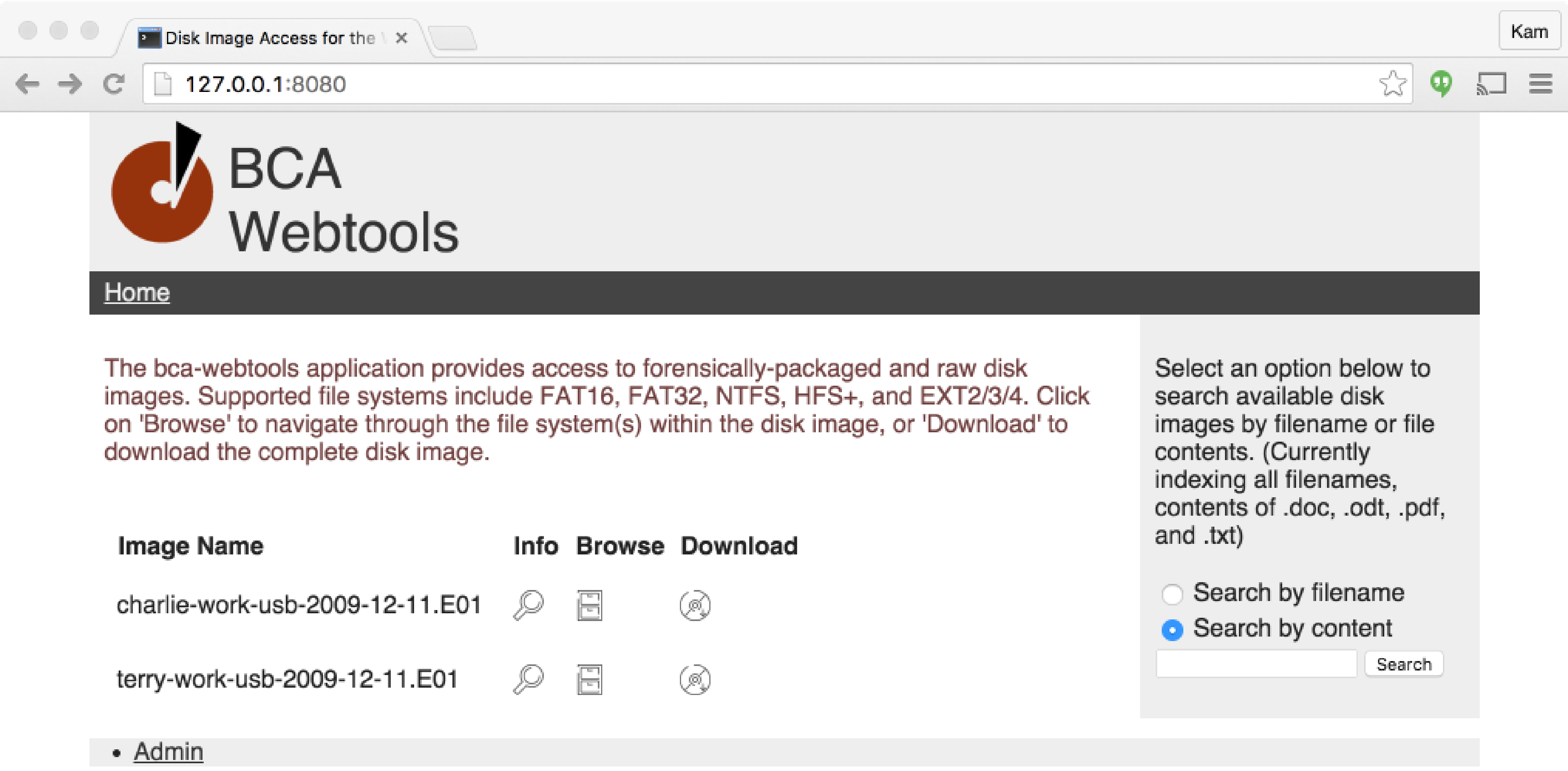
BCA (BitCurator Access) Web Tools

- Integrates digital forensics software libraries and lightweight web-services tools
- Drop disk images in a local or network-accessible location, start up the service, and start browsing
- Most analysis runs server-side (via Sleuthkit and DFXML Python bindings, among others)
- Service is database-agnostic (we're using postgres)
- Automatic metadata production – Digital Forensics XML (DFXML), PREMIS, others)

https://github.com/kamwoods/bca-webtools

Sunitha Misra, Christopher A. Lee, and Kam Woods, "A Web Service for File-Level Access to Disk Images," *Code4Lib Journal 25* (2014), http://journal.code4lib.org/articles/9773


