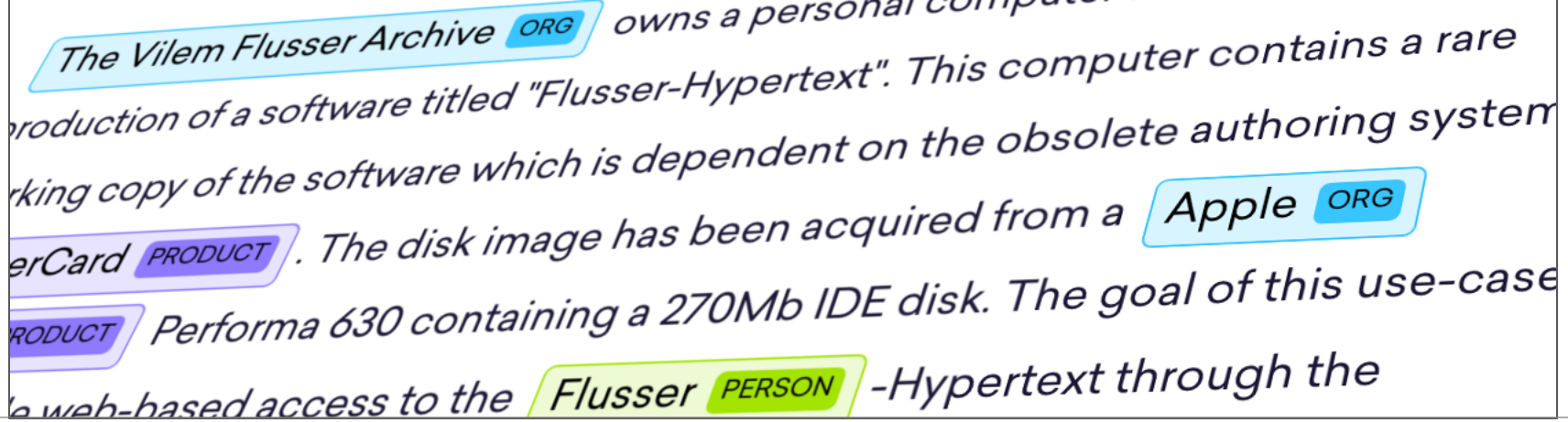


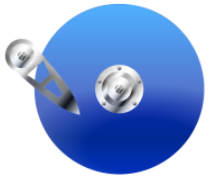
BitCurator NLP: Natural Language Processing for Archivists

