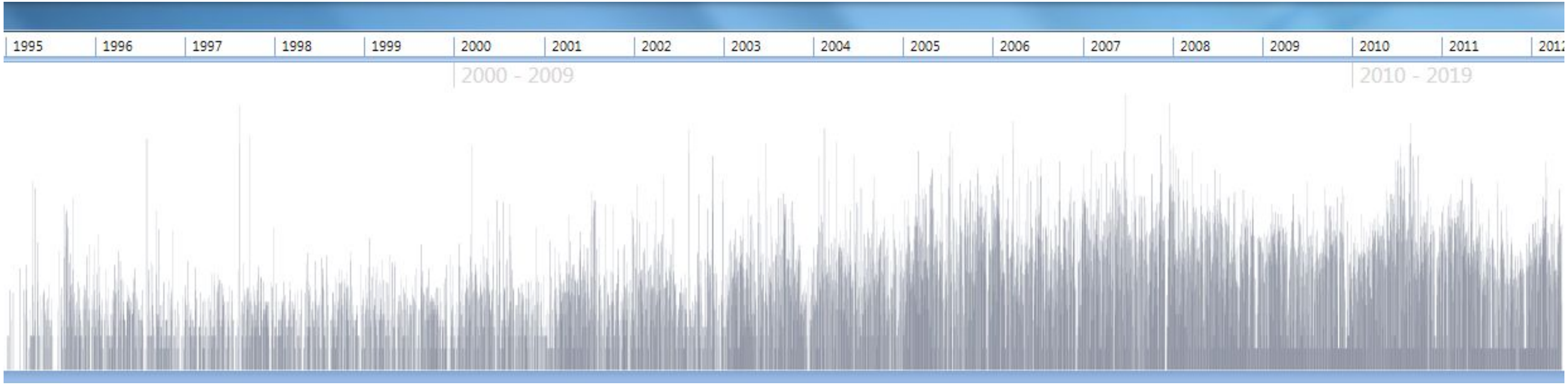
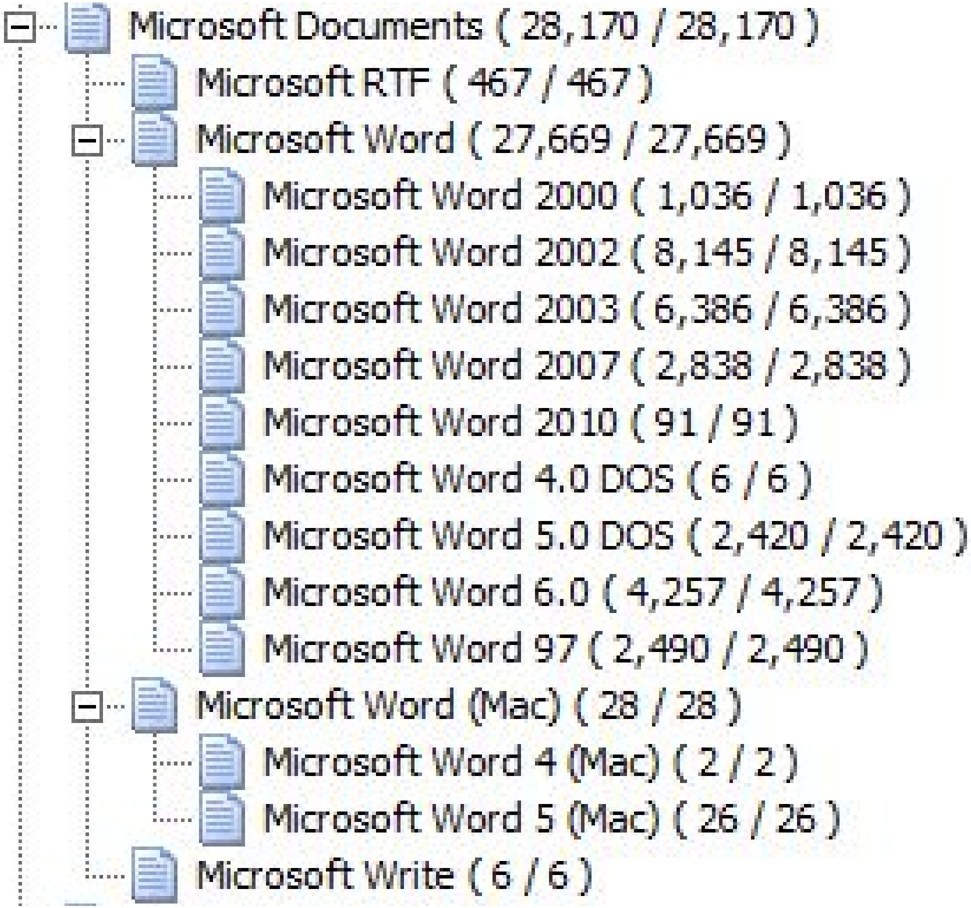
NLP in Archival Processing

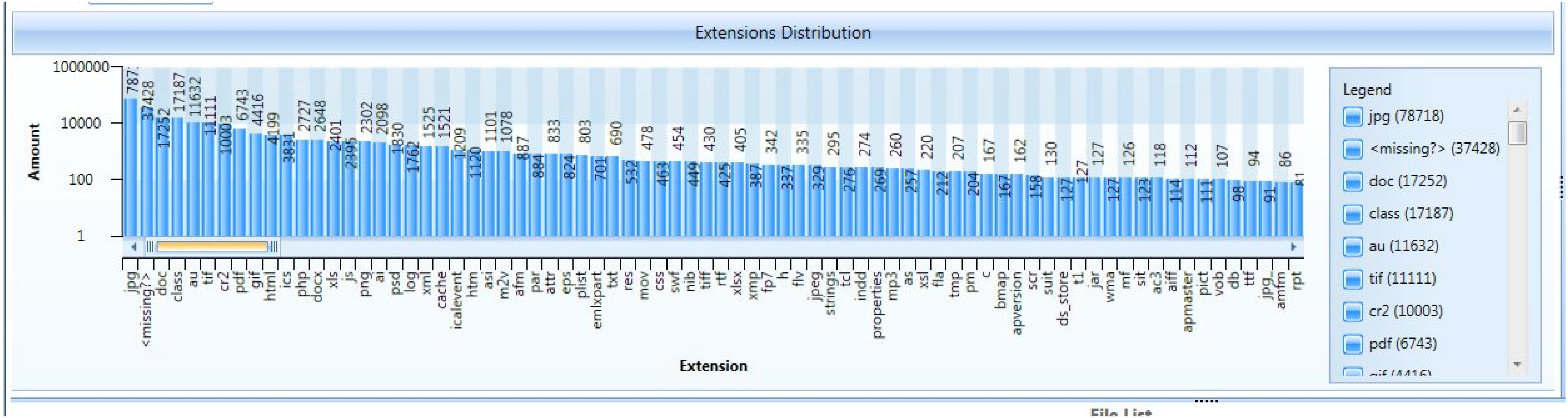
Donald Mennerich, NYU Libraries

Scale


1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2000 - 2009
2010 - 2019

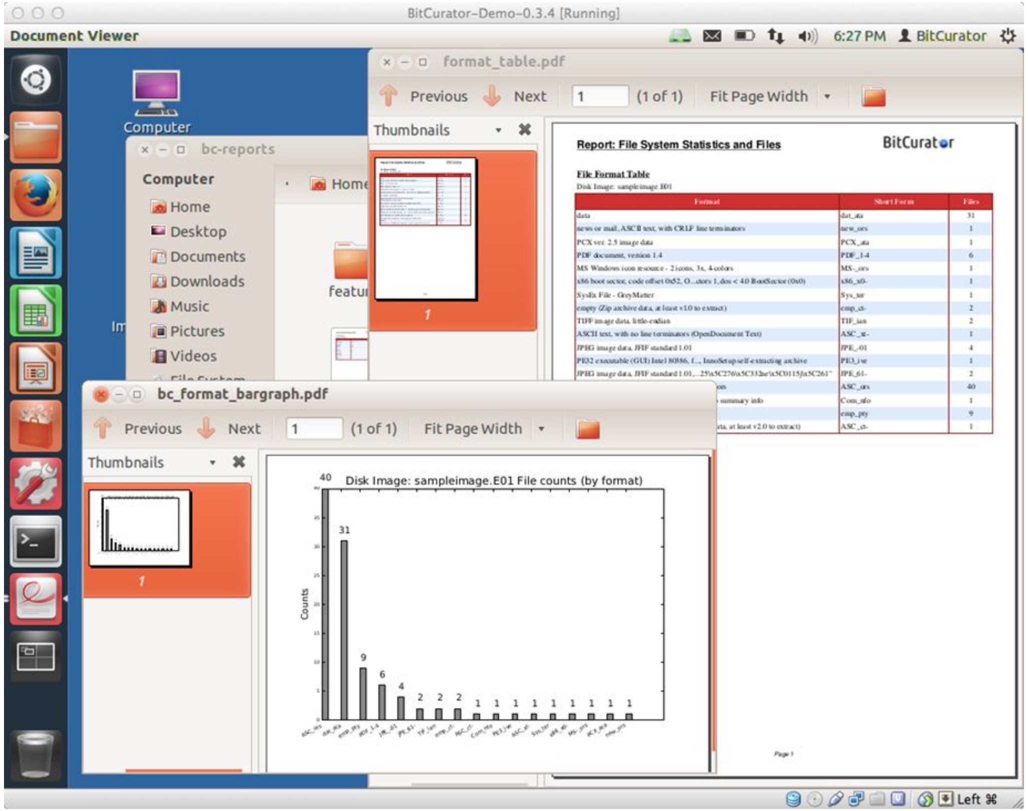


```
Microsoft Documents ( 28,170 / 28,170 )
    Microsoft RTF ( 467 / 467 )
    Microsoft Word ( 27,669 / 27,669 )
        Microsoft Word 2000 ( 1,036 / 1,036 )
        Microsoft Word 2002 ( 8,145 / 8,145 )
        Microsoft Word 2003 ( 6,386 / 6,386 )
        Microsoft Word 2007 ( 2,838 / 2,838 )
        Microsoft Word 2010 ( 91 / 91 )
        Microsoft Word 4.0 DOS ( 6 / 6 )
        Microsoft Word 5.0 DOS ( 2,420 / 2,420 )
        Microsoft Word 6.0 ( 4,257 / 4,257 )
        Microsoft Word 97 ( 2,490 / 2,490 )
    Microsoft Word (Mac) ( 28 / 28 )
        Microsoft Word 4 (Mac) ( 2 / 2 )
        Microsoft Word 5 (Mac) ( 26 / 26 )
    Microsoft Write ( 6 / 6 )
```