Primary Server
Application Stack
The Sleuth Kit (TSK)
fiwalk
pytsk3
DFXML tools
bca-webtools
Flask (Python microframework) application
Jinja2 templating engine
PyLucene
Dynamically generated HTML file system views
Apache Lucene
SQLAlchemy
Web Service Stack
uWSGI server
NGINX
Search Index
Search terms extracted from text-rich document types (Word, ODT, PDT, TXT)
Disk Image Store
Raw and forensically packaged disk images
Database Server
PostgreSQL
Application DB
Disk image metadata
Admin and access control data
Internet or Intranet
Client
Web browser
Local / network storage for downloaded items



BCA Webtools

Home

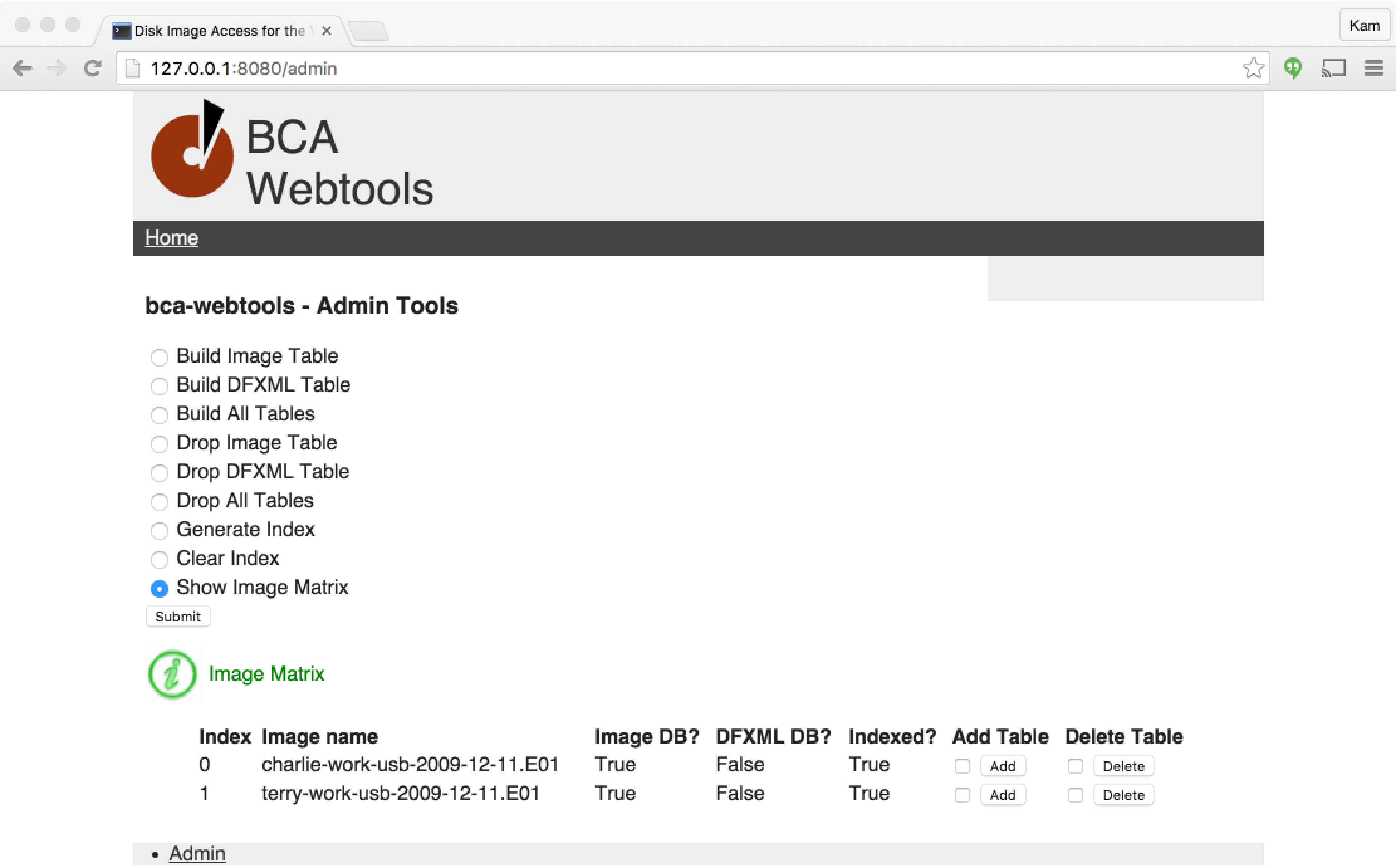
The bca-webtools application provides access to forensically-packaged and raw disk images. Supported file systems include FAT16, FAT32, NTFS, HFS+, and EXT2/3/4. Click on 'Browse' to navigate through the file system(s) within the disk image, or 'Download' to download the complete disk image.

