How Web Archives and Tweets are Reshaping Today's Historical Record

University of Saskatchewan

11 March 2016

Ian Milligan
Assistant Professor
@ianmilligan1

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Historians are largely unprepared to engage with the quantity of digital sources that will fundamentally transform their trade.

... we need to think about big data ...

Today's Talk

- **1. Prologue:** Big Data is everywhere
- **2. The Web Age:** Will accelerate this process
- **3. What can we do with big data?**

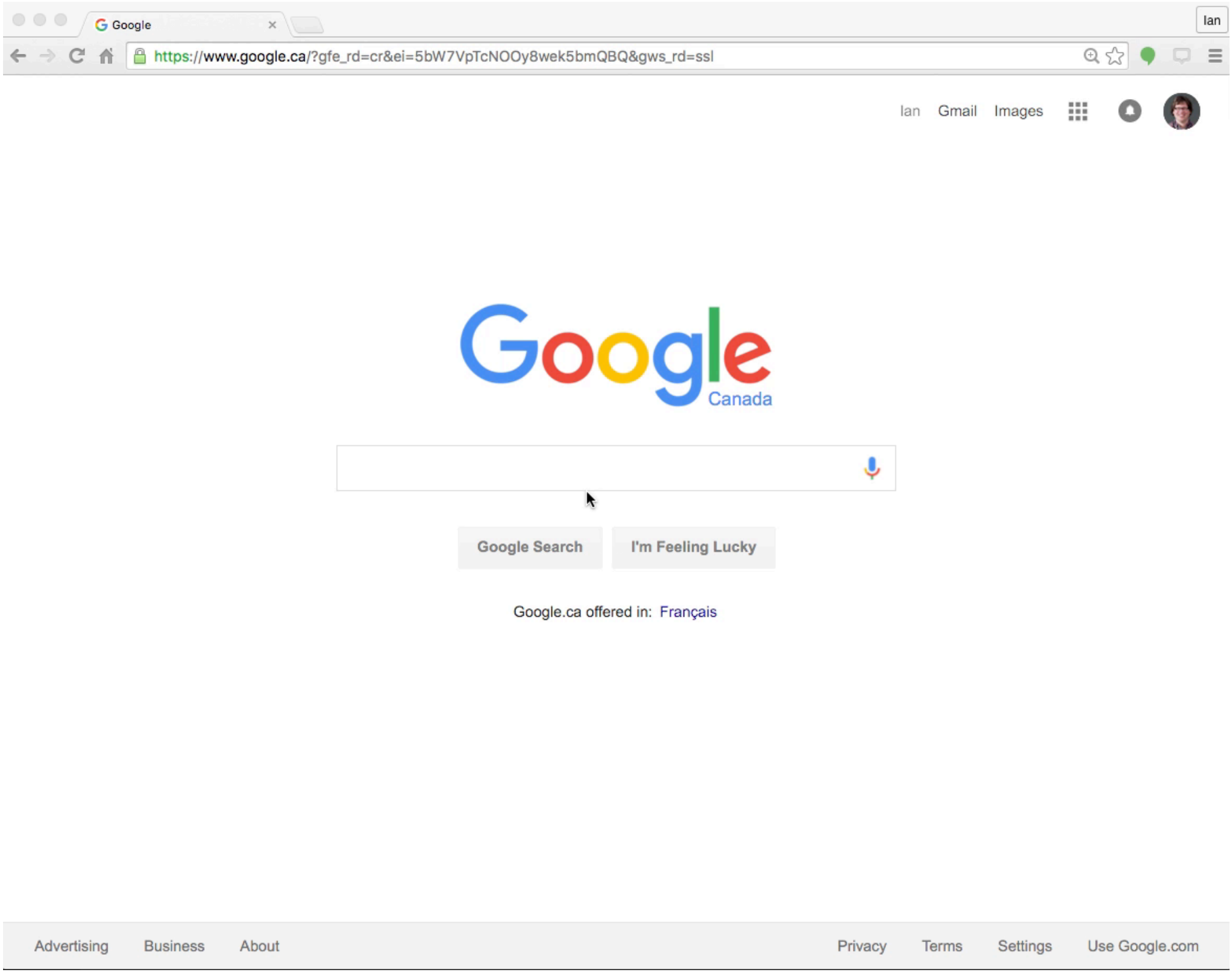
A Prologue: Big Data is Everywhere

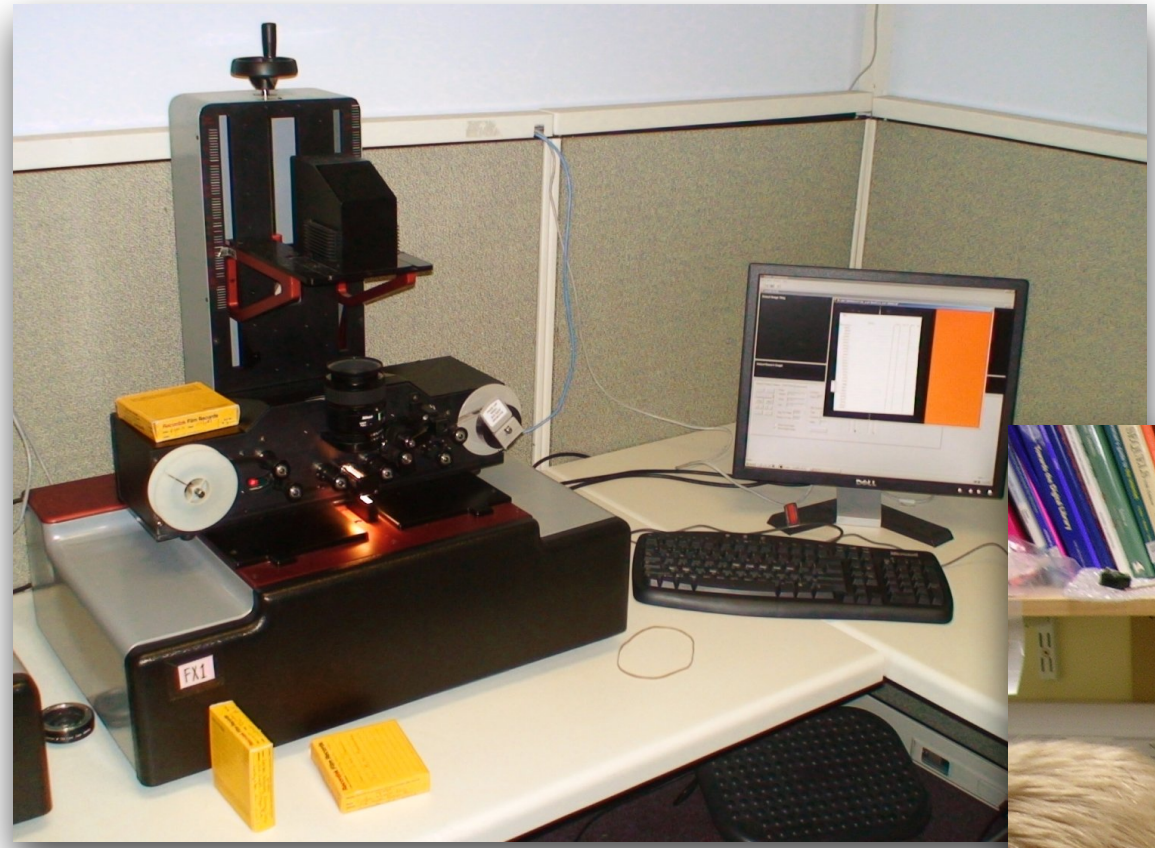
What do we mean by Big Data?

- Computational definition: the 5 Vs (Volume, Velocity, Variety, Veracity, and Value)

- "For us, as humanists, big is in the eye of the beholder. **If it's more data that you could conceivably read yourself in a reasonable amount of time, or that requires computational intervention to make new sense of it, it's big enough!**" (Shawn Graham, Ian Milligan, Scott Weingart, *Exploring Big Historical Data*)

Why is it everywhere?