Extensions Distribution
Amount
1000000
10000
100
1
787
37428
17252
17187
11632
11111
10003
6743
4416
4199
3831
2727
2648
2401
2395
2302
2098
1830
1762
1525
1521
1209
1120
1101
1078
887
884
833
824
803
701
690
532
478
463
454
449
430
425
405
387
342
337
335
329
295
276
274
269
260
257
220
212
207
204
167
167
162
158
130
127
127
127
127
126
123
118
114
112
111
107
98
94
91
86
81
jpg
<missing?>
doc
class
au
tif
cr2
pdf
gif
html
ics
php
docx
xls
js
png
ai
psd
log
xml
cache
icalevent
htm
asi
m2v
afm
par
attr
eps
plist
emlxpart
txt
res
mov
css
swf
nib
tiff
rtf
xlsx
xmp
fp7
h
flv
jpeg
strings
tcl
indd
properties
mp3
as
xsl
fla
tmp
pm
c
bmap
apversion
scr
suit
ds_store
t1
jar
wma
mf
sit
ac3
aiff
apmaster
pict
vob
db
ttf
jpg_
amfm
rpt
Extension
Legend
jpg (78718)
<missing?> (37428)
doc (17252)
class (17187)
au (11632)
tif (11111)
cr2 (10003)
pdf (6743)
gif (4416)
File List

Forensics

```xml
<!-- plugin_process -->

<pronomPuid>x-fmt/391</pronomPuid>

<pronomFormatName>Exchangeable Image File Format (Compressed)</pronomFormatName>

<pronomSignatureName>EXIF Compressed Image 2.2</pronomSignatureName>

<pronomMimeType>image/jpeg</pronomMimeType>

<pronomMatchType>signature</pronomMatchType>

<pronomSignatureFileVersion>formats-v70.xml</pronomSignatureFileVersion>

<pronomContainerFileVersion>20130501.xml</pronomContainerFileVersion>

<fidoVersion>1.3.1</fidoVersion>

<identificationUuid>49050200-e308-4060-886a-14a8efd82078</identificationUuid>

<scanStatus>PASSED</scanStatus>

<clamAVVersion>ClamAV 0.98.1</clamAVVersion>

<virusScanUuid>6dea8d54-107a-43d0-a1fe-beb9c6bc4a21</virusScanUuid>
```



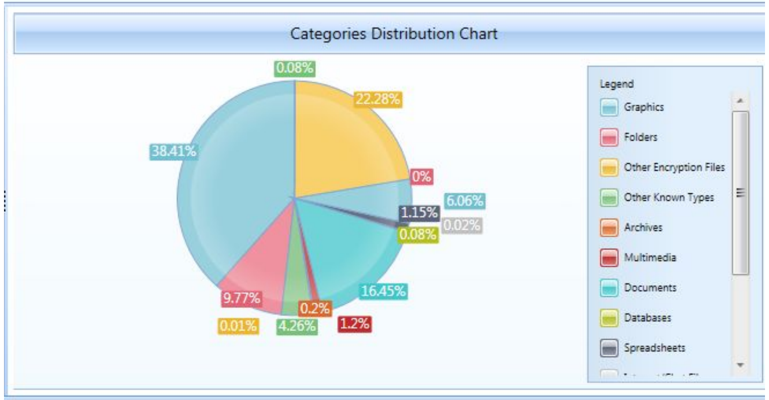
Report: File System Statistics and Files

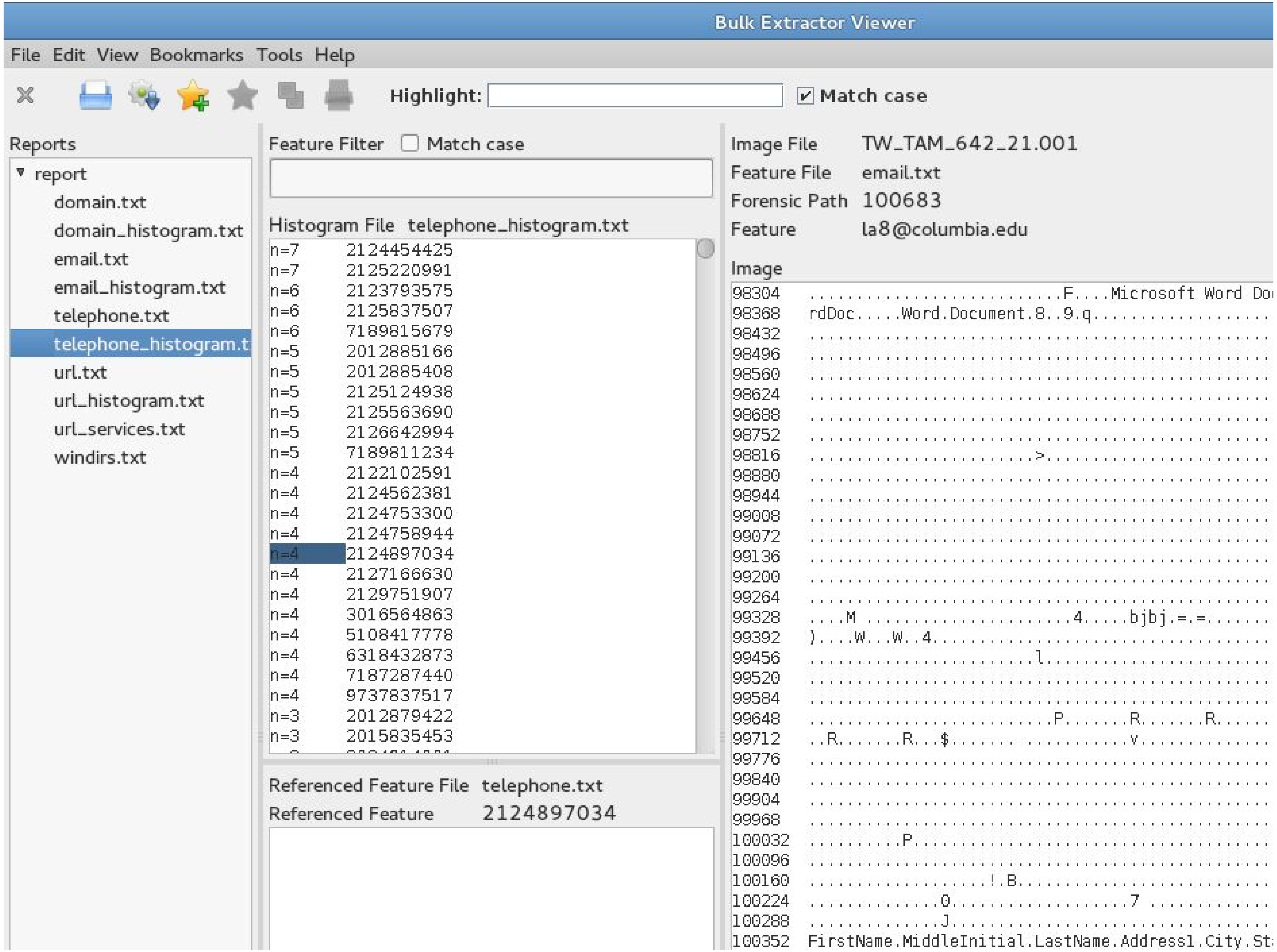
BitCurator

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\x5C332he\x5C0115j\x5C261"	JPE_61-	2
ons	ASC_ors	40
summary info	Com_nfo	1
	emp_pty	9
ata, at least v2.0 to extract)	ASC_ct-	1





▾ report
- domain.txt
- domain_histogram.txt
- email.txt
- email_histogram.txt
- telephone.txt
- telephone_histogram.t
- url.txt
- url_histogram.txt
- url_services.txt
- windirs.txt