SAA Research Forum
August 14, 2018
Washington, DC

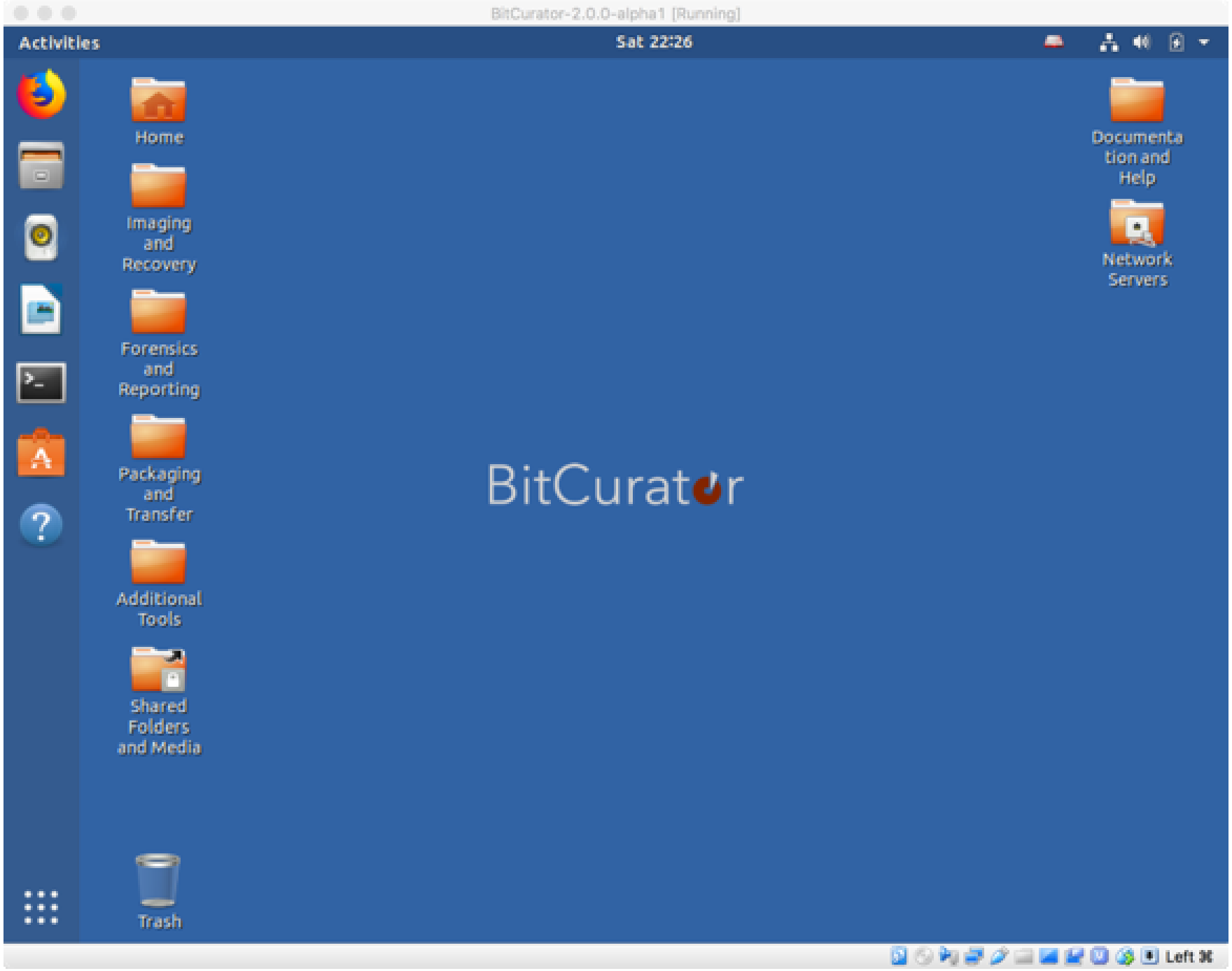
Cal Lee
University of North Carolina at Chapel Hill
School of Information and Library Science

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

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-2.0.0-alpha1 [Running]
Activities
Sat 22:26
Home
Imaging and Recovery
Forensics and Reporting
Packaging and Transfer
Additional Tools
Shared Folders and Media
Trash
Documentation and Help
Network Servers
BitCurator
Left ⌘

BitCurator Access

- 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
- Developed open-source software to support access to disk images. Core areas of focus:
  - Web access services for disk images
  - 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