Image Name	Info	Browse	Download
charlie-work-usb-2009-12-11.E01	🔍	🗄	💿
terry-work-usb-2009-12-11.E01	🔍	🗄	💿

Select an option below to search available disk images by filename or file contents. (Currently indexing all filenames, contents of .doc, .odt, .pdf, and .txt)

○ Search by filename
● Search by content

- Admin

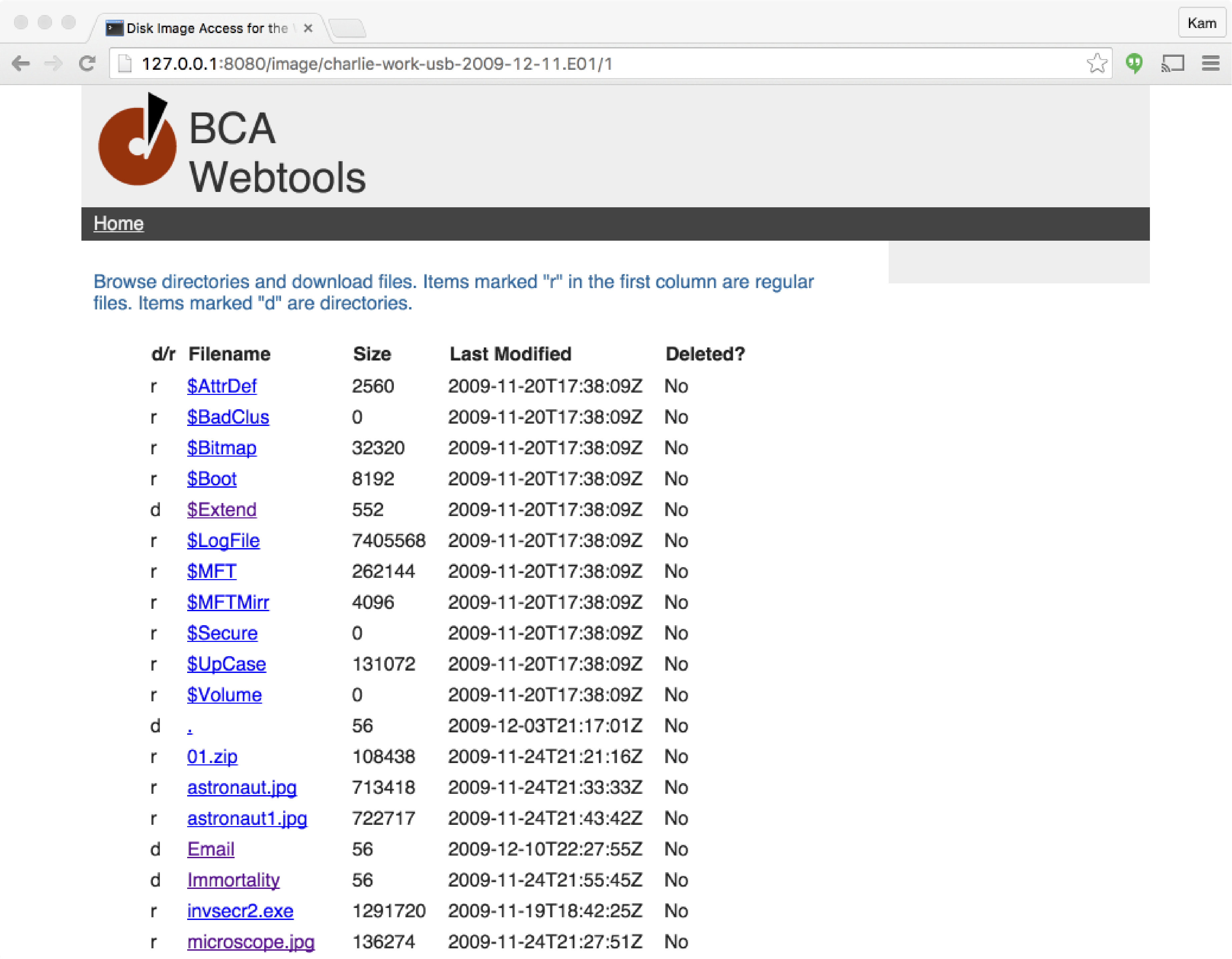


bca-webtools - Admin Tools

○ Build Image Table
○ Build DFXML Table
○ Build All Tables
○ Drop Image Table
○ Drop DFXML Table
○ Drop All Tables
○ Generate Index
○ Clear Index
● Show Image Matrix

Image Matrix

Index	Image name	Image DB?	DFXML DB?	Indexed?	Add Table	Delete Table
0	charlie-work-usb-2009-12-11.E01	True	False	True	□ Add	□ Delete
1	terry-work-usb-2009-12-11.E01	True	False	True	□ Add	□ Delete



Disk Image Access for the \
Kam
127.0.0.1:8080/image/charlie-work-usb-2009-12-11.E01/1

BCA Webtools

Home

Browse directories and download files. Items marked "r" in the first column are regular files. Items marked "d" are directories.