Histogram File telephone_histogram.txt

```
n=7     2124454425
n=7     2125220991
n=6     2123793575
n=6     2125837507
n=6     7189815679
n=5     2012885166
n=5     2012885408
n=5     2125124938
n=5     2125563690
n=5     2126642994
n=5     7189811234
n=4     2122102591
n=4     2124562381
n=4     2124753300
n=4     2124758944
n=4     2124897034
n=4     2127166630
n=4     2129751907
n=4     3016564863
n=4     5108417778
n=4     6318432873
n=4     7187287440
n=4     9737837517
n=3     2012879422
n=3     2015835453
```

Referenced Feature File telephone.txt

Referenced Feature 2124897034

Image File TW_TAM_642_21.001

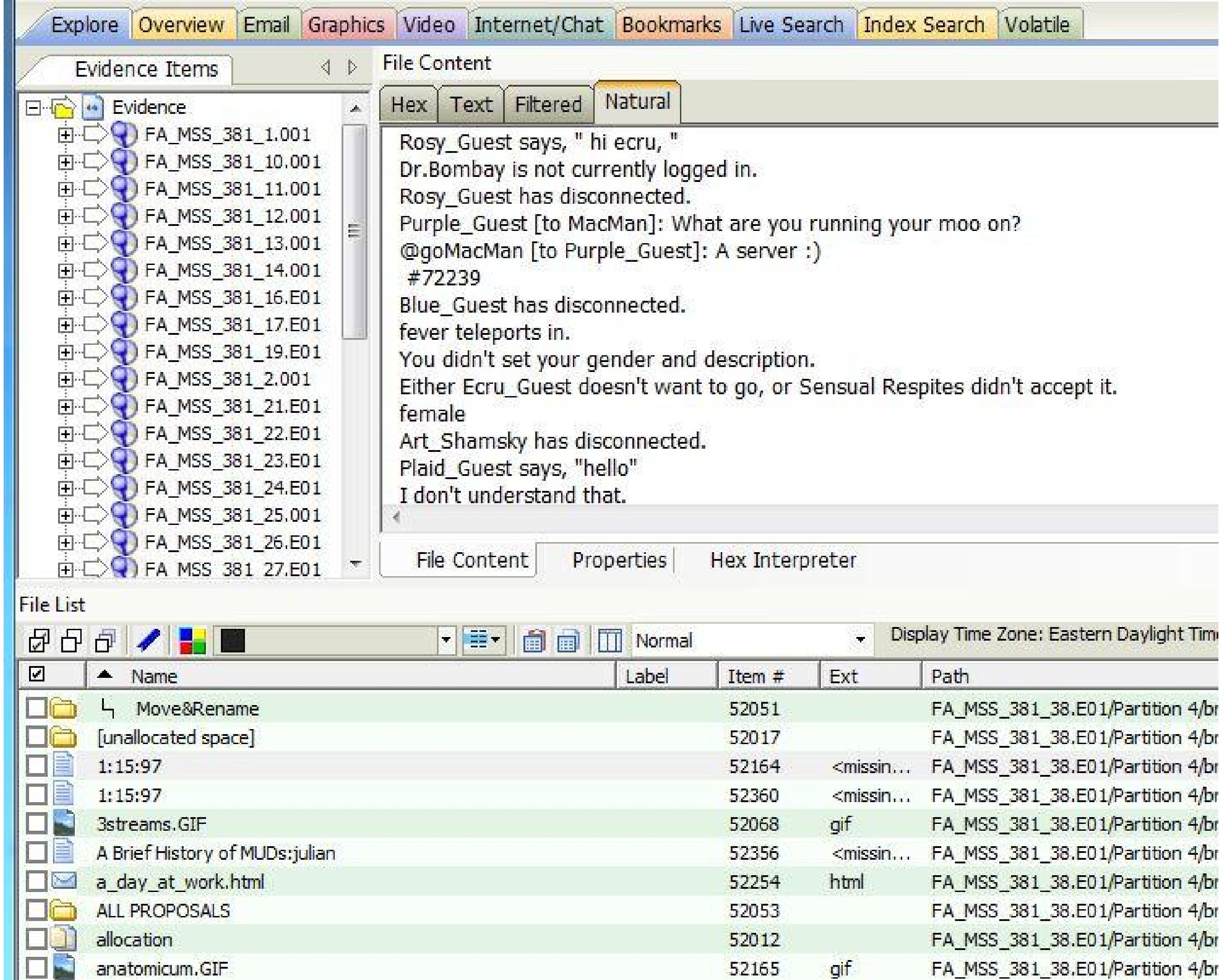
Feature File email.txt

Forensic Path 100683

Feature la8@columbia.edu

Image

```
98304   ..........................F....Microsoft Word Do
98368   rdDoc.....Word.Document.8..9.q..................
98432   ................................................
98496   ................................................
98560   ................................................
98624   ................................................
98688   ................................................
98752   ................................................
98816   ........................>.......................
98880   ................................................
98944   ................................................
99008   ................................................
99072   ................................................
99136   ................................................
99200   ................................................
99264   ................................................
99328   ....M ......................4....bjbj.=.=.......
99392   )....W...W..4...................................
99456   ........................l.......................
99520   ................................................
99584   ................................................
99648   ..........................P.......R.......R.....
99712   ..R.......R...$...................v.............
99776   ................................................
99840   ................................................
99904   ................................................
99968   ................................................
100032  ..........P.....................................
100096  ................................................
100160  ...................!.B..........................
100224  ..............0...................7 ............
100288  ..............J.................................
100352  FirstName.MiddleInitial.LastName.Address1.City.St
```



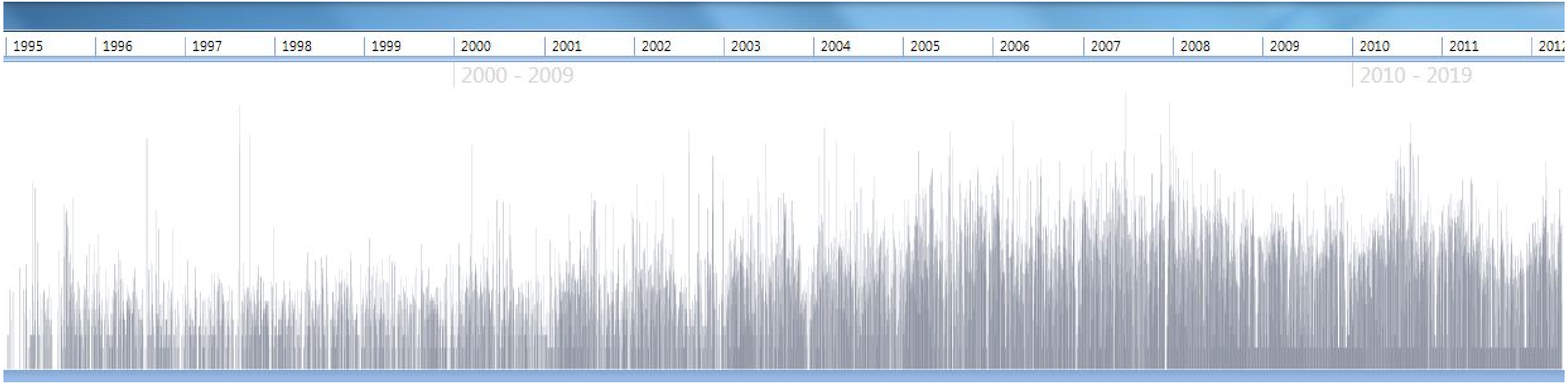
Evidence
- FA_MSS_381_1.001
- FA_MSS_381_10.001
- FA_MSS_381_11.001
- FA_MSS_381_12.001
- FA_MSS_381_13.001
- FA_MSS_381_14.001
- FA_MSS_381_16.E01
- FA_MSS_381_17.E01
- FA_MSS_381_19.E01
- FA_MSS_381_2.001
- FA_MSS_381_21.E01
- FA_MSS_381_22.E01
- FA_MSS_381_23.E01
- FA_MSS_381_24.E01
- FA_MSS_381_25.001
- FA_MSS_381_26.E01
- FA_MSS_381_27.E01