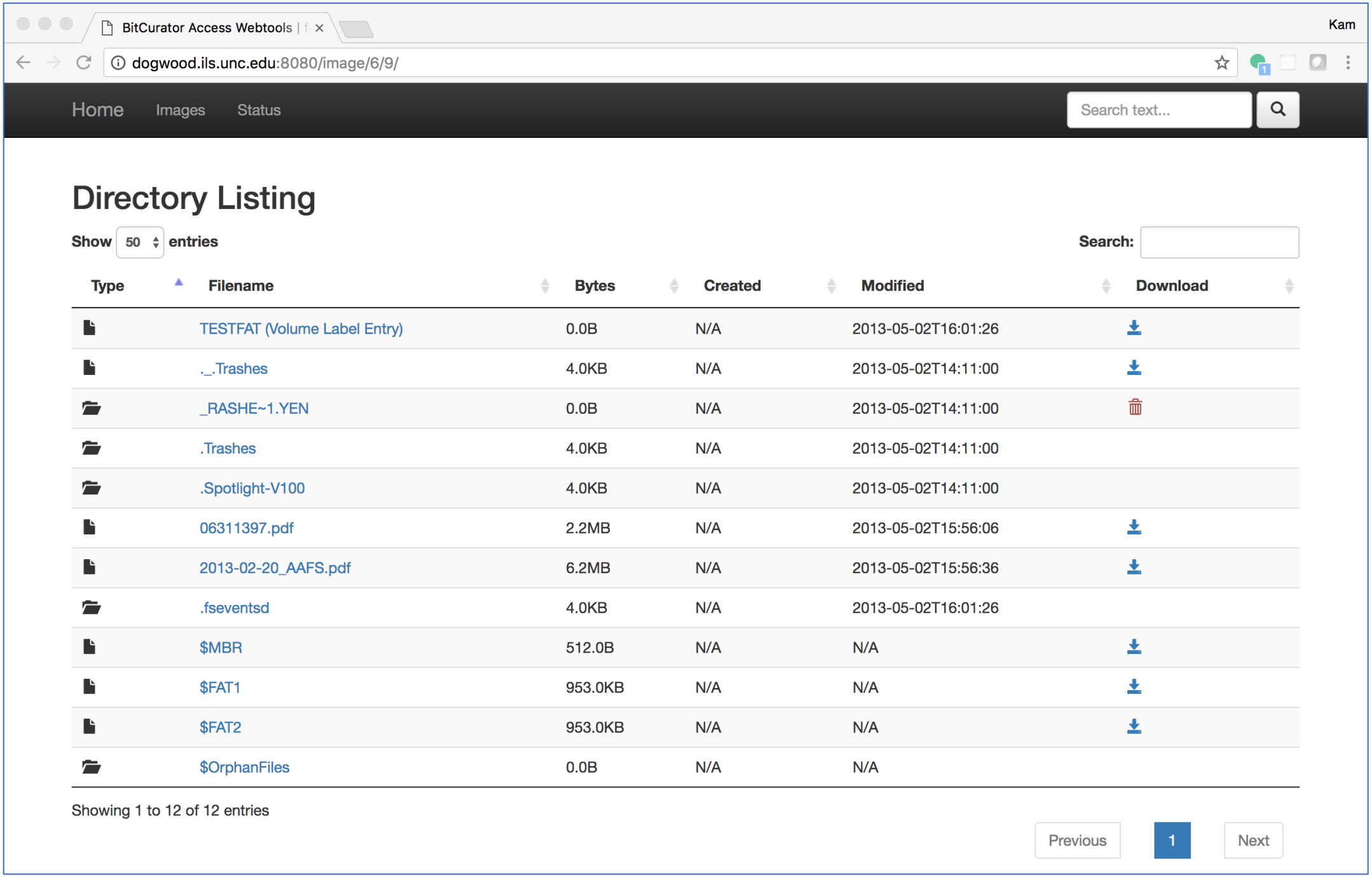
BitCurator NLP

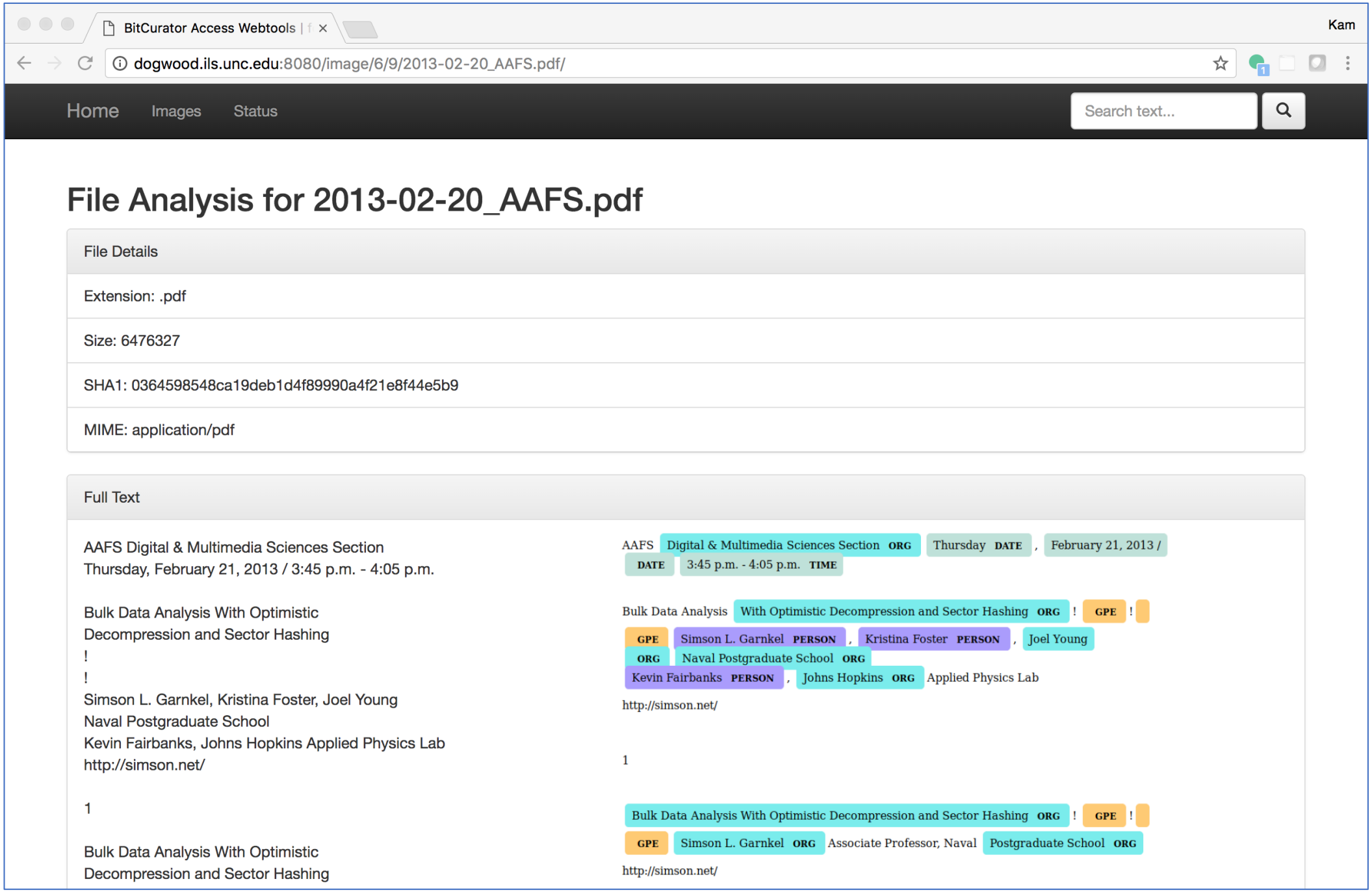
- Funded by Andrew W. Mellon Foundation: October 1, 2016 – September 30, 2018
- Use natural language processing (NLP) to identify and report on items relevant to preservation, information organization, and access
  - Entities (e.g. persons, places, and organizations)
  - Relationships among entities (e.g. appear together within documents or set of documents)
  - Topic models to provide insight into how concepts are clustered