Google
Ian
https://www.google.ca/?gfe_rd=cr&ei=5bW7VpTcNOOy8wek5bmQBQ&gws_rd=ssl
Ian
Gmail
Images
Google
Canada
Google Search
I'm Feeling Lucky
Google.ca offered in: Français
Advertising
Business
About
Privacy
Terms
Settings
Use Google.com


FX1


Canon
Microfilm Scanner 300 II
The Shetland Times
S. C. W. S.
FASHION GOODS
SALE

Advanced Search - ProQue

search.proquest.com/hnptorontostar/advanced/accountid=14906

Searching: 1 database | 0 Recent searches | 0 Selected items | My Research | Exit

« All databases | News & Newspapers databases | Preferences | English | Help

ProQuest Historical Newspapers: Toronto Star (1894-2011)

Basic Search | Advanced | Obituaries | Publications

Advanced Search

Look Up Citation | Command Line | Find Similar

Field codes | Search tips

in Anywhere

AND (OR) in Anywhere

AND (OR) in Anywhere

Add a row | Remove a row

Search Clear form

Search options

Publication date: All dates

Sort results by: Publication date (most recent first)

Items per page: 50

Duplicates: Include duplicate documents

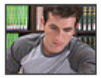
Search Clear form

Search subject areas

Use search forms customized for each subject.


The Arts


Business

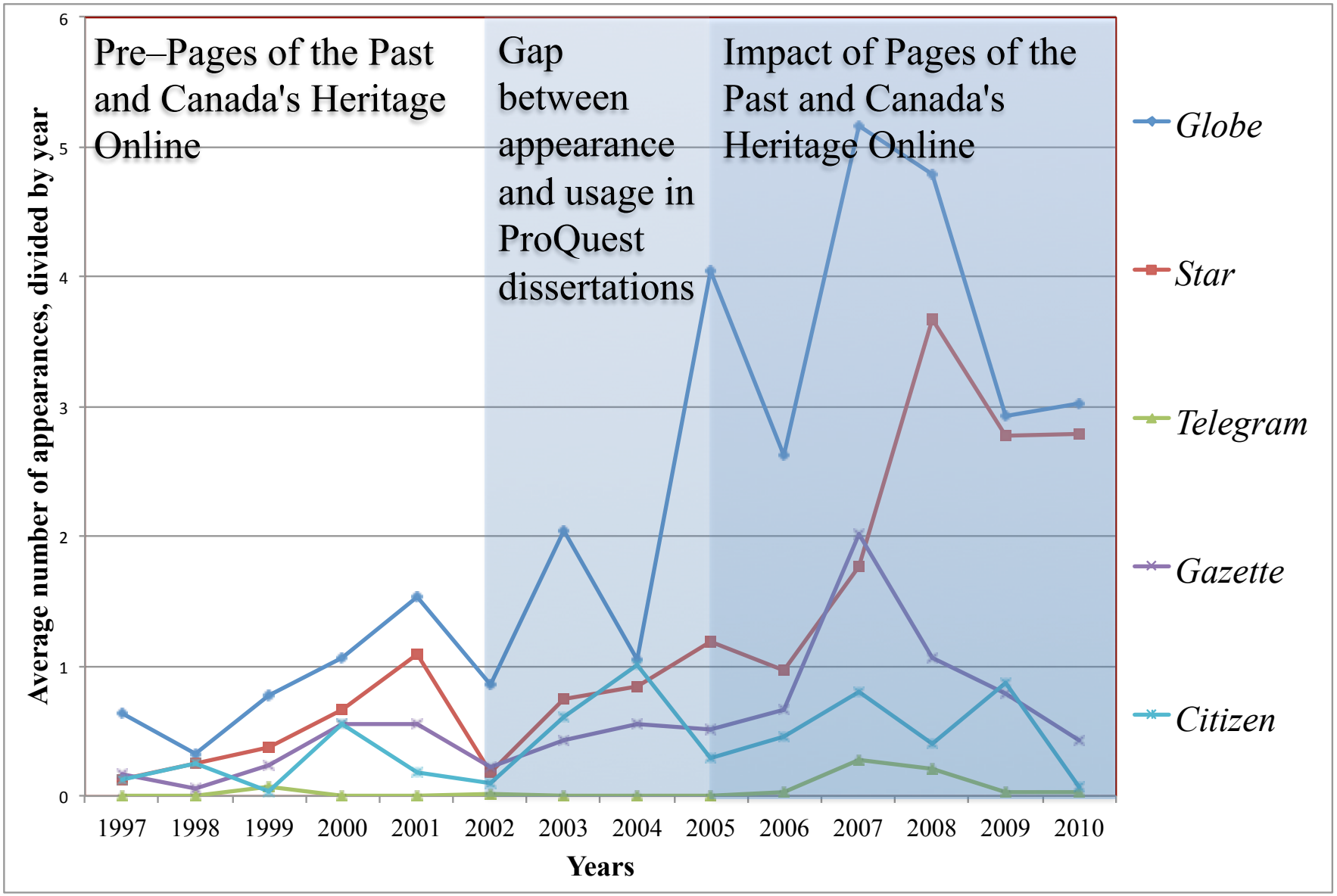

Dissertations & Theses


Health & Medicine


History


Literature & Language


News & Newspapers



Ian Milligan, "Online Databases, Optical Character Recognition, and Canadian History, 1997–2010," *Canadian Historical Review*, 94.4 (December 2013): 540-569.

... this is our long-term track record w/ digital resources ...

Our history with digital sources is the unreflective use of technology.

… we've become, in some ways, a discipline defined by the keyword …

A process that is only now beginning to accelerate.

First - more data than ever before being preserved;

Second - it'll be saved/delivered to us in very different ways


CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Subseries F Democratic Party
BOX 28
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
373
CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series V
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Subseries F Democratic Party
BOX 29
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
374
CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series V
PARTY LEADERSHIP / ADMINISTRATIVE FILE
Subseries F Democratic Party
BOX 30
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
375