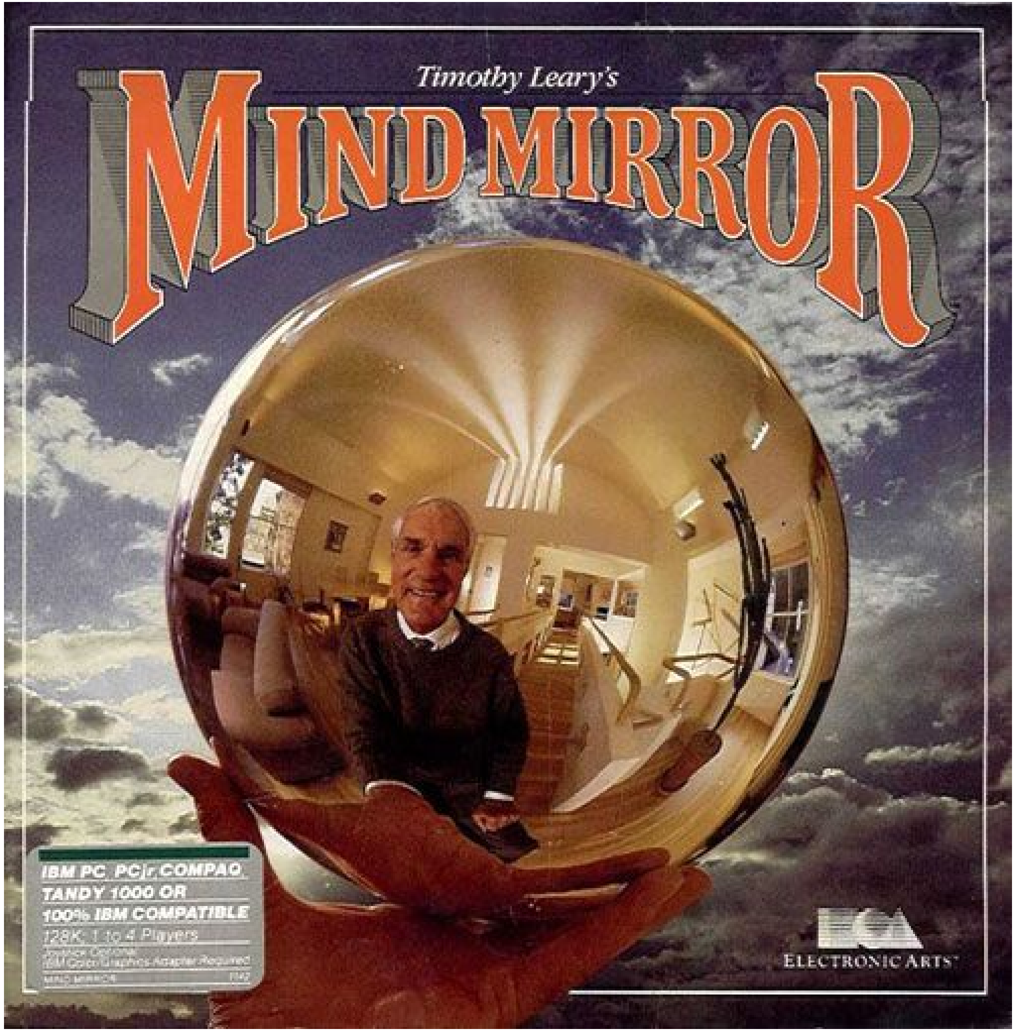
```
Rosy_Guest says, " hi ecru, "
Dr.Bombay is not currently logged in.
Rosy_Guest has disconnected.
Purple_Guest [to MacMan]: What are you running your moo on?
@goMacMan [to Purple_Guest]: A server :)
 #72239
Blue_Guest has disconnected.
fever teleports in.
You didn't set your gender and description.
Either Ecru_Guest doesn't want to go, or Sensual Respites didn't accept it.
female
Art_Shamsky has disconnected.
Plaid_Guest says, "hello"
I don't understand that.
```

Display Time Zone: Eastern Daylight Tim

☑	Name	Label	Item #	Ext	Path
□	Move&Rename		52051		FA_MSS_381_38.E01/Partition 4/br
□	[unallocated space]		52017		FA_MSS_381_38.E01/Partition 4/br
□	1:15:97		52164	<missin…	FA_MSS_381_38.E01/Partition 4/br
□	1:15:97		52360	<missin…	FA_MSS_381_38.E01/Partition 4/br
□	3streams.GIF		52068	gif	FA_MSS_381_38.E01/Partition 4/br
□	A Brief History of MUDs:julian		52356	<missin…	FA_MSS_381_38.E01/Partition 4/br
□	a_day_at_work.html		52254	html	FA_MSS_381_38.E01/Partition 4/br
□	ALL PROPOSALS		52053		FA_MSS_381_38.E01/Partition 4/br
□	allocation		52012		FA_MSS_381_38.E01/Partition 4/br
□	anatomicum.GIF		52165	gif	FA_MSS_381_38.E01/Partition 4/br

Scale


1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
201
2000 - 2009
2010 - 2019


Timothy Leary's
MIND MIRROR
IBM PC, PCjr, COMPAQ,
TANDY 1000 OR
100% IBM COMPATIBLE
128K; 1 to 4 Players
Joystick Optional
IBM Color/Graphics Adapter Required
MIND MIRROR
1142
ELECTRONIC ARTS

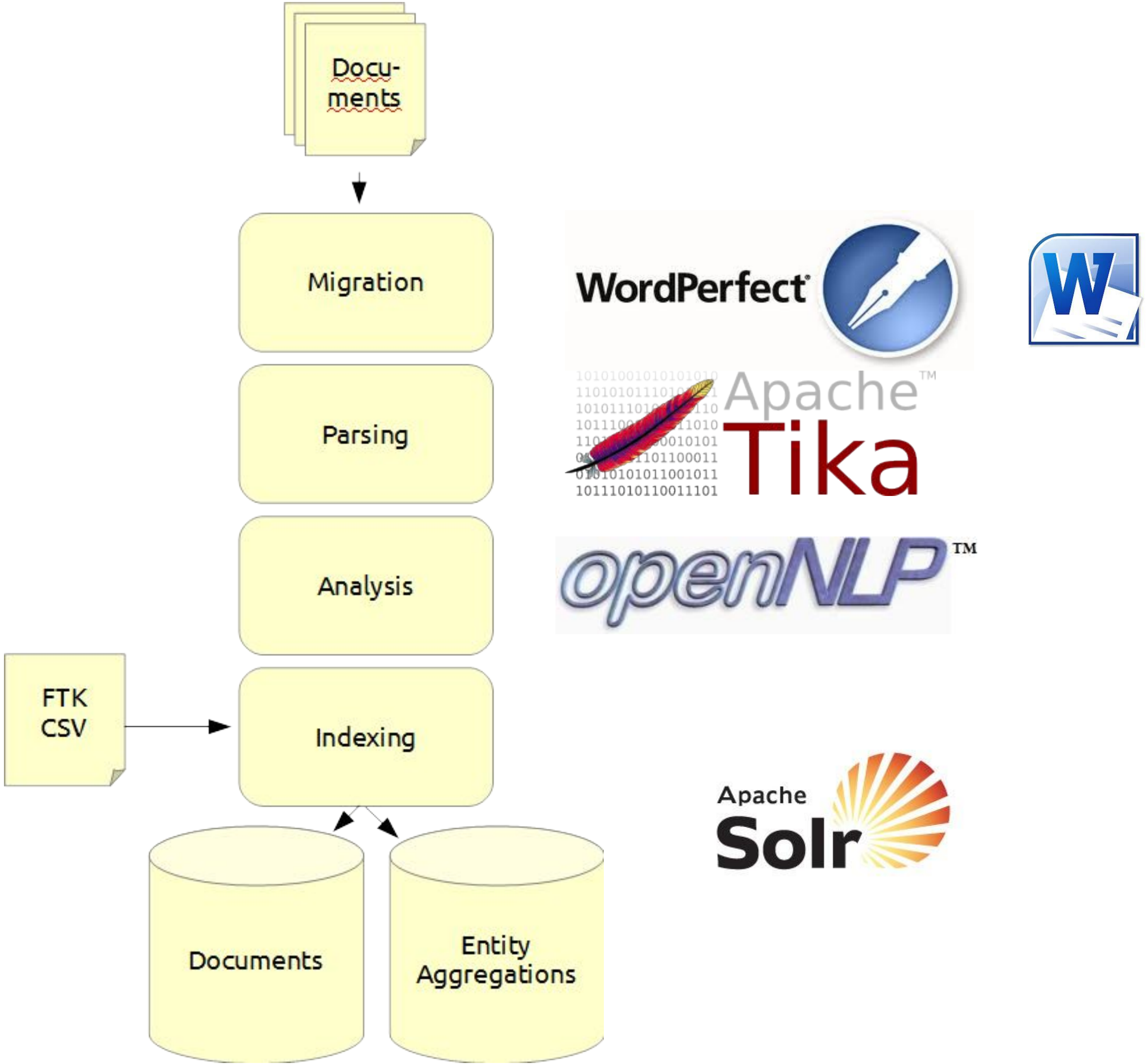
REBEL HERO OF THE WEEK

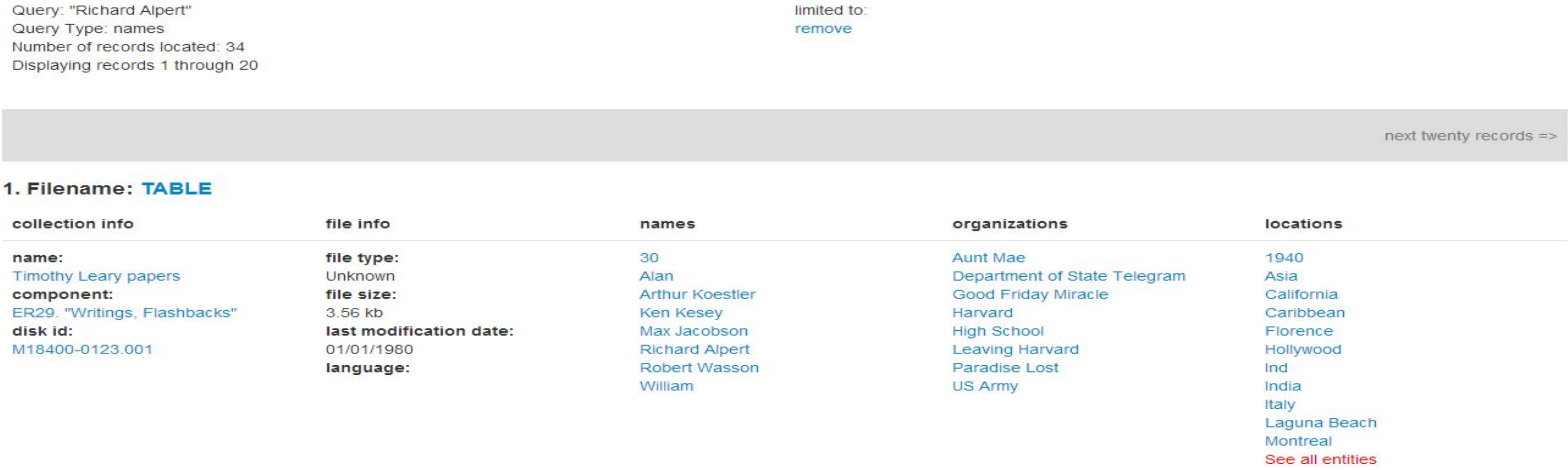
4THversion
Dec.28 1983

CARY GRANT

(Nominated by Timothy Leary)

Thanks to the detergent vigilence of our media, the information we get is laundered, starched and squeaky clean. So very few Americans remember that in the late 1950's the most prominent advocate of LSD and chemical altered states was none other than CARY GRANT. His rapturous descriptions of his 100 trips helped launch the psychedelic movement of the '60's.