d/r	Filename	Size	Last Modified	Deleted?
r	$AttrDef	2560	2009-11-20T17:38:09Z	No
r	$BadClus	0	2009-11-20T17:38:09Z	No
r	$Bitmap	32320	2009-11-20T17:38:09Z	No
r	$Boot	8192	2009-11-20T17:38:09Z	No
d	$Extend	552	2009-11-20T17:38:09Z	No
r	$LogFile	7405568	2009-11-20T17:38:09Z	No
r	$MFT	262144	2009-11-20T17:38:09Z	No
r	$MFTMirr	4096	2009-11-20T17:38:09Z	No
r	$Secure	0	2009-11-20T17:38:09Z	No
r	$UpCase	131072	2009-11-20T17:38:09Z	No
r	$Volume	0	2009-11-20T17:38:09Z	No
d	.	56	2009-12-03T21:17:01Z	No
r	01.zip	108438	2009-11-24T21:21:16Z	No
r	astronaut.jpg	713418	2009-11-24T21:33:33Z	No
r	astronaut1.jpg	722717	2009-11-24T21:43:42Z	No
d	Email	56	2009-12-10T22:27:55Z	No
d	Immortality	56	2009-11-24T21:55:45Z	No
r	invsecr2.exe	1291720	2009-11-19T18:42:25Z	No
r	microscope.jpg	136274	2009-11-24T21:27:51Z	No

Many Professional Decisions – One Example:

- Remote access over the Internet
- Direct access through the reading room
- More highly mediated access through selected surrogates