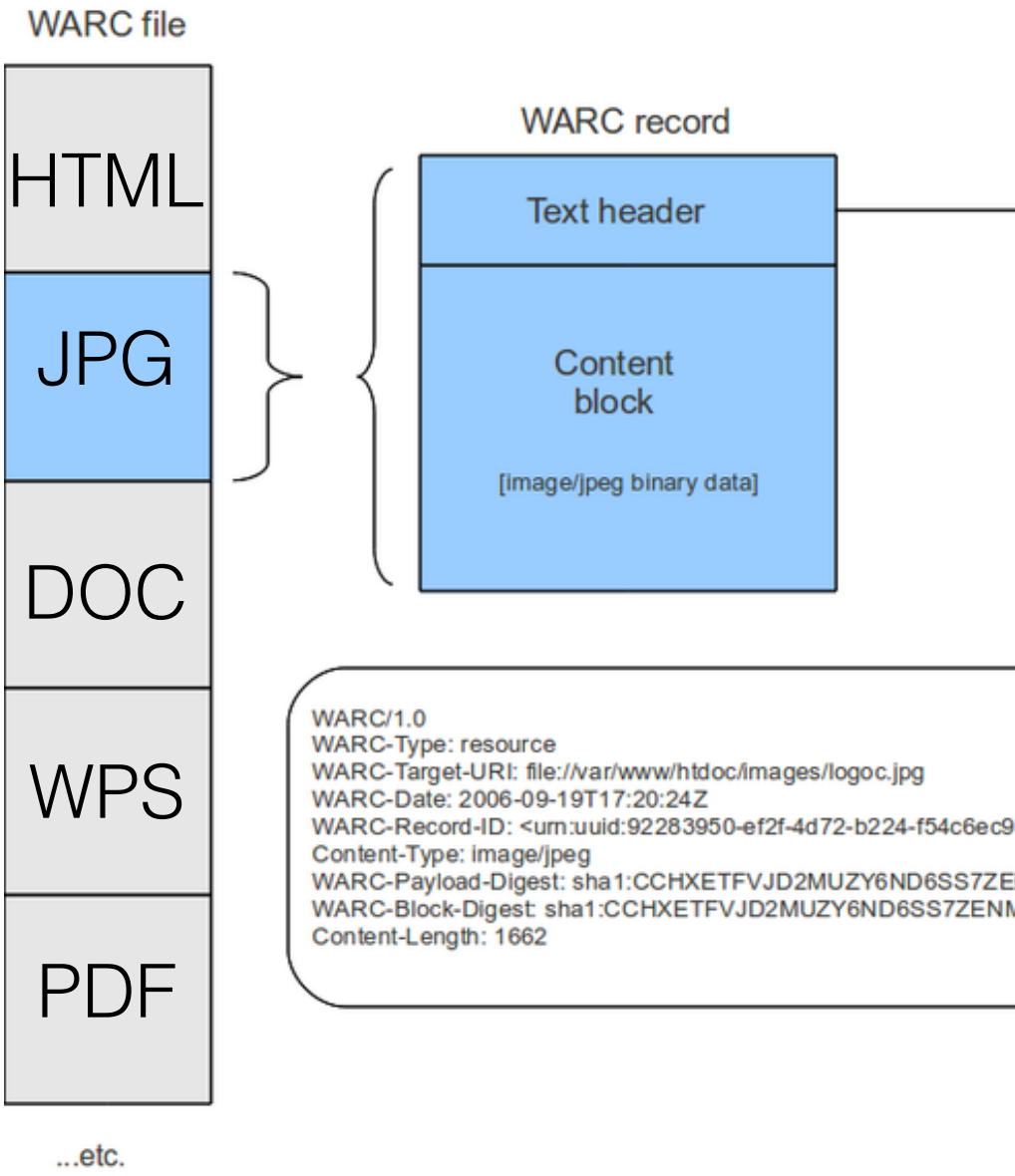

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
374
CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series V
PARTY LEADERSHIP / ADMINISTRATIVE FILE
Democratic Party
BOX 30
JOHN J. BURNS LIBRARY
BOSTON COLLEGE

Scarcity

WebARChive (WARC) File

????

- What does this really mean?



DEPARTMENT of HISTORY

9 Campus Drive
Saskatoon, SK S7N 5A5

UNIVERSITY OF SASKATCHEWAN

Last Update (*this page only*): Wednesday, 22-Jan-97 15:02:45 CST
[See Recent Changes to History Web Files]

- **Telephone: (306)966-5792**
- **Fax: (306)966-5852**
- **E-mail: hist.dept@usask.ca**

- **Department Head: *Bill Waiser***
- **Graduate Director: *Michael Hayden***
- **Undergraduate Director: *Michael Swan***

The History Department of the University of Saskatchewan in Saskatoon is one of the largest departments in the College of Arts and Science. Nevertheless, the three staff and twenty-one faculty members work to create an atmosphere in which students can find the information and advice they need about courses, careers, and the discipline of history.

SOME *HOT SPOTS* TO VISIT . . .

Canadian Journal of History	1912-Present Theses in History at UofS	H-Canada [English] [French]	UofS Library & Univ. Archives Essay Award	NEW Quiz Show	Other Web Sites
Canadian Historical Assoc. 1997	Now Then newsletter	What's UP?	Send a Postcard NEW	Amazon Book Search NEW	. . .

OccupyWallStreet

The revolution continues worldwide!

Farmers Join Occupy Wall Street, Calling for Food Justice

Posted 5 hours ago on Dec. 2, 2011, 6:21 p.m. EST by OccupyWallSt

As Wall Street's corrupt influence on the economy has grown, the corporate ownership of our food system has hurt the health and livelihood's of some of our most vulnerable communities. This *Sunday, December 4th* food justice activists and occupiers will be traveling from as far as Colorado, Iowa, Maine and Upstate New York to join together for the Occupy Wall Street FARMERS' MARCH.Through a day of dialogue, musical performances, and a march, farmers and their urban allies working for food justice in their communities will form alliances to fight and expose corporate control of the food supply.

Events throughout the day will call and inspire participants to fight against the corporate manipulation of the agriculture system. An industry that is responsible for using chemical toxins tied to soaring obesity rates, heart disease and diabetes and limiting access to affordable, wholesome food to the country's poorest citizens.

Read More...

30 Comments

Occupy Wall Street Goes Home

Posted 1 day ago on Dec. 1, 2011, 3:04 p.m. EST by OccupyWallSt

On December 6th Occupy Wall Street will join in solidarity with a Brooklyn community to re-occupy a foreclosed home. The day of action marks a national kick-off for a new frontier for the

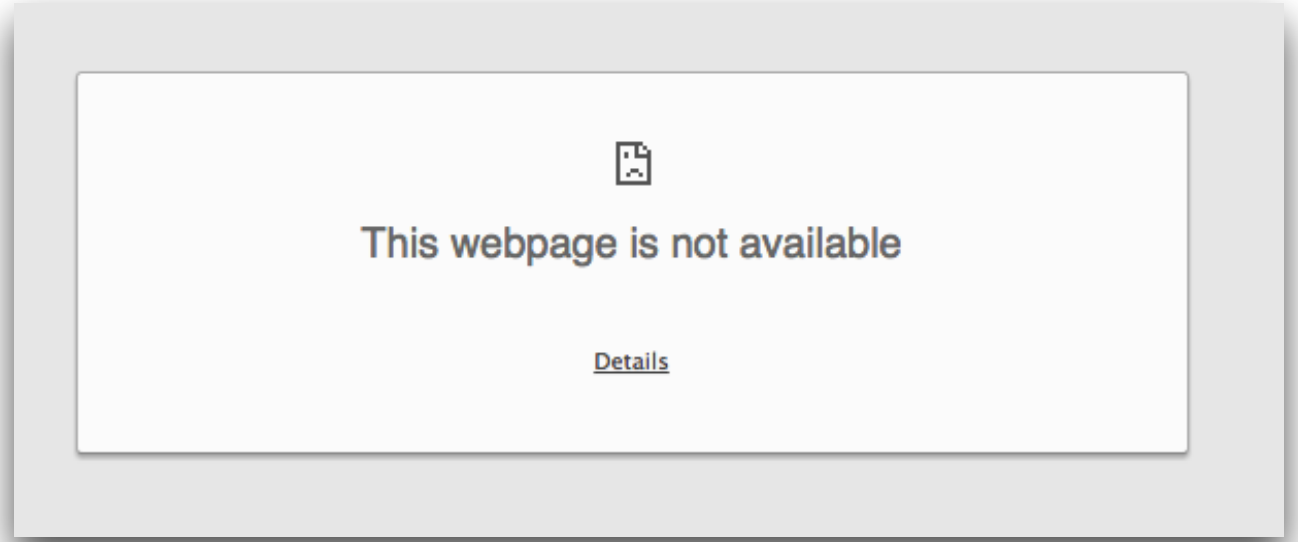
General Inquiries:
general@occupywallst.org
Press Inquiries:
press@occupywallst.org
Press Phone: +1 (347) 292-1444
Help & Directions: +1 (516) 708-4777
Watch: The world we're building
Read: This call to action
Liberty Square Eviction Defense: Text "@occupyalert" to 23559 to receive alerts in the event of imminent emergency.

Occupy Wall Street is leaderless resistance movement with people of many colors, genders and political persuasions. The one thing we all have in common is that We Are The 99% that will no longer tolerate the greed and corruption of the 1%. We are using the revolutionary Arab Spring tactic to achieve our ends and encourage the use of nonviolence to maximize the safety of all participants.

This #ows movement empowers real people to create real change from the bottom up. We want to see a general assembly in every backyard, on every street corner because we don't need Wall Street and we don't need politicians to build a better society.

the only solution is WorldRevolution

Click here for NYCGA committee meeting times.


This webpage is not available
Details

OccupyWallStreet

The revolution continues worldwide!

Farmers Join Occupy Wall Street, Calling for Food Justice

Posted 5 hours ago on Dec. 2, 2011, 6:21 p.m. EST by OccupyWallSt

As Wall Street's corrupt influence on the economy has grown, the corporate ownership of our food system has hurt the health and livelihood's of some of our most vulnerable communities. This *Sunday, December 4th* food justice activists and occupiers will be traveling from as far as Colorado, Iowa, Maine and Upstate New York to join together for the Occupy Wall Street FARMERS' MARCH.Through a day of dialogue, musical performances, and a march, farmers and their urban allies working for food justice in their communities will form alliances to fight and expose corporate control of the food supply.

Events throughout the day will call and inspire participants to fight against the corporate manipulation of the agriculture system. An industry that is responsible for using chemical toxins tied to soaring obesity rates, heart disease and diabetes and limiting access to affordable, wholesome food to the country's poorest citizens.

Read More...