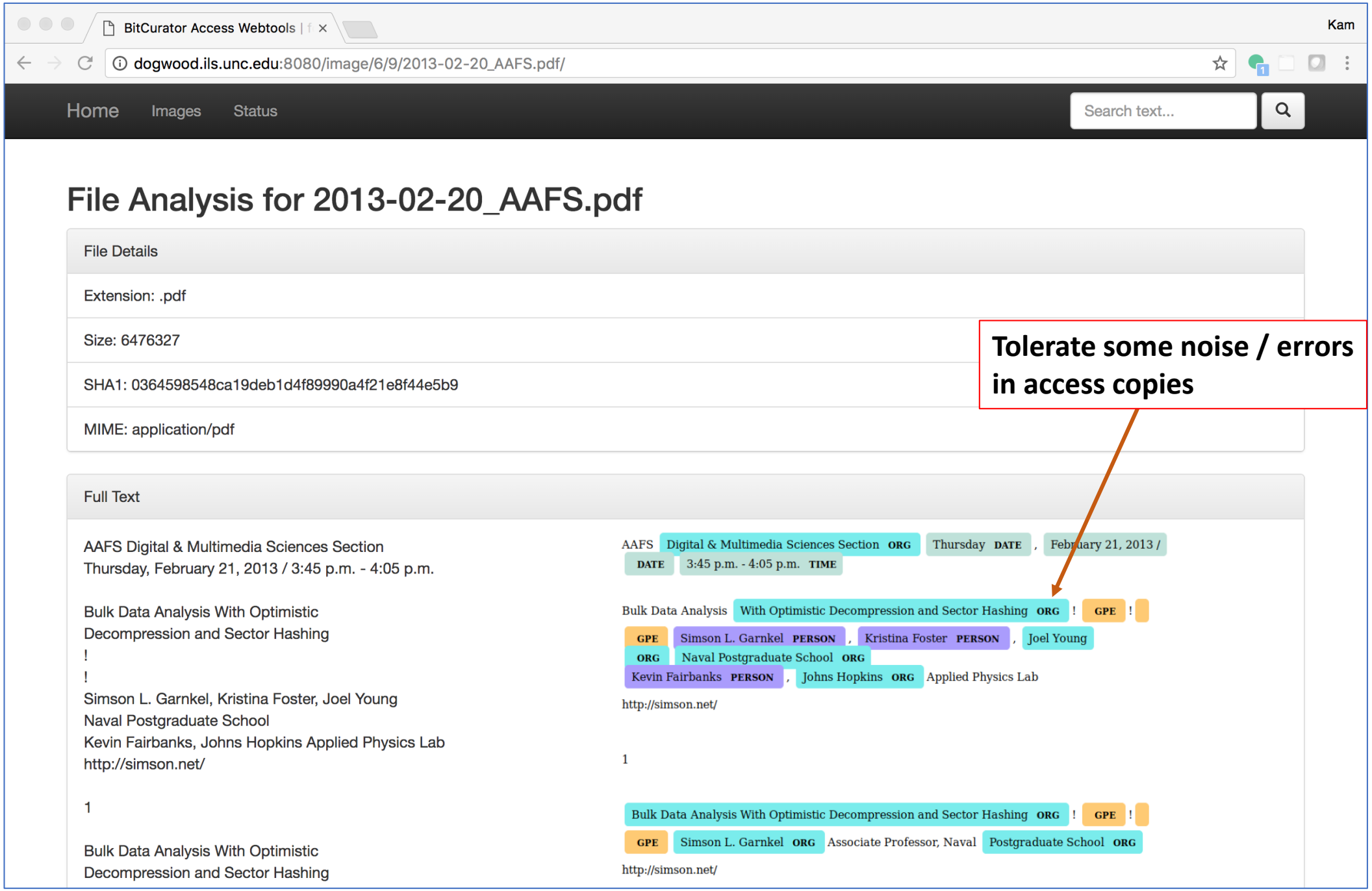
The Challenge of Diverse Archival Collections