Docu-ments
Migration
Parsing
Analysis
Indexing
FTK CSV
Documents
Entity Aggregations
WordPerfect
W
Apache Tika
openNLP
Apache Solr



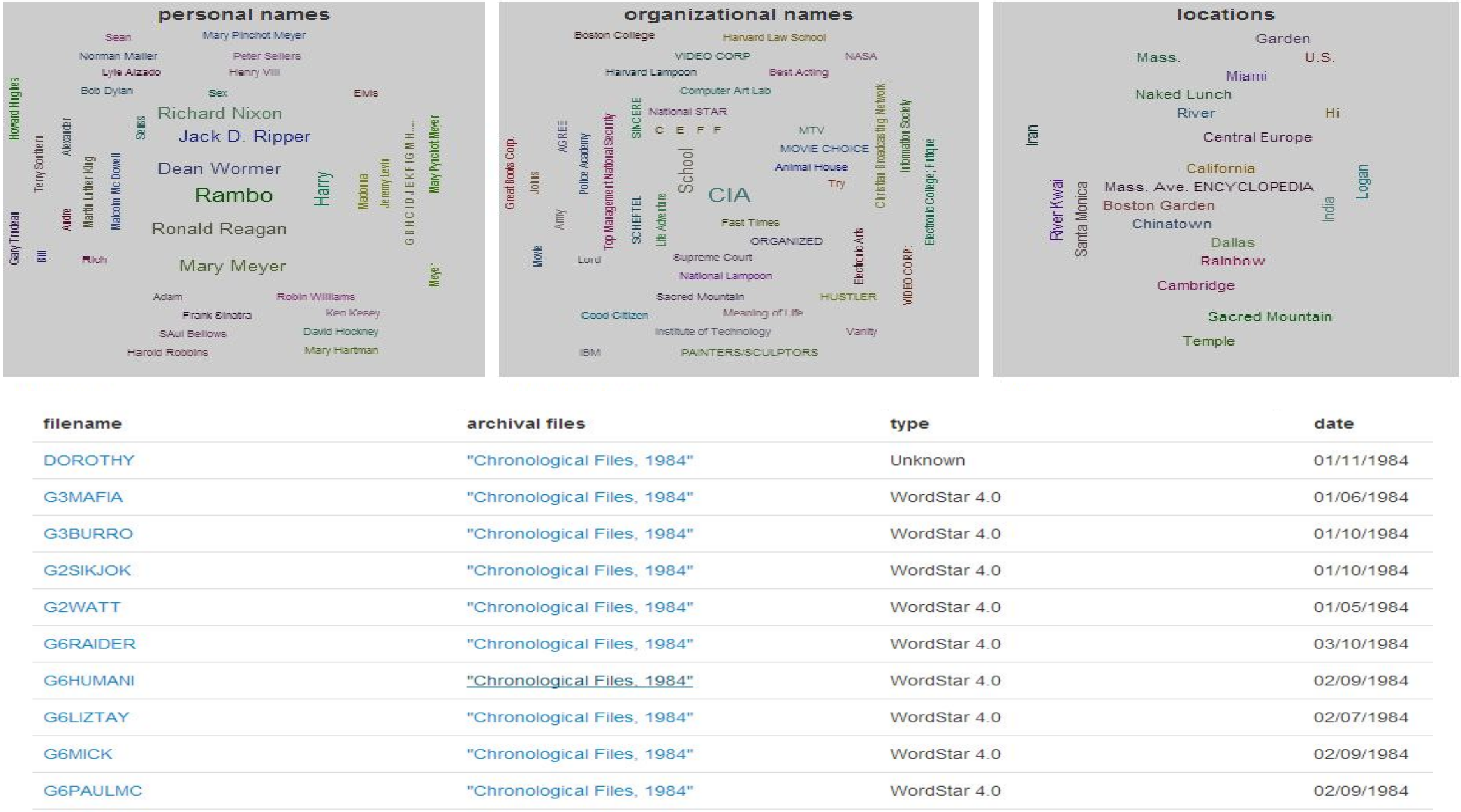
Query: "Richard Alpert"
Query Type: names
Number of records located: 34
Displaying records 1 through 20

limited to:
remove

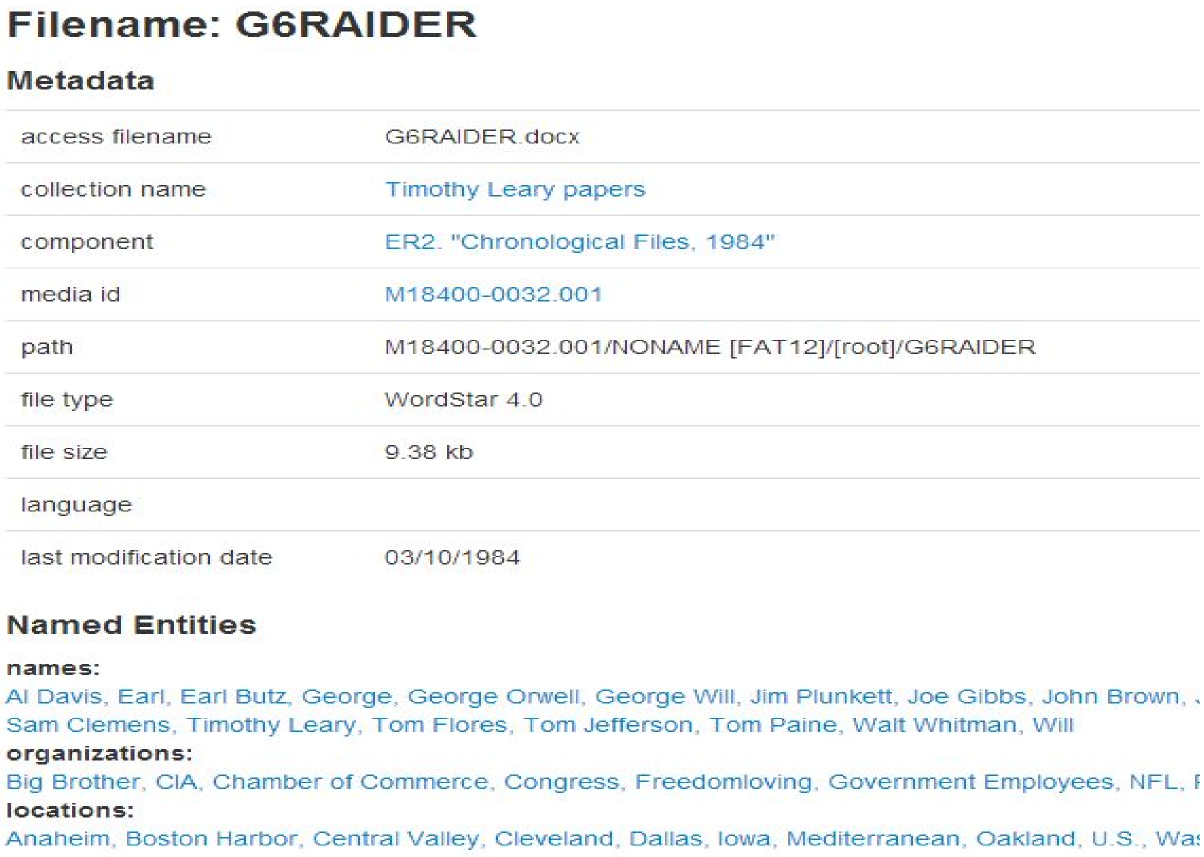
next twenty records =>

1. Filename: TABLE

collection info	file info	names	organizations	locations
name: Timothy Leary papers **component:** ER29. "Writings, Flashbacks" **disk id:** M18400-0123.001	**file type:** Unknown **file size:** 3.56 kb **last modification date:** 01/01/1980 **language:**	30 Alan Arthur Koestler Ken Kesey Max Jacobson Richard Alpert Robert Wasson William	Aunt Mae Department of State Telegram Good Friday Miracle Harvard High School Leaving Harvard Paradise Lost US Army	1940 Asia California Caribbean Florence Hollywood Ind India Italy Laguna Beach Montreal See all entities



filename	archival files	type	date
DOROTHY	"Chronological Files, 1984"	Unknown	01/11/1984
G3MAFIA	"Chronological Files, 1984"	WordStar 4.0	01/06/1984
G3BURRO	"Chronological Files, 1984"	WordStar 4.0	01/10/1984
G2SIKJOK	"Chronological Files, 1984"	WordStar 4.0	01/10/1984
G2WATT	"Chronological Files, 1984"	WordStar 4.0	01/05/1984
G6RAIDER	"Chronological Files, 1984"	WordStar 4.0	03/10/1984
G6HUMANI	"Chronological Files, 1984"	WordStar 4.0	02/09/1984
G6LIZTAY	"Chronological Files, 1984"	WordStar 4.0	02/07/1984
G6MICK	"Chronological Files, 1984"	WordStar 4.0	02/09/1984
G6PAULMC	"Chronological Files, 1984"	WordStar 4.0	02/09/1984