30 Comments

Occupy Wall Street Goes Home

Posted 1 day ago on Dec. 1, 2011, 3:04 p.m. EST by OccupyWallSt

On December 6th Occupy Wall Street will join in solidarity with a Brooklyn community to re-occupy a foreclosed home. The day of action marks a national kick-off for a new frontier for the

General Inquiries:
general@occupywallst.org
Press Inquiries:
press@occupywallst.org
Press Phone: +1 (347) 292-1444
Help & Directions: +1 (516) 708-4777
Watch: The world we're building
Read: This call to action
Liberty Square Eviction Defense: Text "@occupyalert" to 23559 to receive alerts in the event of imminent emergency.

Occupy Wall Street is leaderless resistance movement with people of many colors, genders and political persuasions. The one thing we all have in common is that We Are The 99% that will no longer tolerate the greed and corruption of the 1%. We are using the revolutionary Arab Spring tactic to achieve our ends and encourage the use of nonviolence to maximize the safety of all participants.

This #ows movement empowers real people to create real change from the bottom up. We want to see a general assembly in every backyard, on every street corner because we don't need Wall Street and we don't need politicians to build a better society.

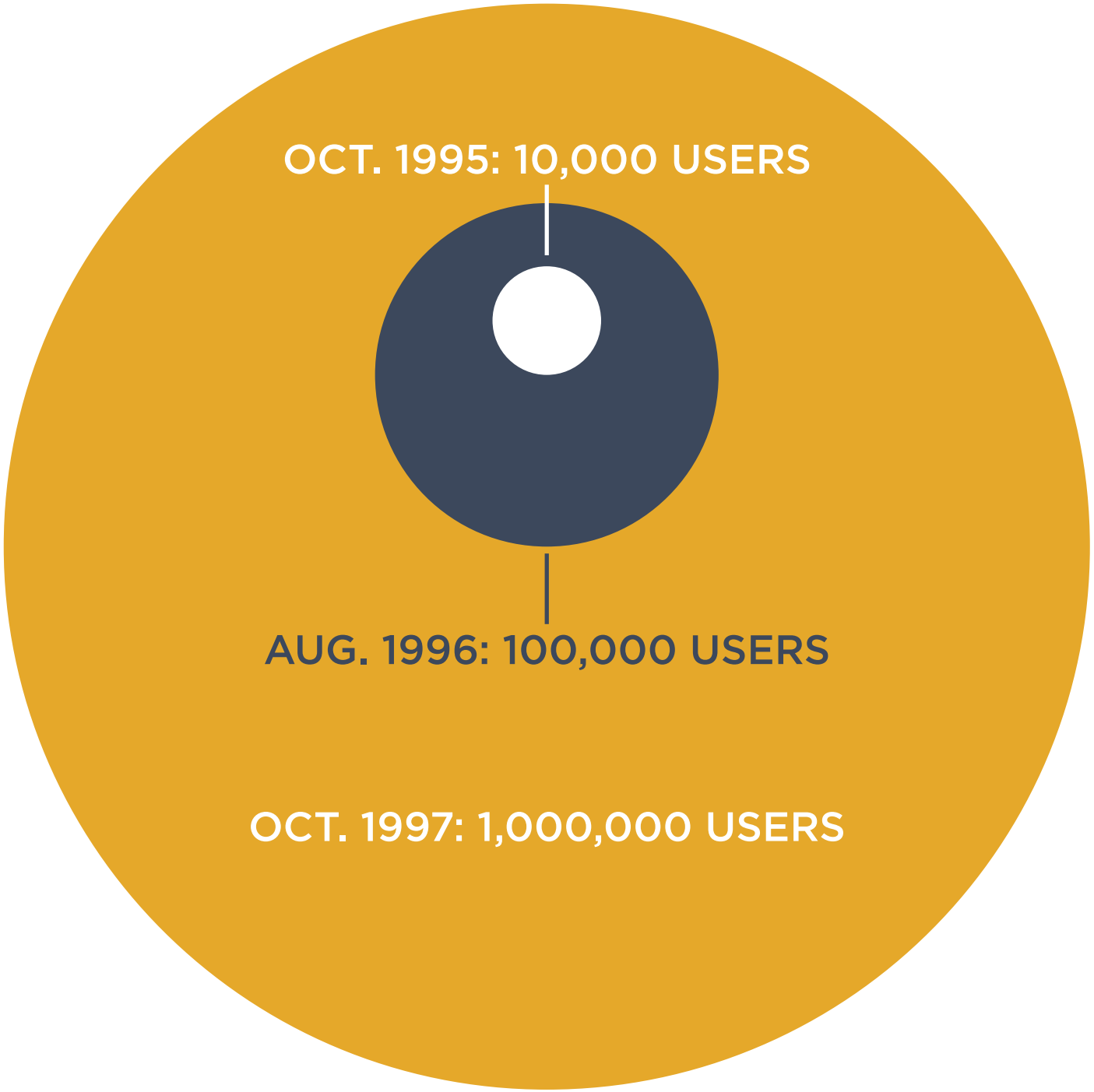
the only solution is WorldRevolution

Click here for NYCGA committee meeting times.

~~Scarcity~~

Abundance

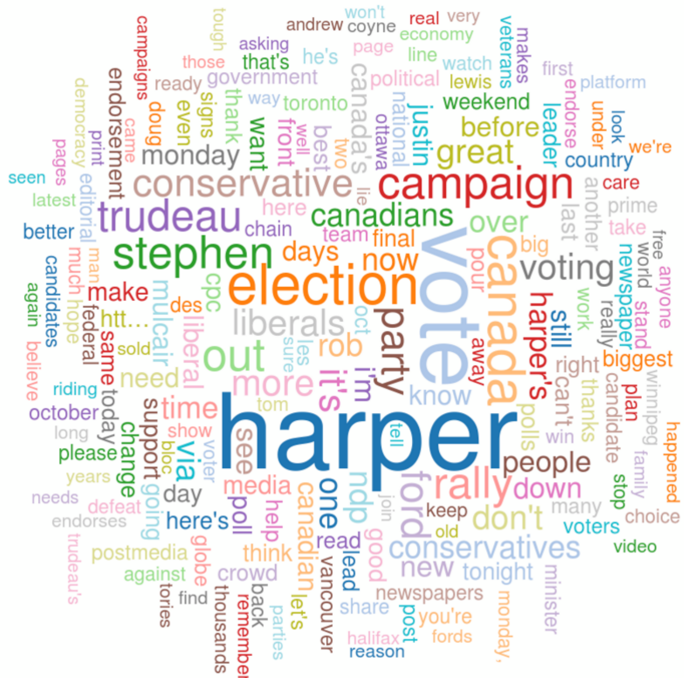
GEOCITIES USERS:



This is a scale that **boggles the mind** - compare it to the Old Bailey (197,745 trials between 1674 and 1913)