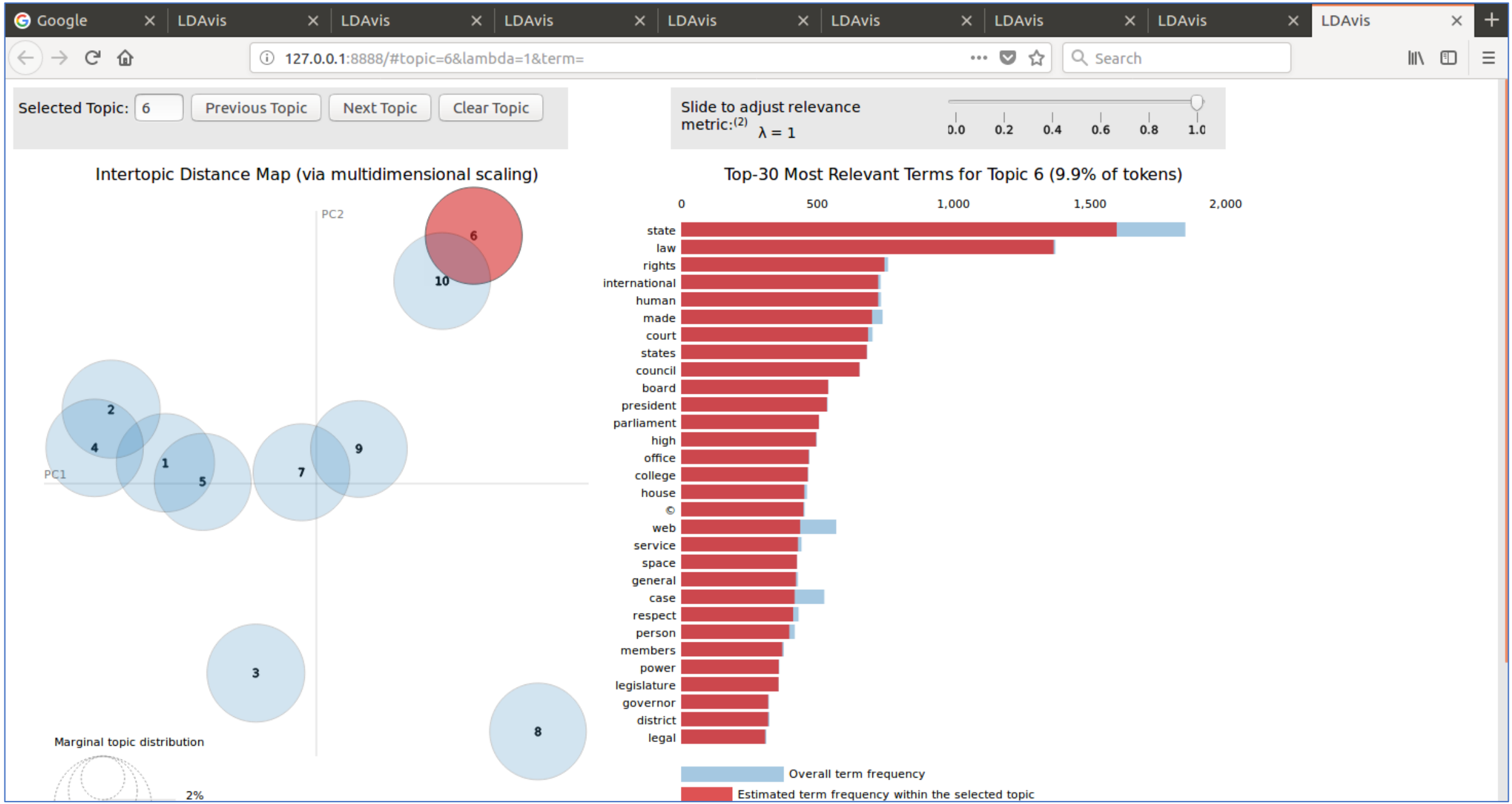
- Most application of NLP is based on relatively consistent, clean collections, but…

- Archival collections can contain:
  - Raw or forensically packaged disk images
  - Many different file types
  - Limited or inconsistent metadata