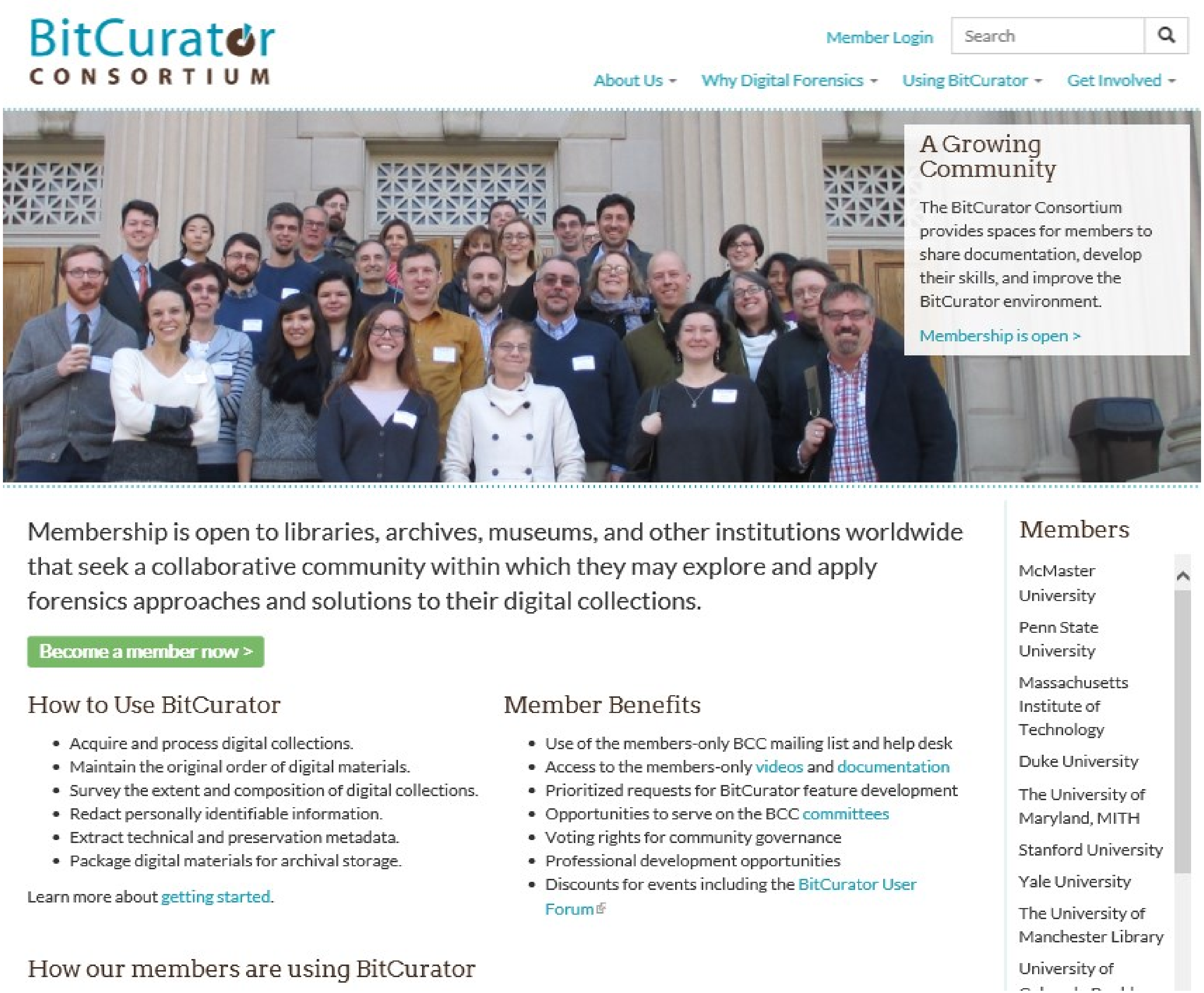
Research questions also abound. A couple examples:

Block-Level vs. File-Level Redaction

Block-Level	File-Level
Finding the bits based on where they appear on the disk, without regard for what files or folders they’re in	Identifying bits to remove based on where they appear in specific files or folders
Usually much faster and simpler to implement than file-level redaction and can find data in unallocated space (deleted content)	Requires more time and more knowledge of underlying data structures (filesystems, file formats), and often requires format-specific tools
Potential to hinder the mounting and navigation of file systems and ability to use/render files	Can require much more human effort, including identification, installation and running of format-specific redaction tools

More Product, Less Process (MPLP) Mean with Born-Digital Materials

- Options I’ve discussed represent significantly different levels of human and computational resources
- Different materials and use cases will warrant different strategies
- How do we decide which to adopt (i.e. what are the decision rules or heuristics)?



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

https://bitcuratorconsortium.org/