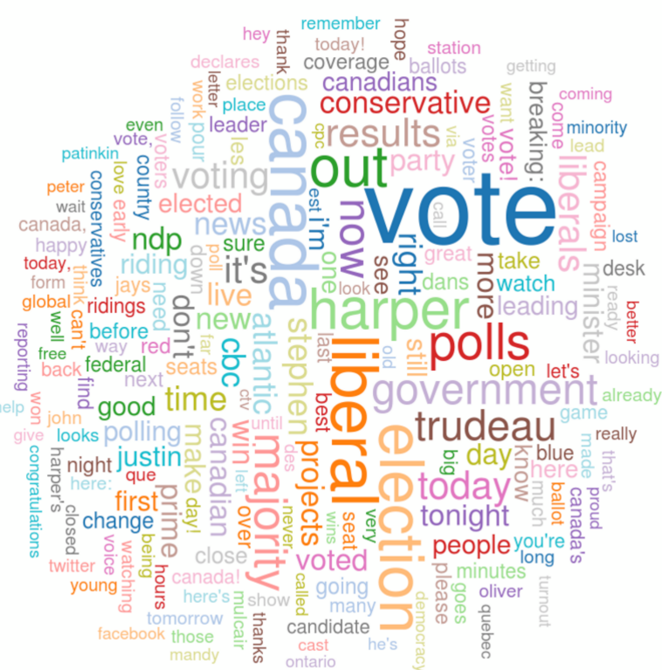

RIC'S GRILL
STEAK SEAFOOD & CHOP HOUSE
Wharf
IDLE NO MORE

October 17, 2015



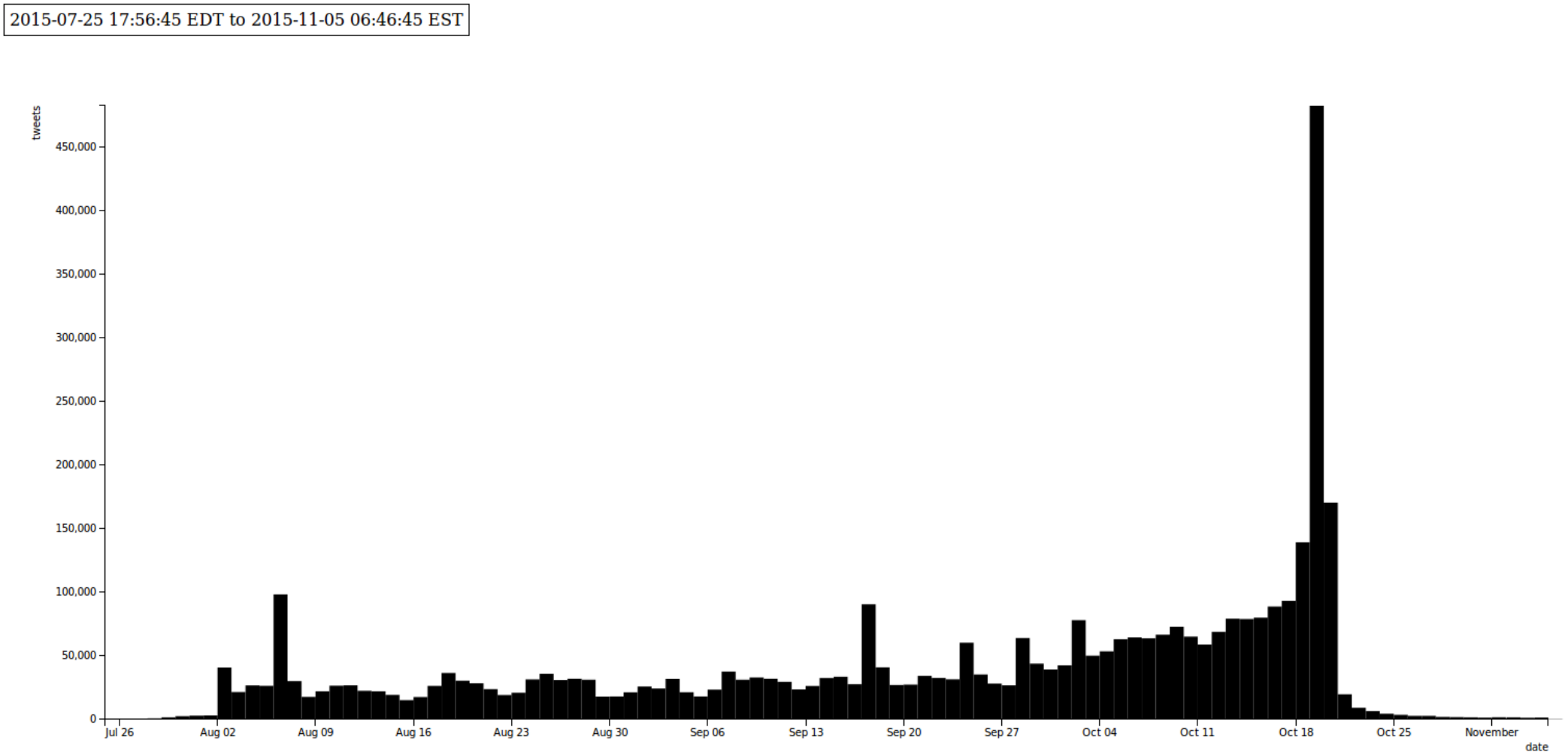
October 18, 2015

October 19, 2015



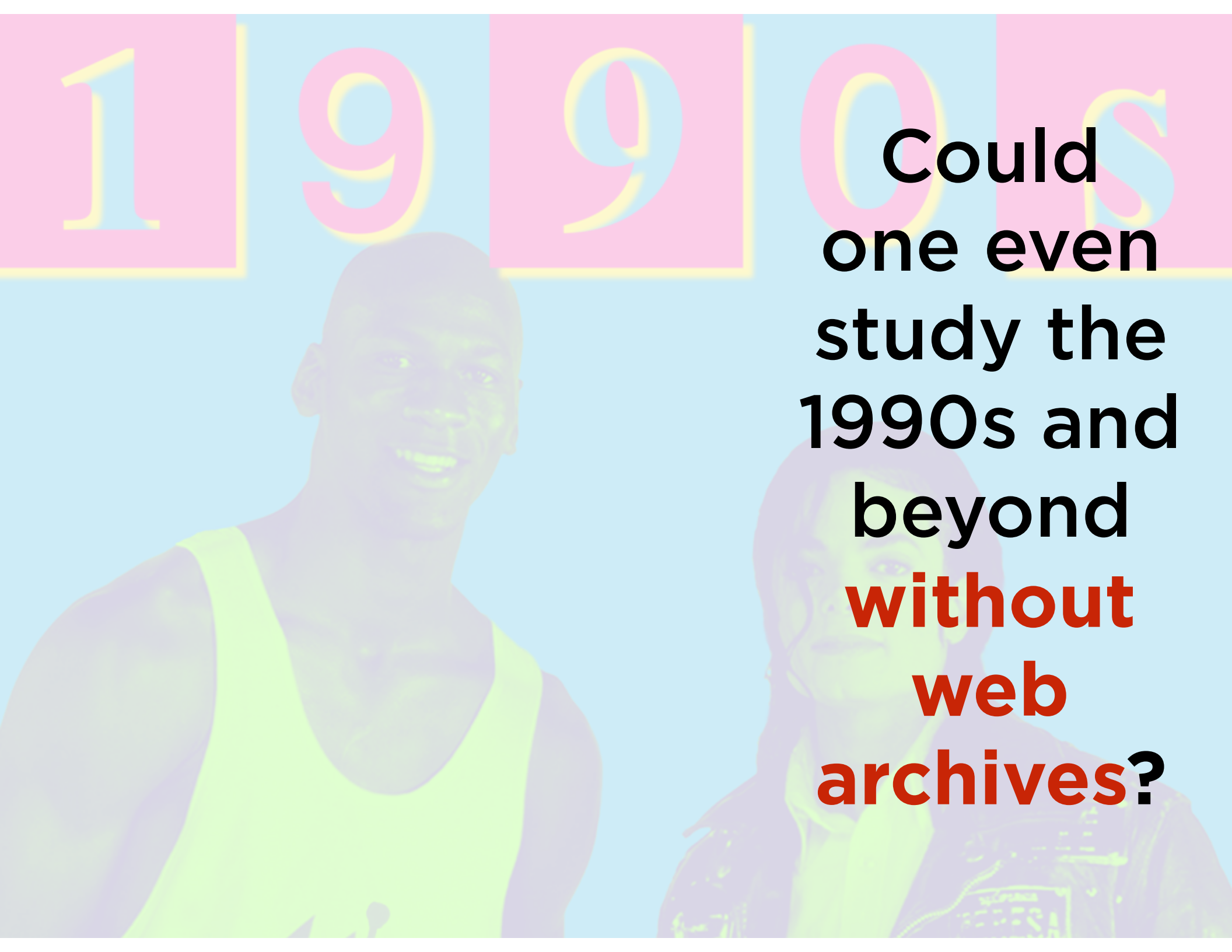
October 20, 2015

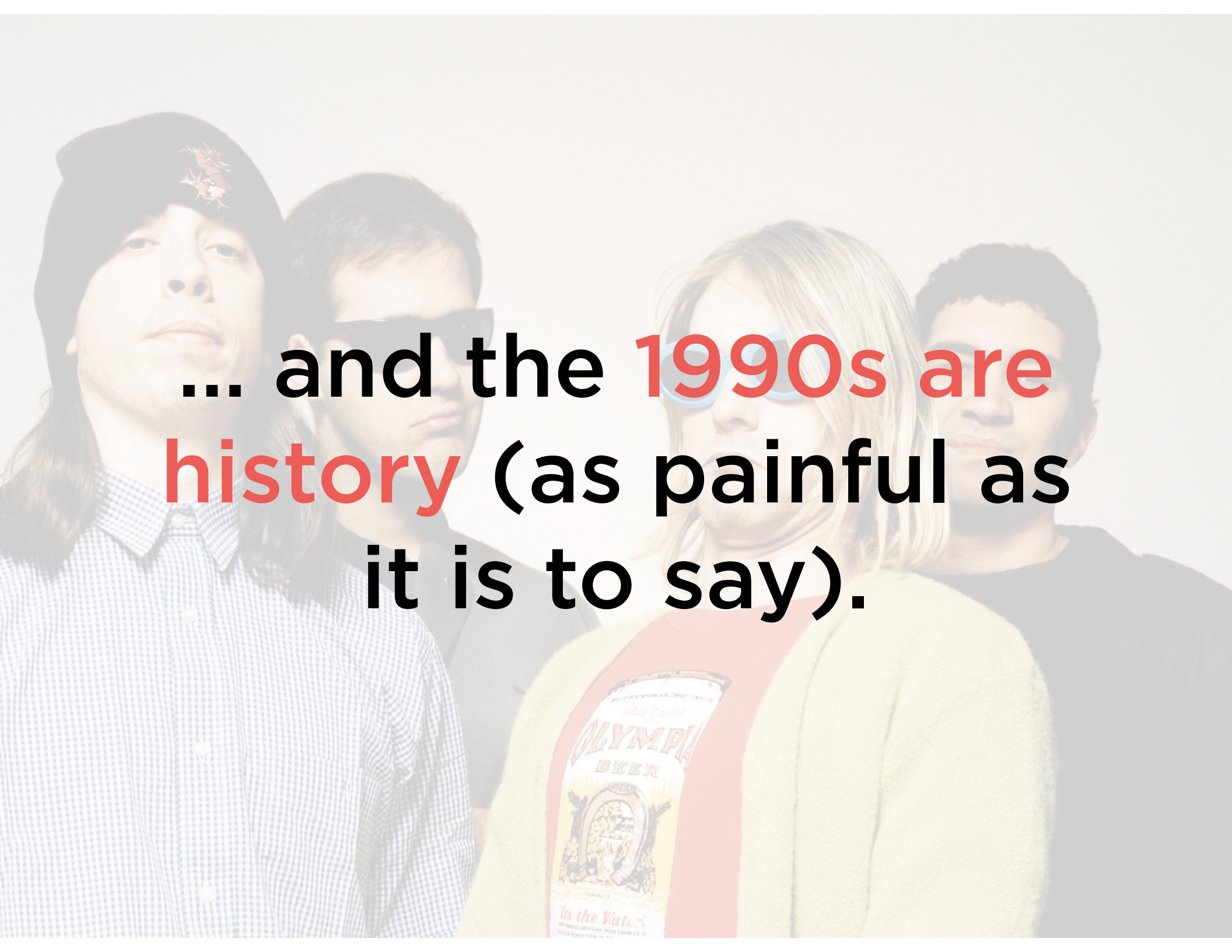
Frequency on the #ELXN42 Hashtag