Filename: G6RAIDER

Metadata

access filename	G6RAIDER.docx
collection name	Timothy Leary papers
component	ER2. "Chronological Files, 1984"
media id	M18400-0032.001
path	M18400-0032.001/NONAME [FAT12]/[root]/G6RAIDER
file type	WordStar 4.0
file size	9.38 kb
language	
last modification date	03/10/1984

Named Entities

names:
Al Davis, Earl, Earl Butz, George, George Orwell, George Will, Jim Plunkett, Joe Gibbs, John Brown,
Sam Clemens, Timothy Leary, Tom Flores, Tom Jefferson, Tom Paine, Walt Whitman, Will

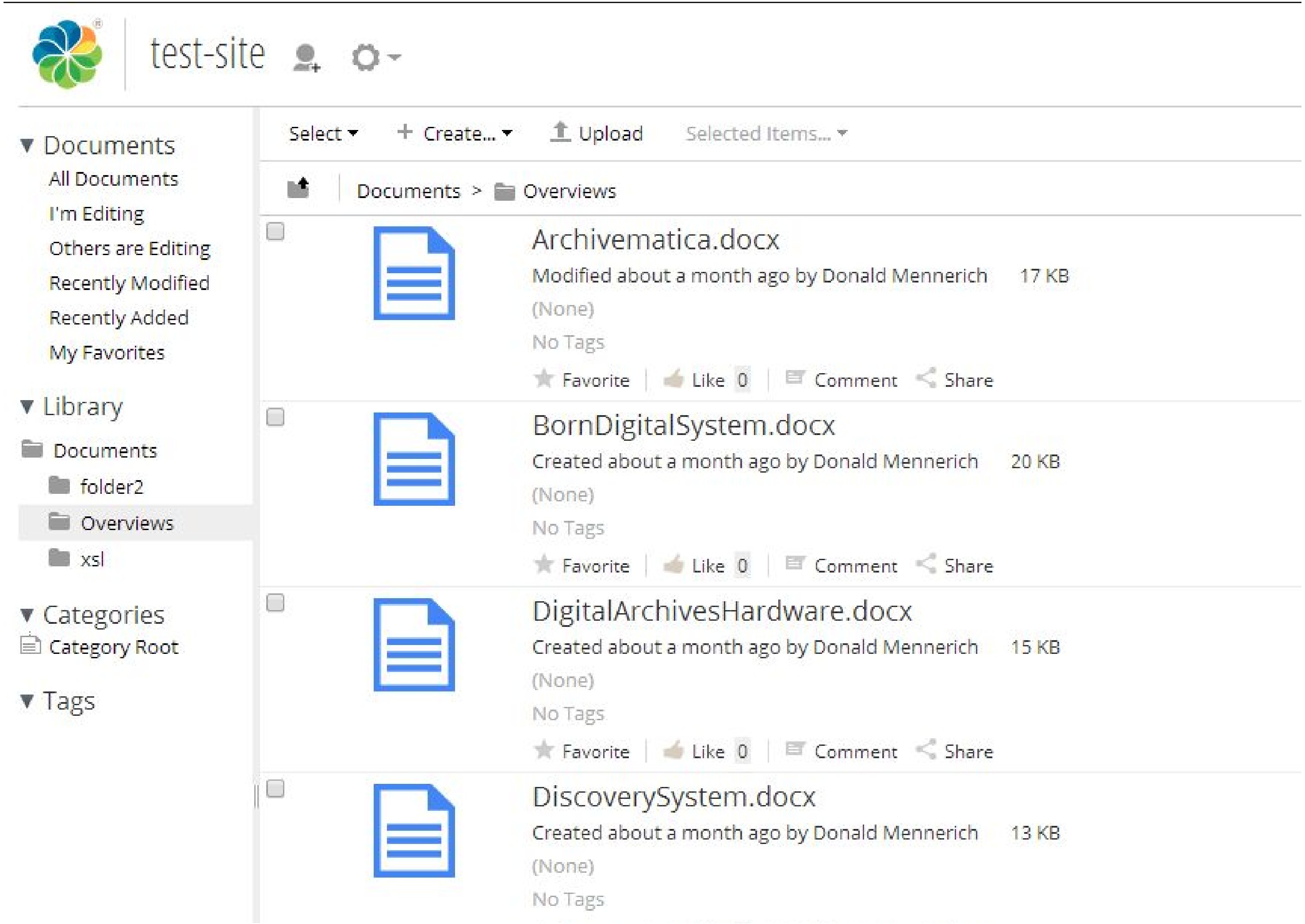
organizations:
Big Brother, CIA, Chamber of Commerce, Congress, Freedomloving, Government Employees, NFL,

locations:
Anaheim, Boston Harbor, Central Valley, Cleveland, Dallas, Iowa, Mediterranean, Oakland, U.S., Was

improvements

- Better infrastructure, distributed processing, machine Learning
- Topic modeling, cluster analysis, document similarity
- Visualizations
- Integration with discovery, dissemination and access systems, Linked open data

Beyond the obsolete…

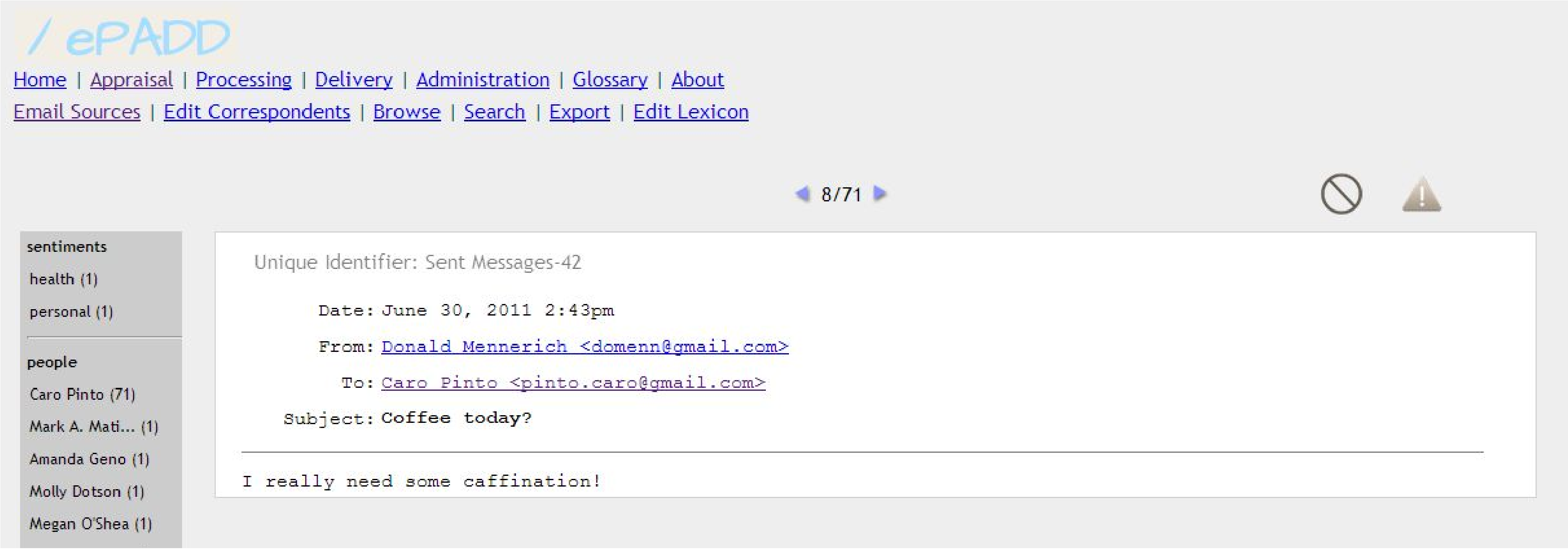


Archivematica.docx
Modified about a month ago by Donald Mennerich 17 KB
(None)
No Tags
Favorite | Like 0 | Comment Share

BornDigitalSystem.docx
Created about a month ago by Donald Mennerich 20 KB
(None)
No Tags
Favorite | Like 0 | Comment Share

DigitalArchivesHardware.docx
Created about a month ago by Donald Mennerich 15 KB
(None)
No Tags
Favorite | Like 0 | Comment Share