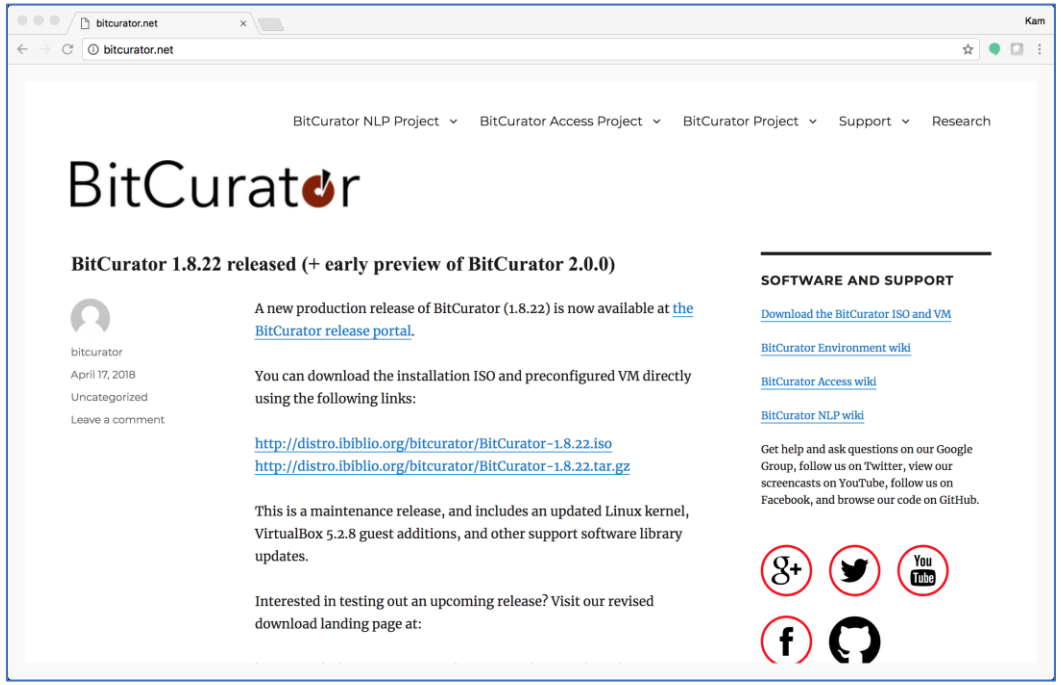

BitCurator Access Webtools
Kam
dogwood.ils.unc.edu:8080/image/6/9/
Home
Images
Status
Search text...
Directory Listing
Show 50 entries
Search:
Type
Filename
Bytes
Created
Modified
Download
TESTFAT (Volume Label Entry) 0.0B N/A 2013-05-02T16:01:26
._.Trashes 4.0KB N/A 2013-05-02T14:11:00
_RASHE~1.YEN 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
Previous
1
Next


BitCurator Access Webtools
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/

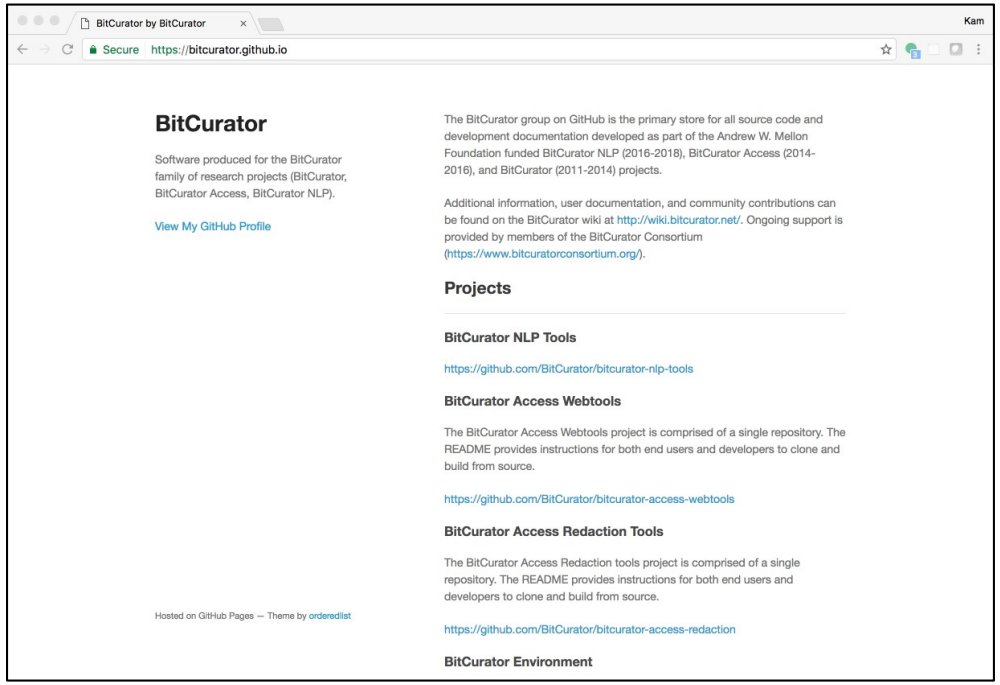

BitCurator Access Webtools
Kam
dogwood.ils.unc.edu:8080/image/6/9/2013-02-20_AAFS.pdf/
Home
Images
Status
Search text...
File Analysis for 2013-02-20_AAFS.pdf
File Details
Extension: .pdf
Size: 6476327
SHA1: 0364598548ca19deb1d4f89990a4f21e8f44e5b9
MIME: application/pdf
Tolerate some noise / errors in access copies
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/


Google
LDAvis
127.0.0.1:8888/#topic=6&lambda=1&term=
Search
Selected Topic: 6
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
6
10
2
4
1
5
7
9
3
8
Marginal topic distribution
2%
Top-30 Most Relevant Terms for Topic 6 (9.9% of tokens)
0
500
1,000
1,500
2,000
state
law
rights
international
human
made
court
states
council
board
president
parliament
high
office
college
house
©
web
service
space
general
case
respect
person
members
power
legislature
governor
district
legal
Overall term frequency
Estimated term frequency within the selected topic

Thank you!



https://bitcurator.net/



https://bitcurator.github.io/