“…. [n]ow expectations have inverted. Everything may be recorded and preserved, at least potentially.”

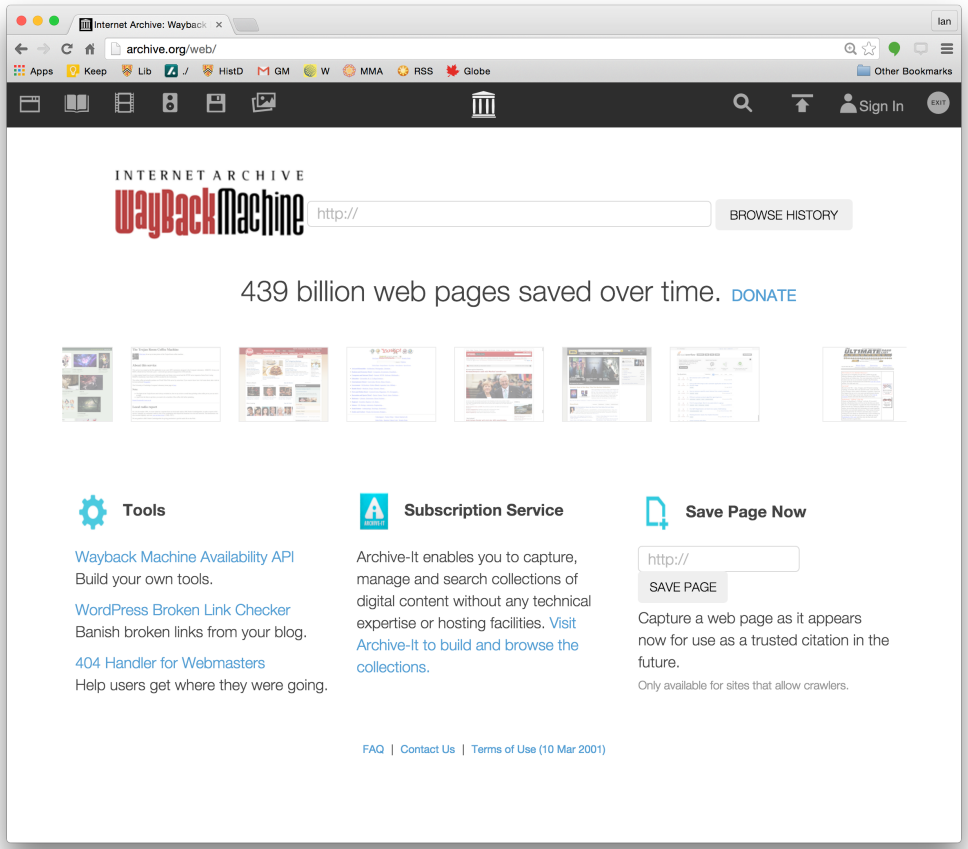
- James Gleick


1990s
Could one even study the 1990s and beyond without web archives?



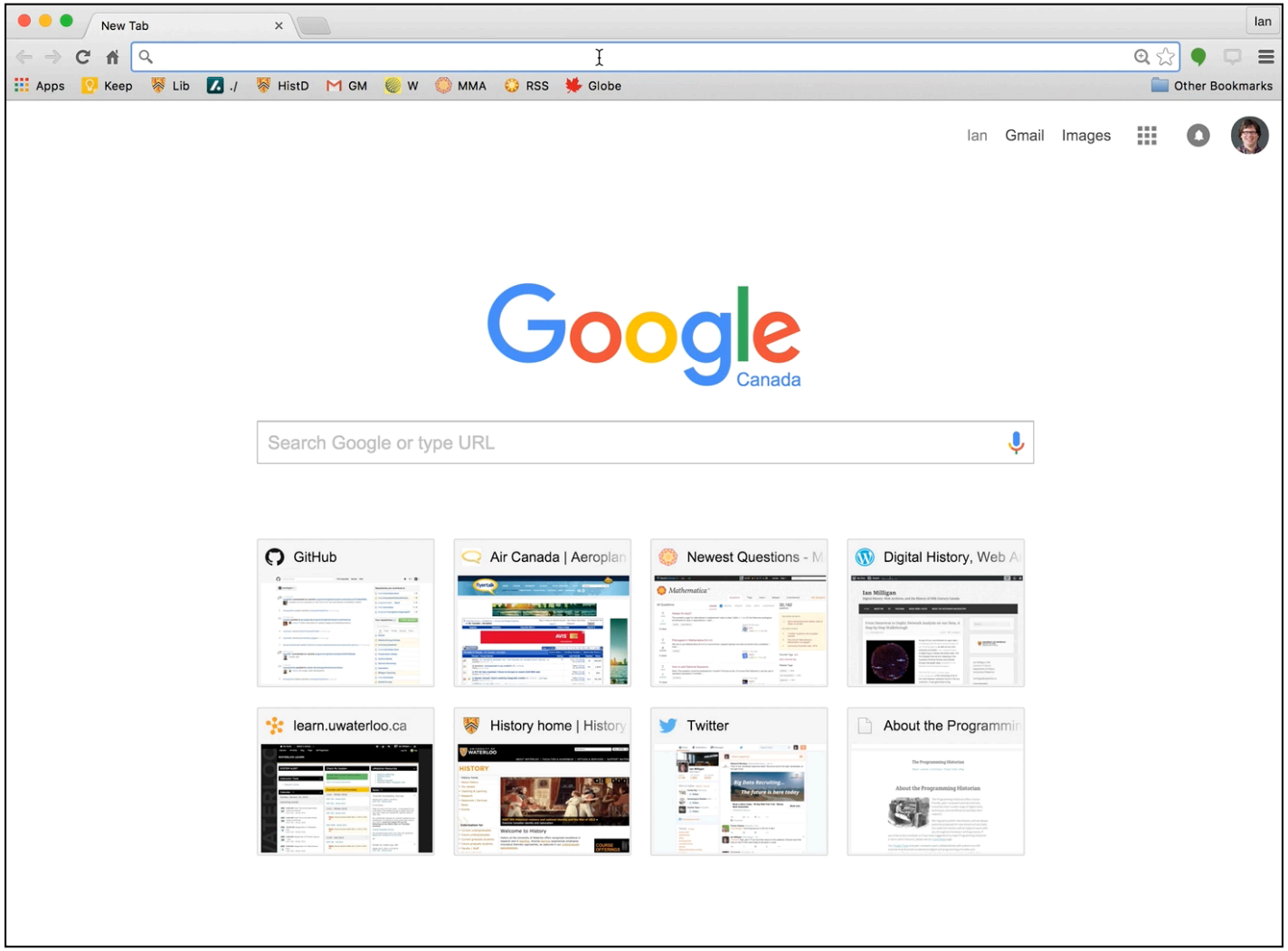
... and the 1990s are history (as painful as it is to say).