DiscoverySystem.docx
Created about a month ago by Donald Mennerich 13 KB
(None)
No Tags



/ ePADD
Home | Appraisal | Processing | Delivery | Administration | Glossary | About
Email Sources | Edit Correspondents | Browse | Search | Export | Edit Lexicon
8/71
sentiments
health (1)
personal (1)
people
Caro Pinto (71)
Mark A. Mati... (1)
Amanda Geno (1)
Molly Dotson (1)
Megan O'Shea (1)

Unique Identifier: Sent Messages-42

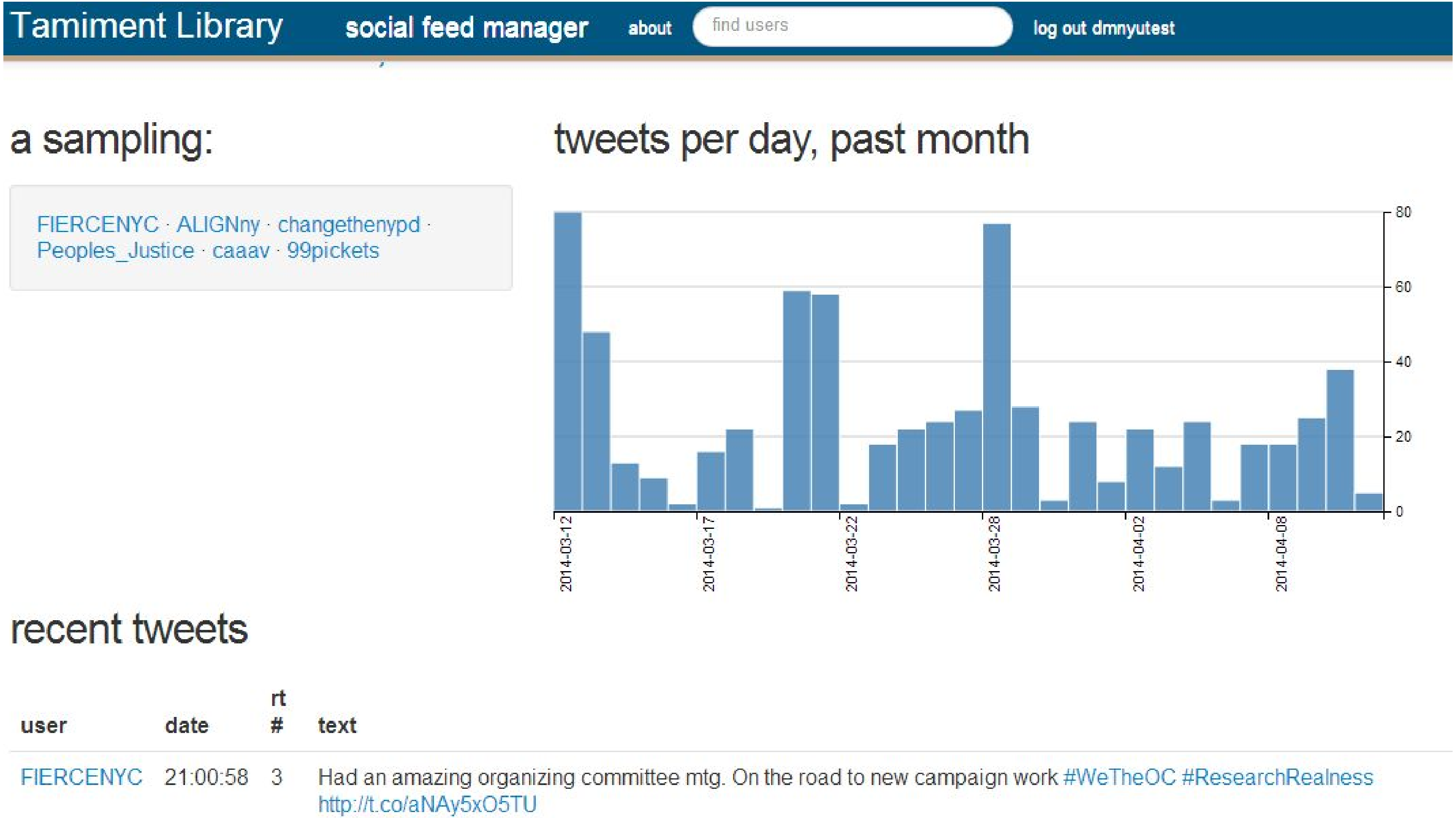
Date: June 30, 2011 2:43pm

From: Donald Mennerich <domenn@gmail.com>

To: Caro Pinto <pinto.caro@gmail.com>

Subject: Coffee today?

I really need some caffination!



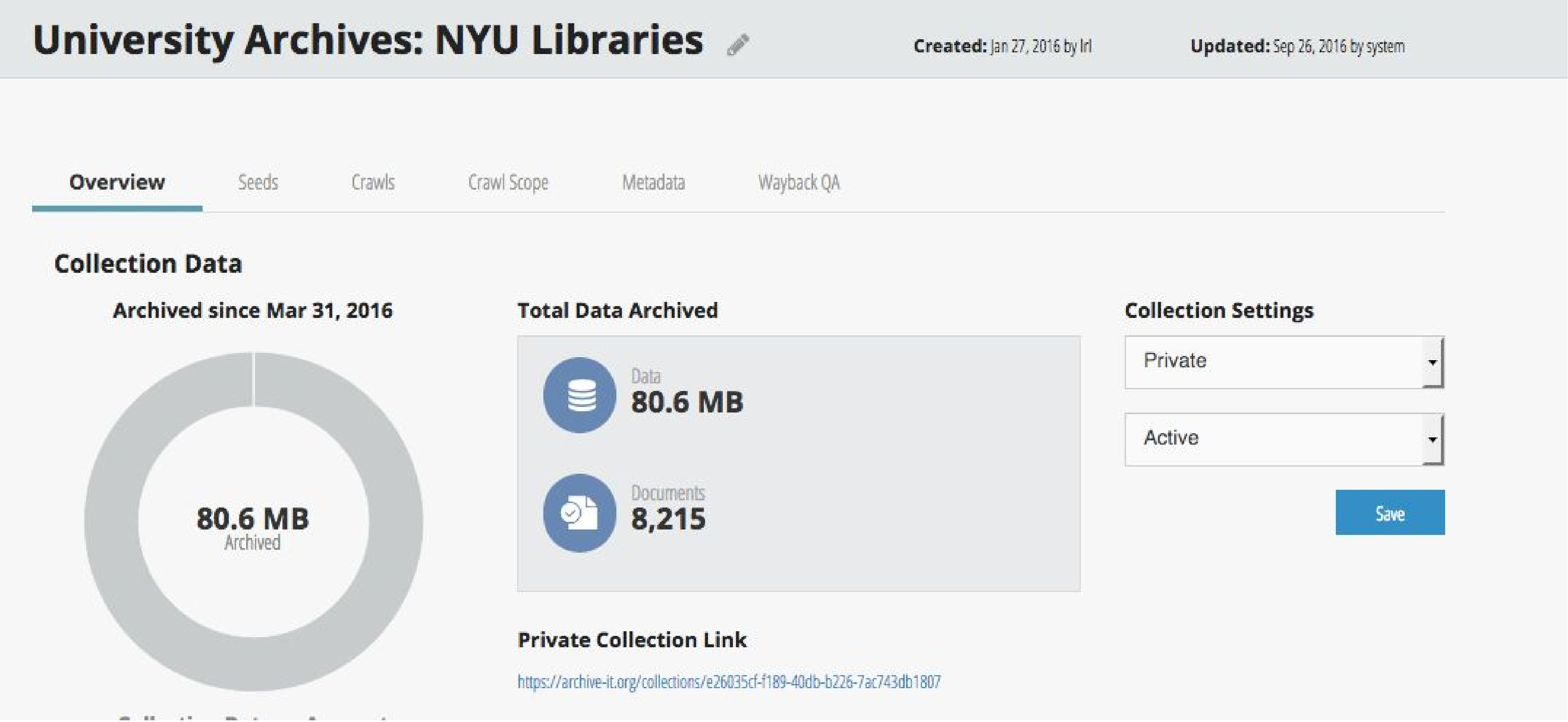
a sampling:

FIERCENYC · ALIGNny · changethenypd · Peoples_Justice · caaav · 99pickets

tweets per day, past month

recent tweets

user	date	rt #	text
FIERCENYC	21:00:58	3	Had an amazing organizing committee mtg. On the road to new campaign work #WeTheOC #ResearchRealness http://t.co/aNAy5xO5TU