Nightmare Scenario



This won't be enough!

Nightmare Scenario

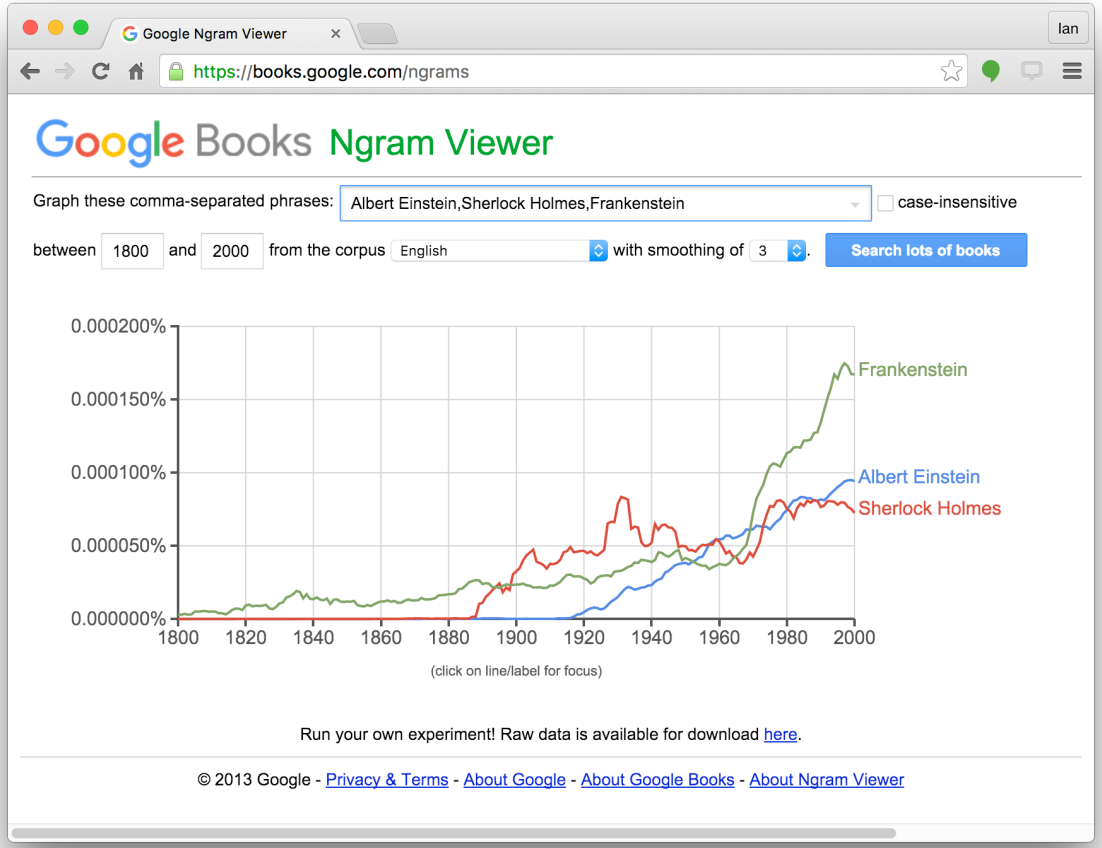


This won't be enough!

… but what will our search engines look like?

Nightmare Scenario

- Historians rely uncritically on **date-ordered keyword search results**, putting them at mercy of search algorithms they do not understand (similar to digitized newspapers);



My deepest fear:

Historians are completely left out of post-1996 research, letting everybody else do the work (a la Culturomics project/ Science magazine article);
Our profession gets left behind...

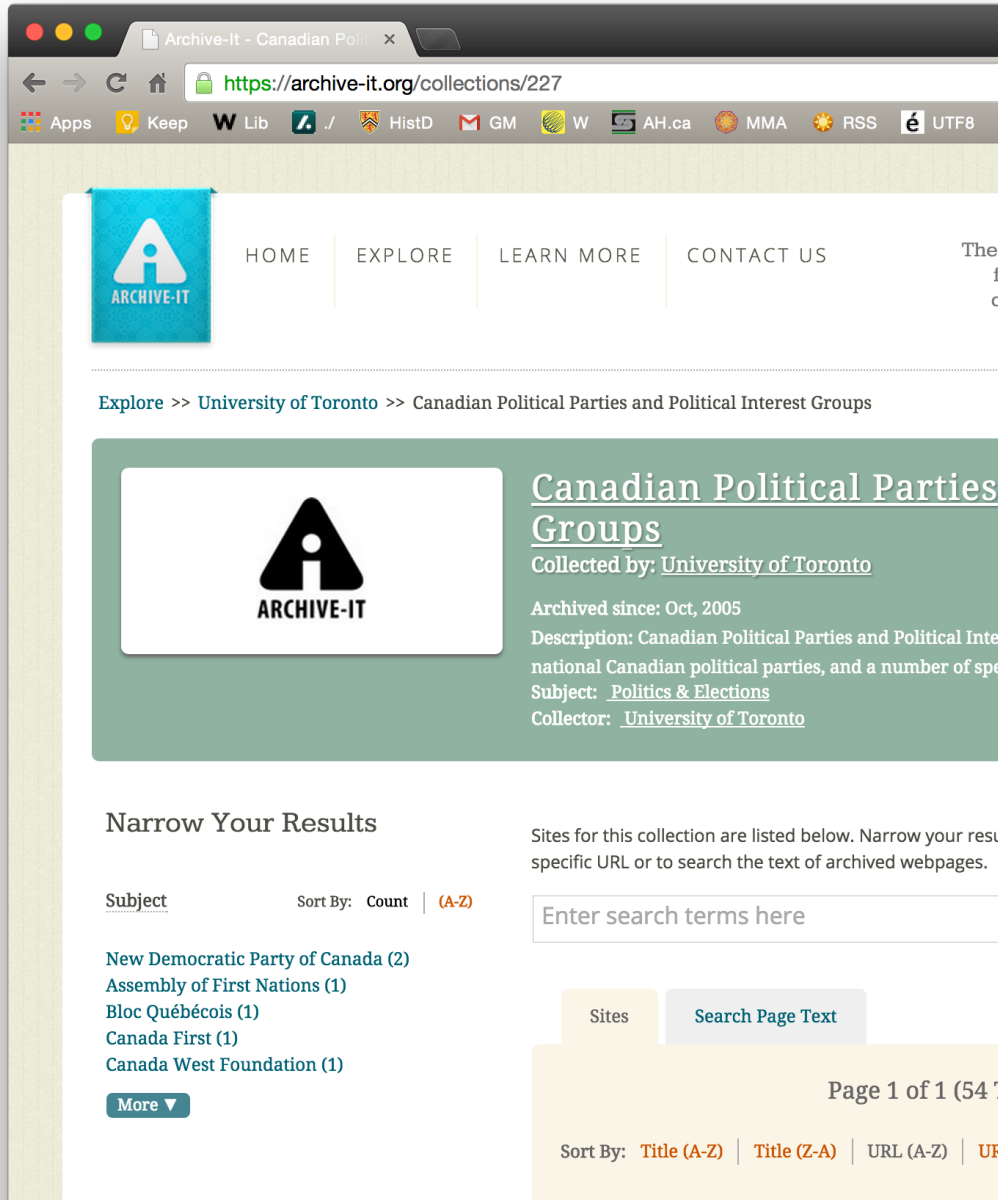
The historians who came to the meeting were intelligent, kind, and encouraging. But they didn't seem to have a good sense of how to wield quantitative data to answer questions, didn't have relevant computational skills, and didn't seem to have the time to dedicate to a big multiauthor collaboration. It's not their fault: these things don't appear to be taught or encouraged in history departments right now.