University Archives: NYU Libraries

Created: Jan 27, 2016 by lrl

Updated: Sep 26, 2016 by system

Collection Data

Archived since Mar 31, 2016

80.6 MB
Archived

Total Data Archived

Data
80.6 MB

Documents
8,215

Collection Settings

Private

Active

Private Collection Link

https://archive-it.org/collections/e26035cf-f189-40db-b226-7ac743db1807

NLP

Named entity extraction

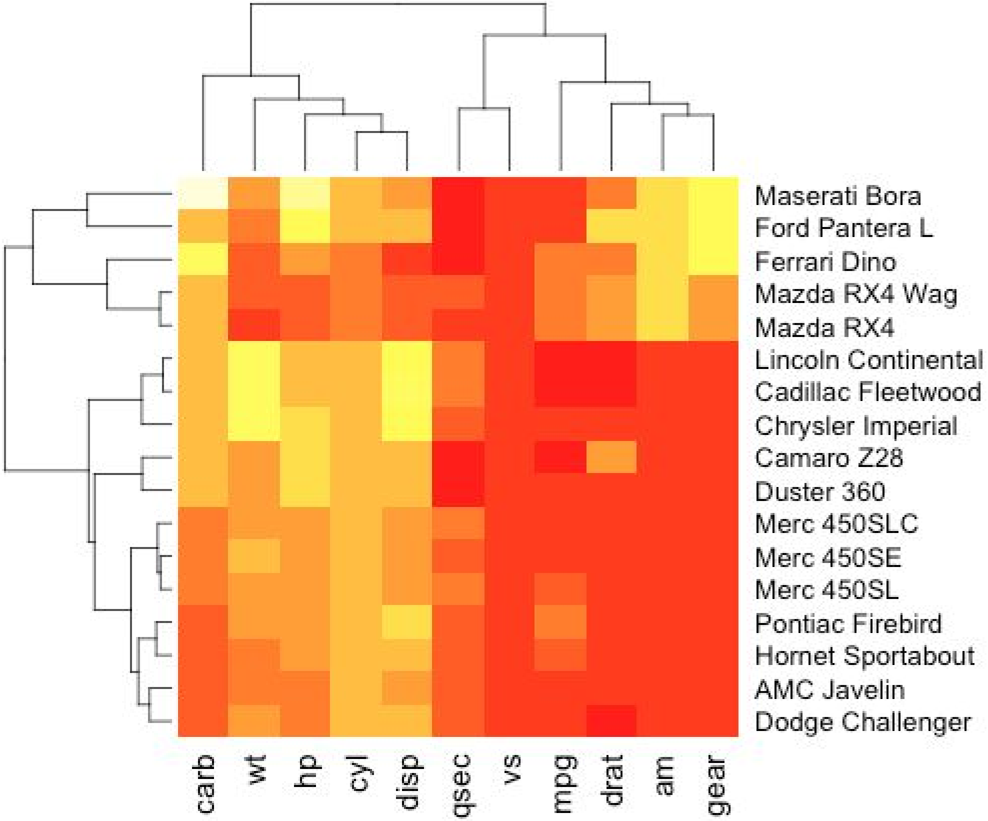
Topic modeling

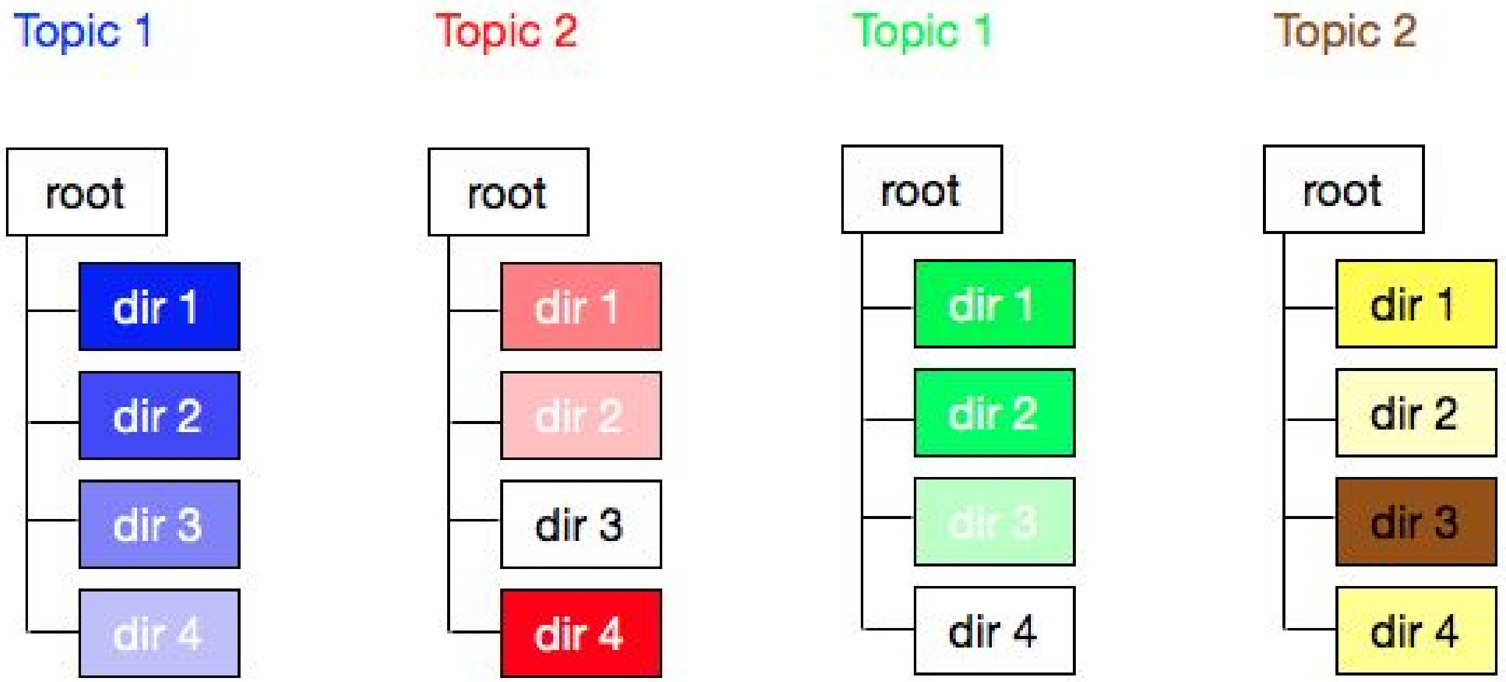
Clustering

Classification

Collaborative filtering

Language detection


Maserati Bora
Ford Pantera L
Ferrari Dino
Mazda RX4 Wag
Mazda RX4
Lincoln Continental
Cadillac Fleetwood
Chrysler Imperial
Camaro Z28
Duster 360
Merc 450SLC
Merc 450SE
Merc 450SL
Pontiac Firebird
Hornet Sportabout
AMC Javelin
Dodge Challenger
carb
wt
hp
cyl
disp
qsec
vs
mpg
drat
am
gear



Topic 1

root

dir 1

dir 2

dir 3

dir 4

Topic 2

root

dir 1

dir 2

dir 3

dir 4

Topic 1

root

dir 1

dir 2

dir 3

dir 4

Topic 2

root

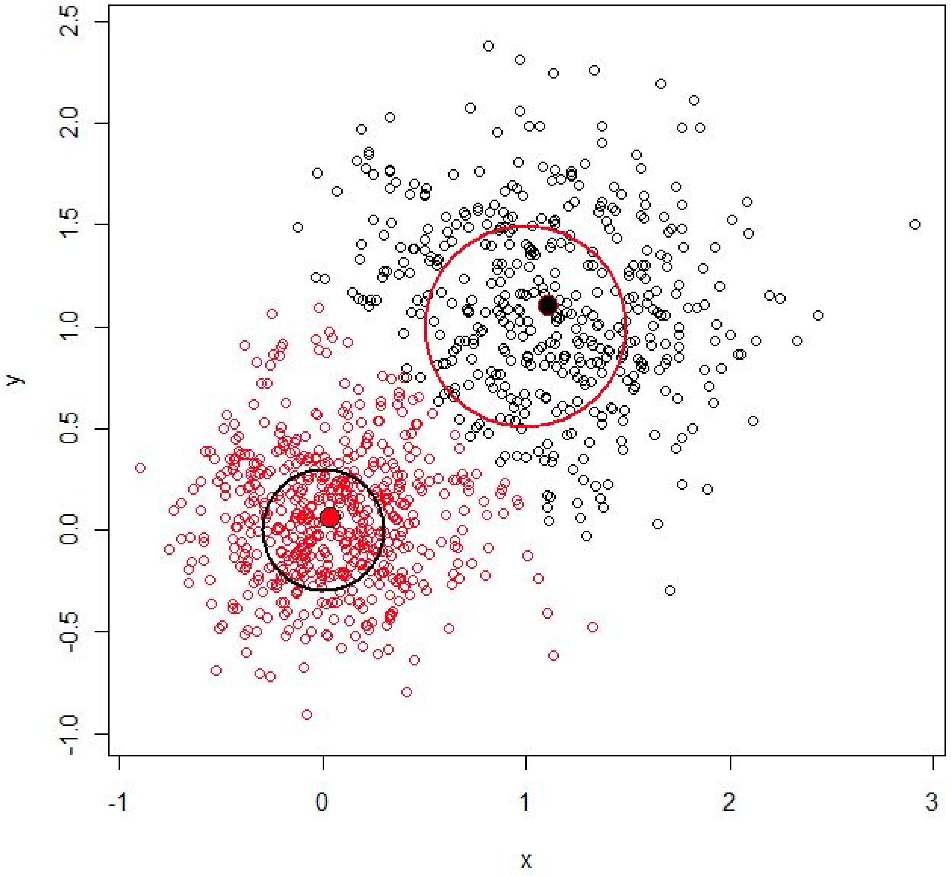
dir 1

dir 2

dir 3

dir 4

Lloyd k-means Clustering: iterations



Thanks.