- Erez Leiberman Aiden and Jean-Baptiste Michel

What can we do to access this information and avoid my nightmare?

Building Portals

- Democratizing access so that historians can use them.
- Building **transparent indexes**.
- But they have to be useful and tested...



Pivotal Changes in Canadian Politics, 2005-2015

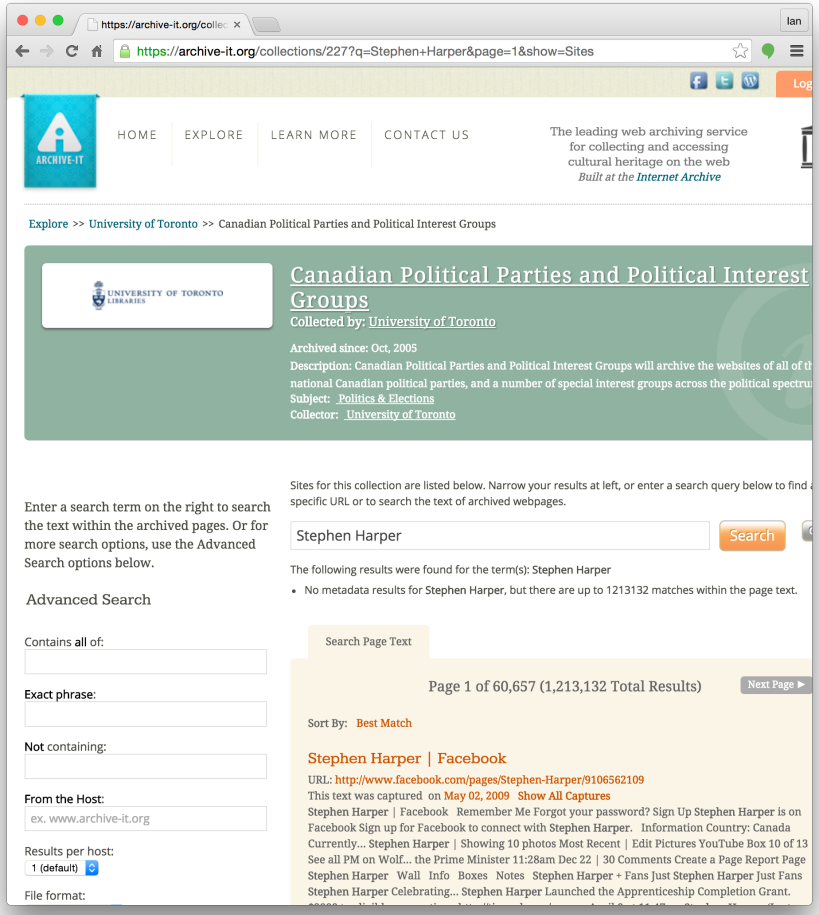
- Militarization of Canadian society?
- Change from ‘natural governing party’ of Liberals to Conservatives
- Major policy changes on foreign policy, environment, etc.
- Sweeping change to how we understand our history
- How to measure?

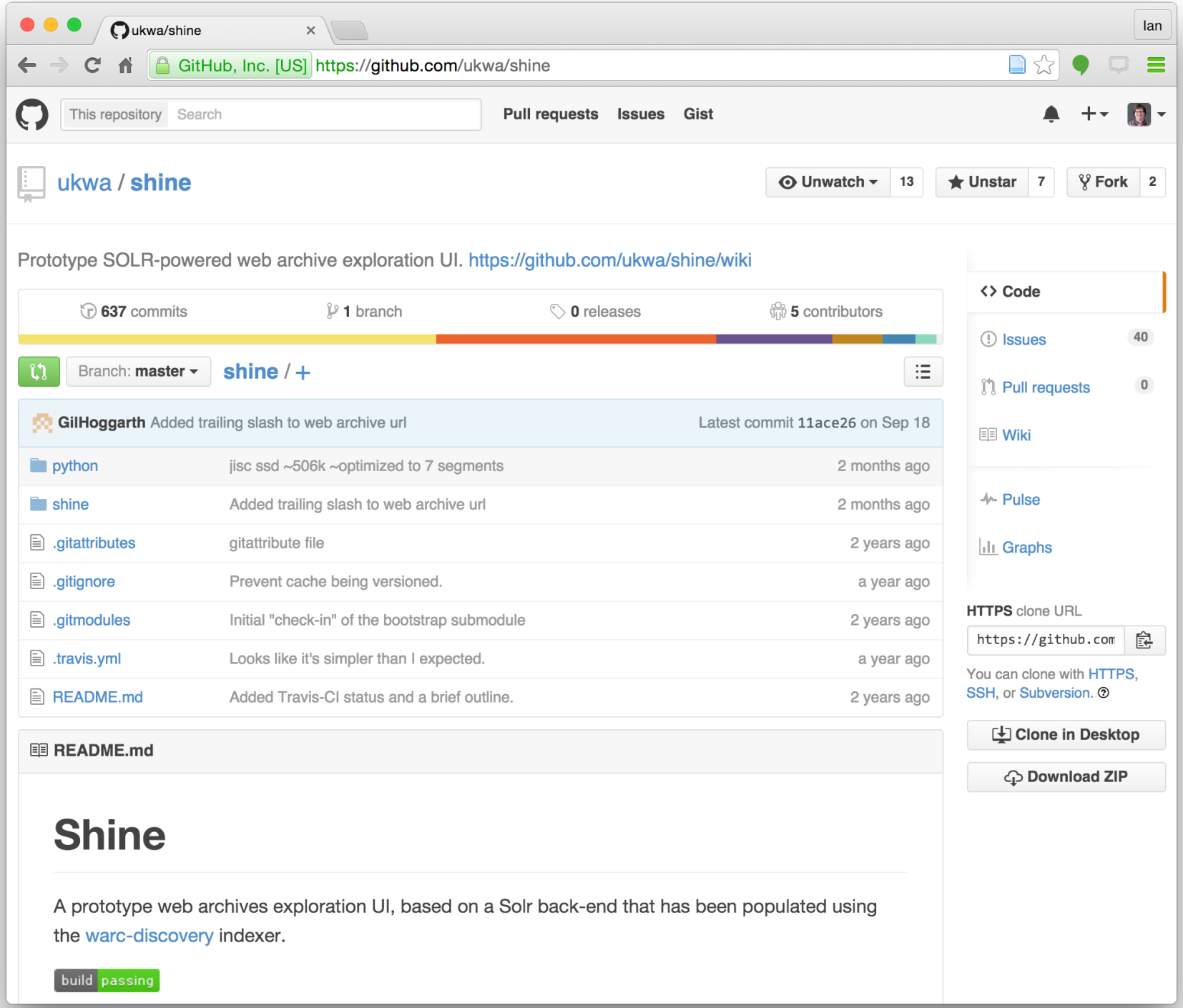
Canadian Political Parties & Political Interest Group Collection

- 50 Websites
  - All major political parties
  - Minor political parties
  - Political interest groups
- Collected quarterly between 2005 & present.

Current Interface

- Very limited - simple search engine, some advanced options; no facets
- Great collections.. but nobody uses them!




ukwa/shine
Ian
GitHub, Inc. [US] https://github.com/ukwa/shine
This repository Search
Pull requests Issues Gist
ukwa / shine
Unwatch 13 Unstar 7 Fork 2
Prototype SOLR-powered web archive exploration UI. https://github.com/ukwa/shine/wiki
637 commits 1 branch 0 releases 5 contributors
Branch: master shine / +
GilHoggarth Added trailing slash to web archive url
Latest commit 11ace26 on Sep 18
python jisc ssd ~506k ~optimized to 7 segments 2 months ago
shine Added trailing slash to web archive url 2 months ago
.gitattributes gitattribute file 2 years ago
.gitignore Prevent cache being versioned. a year ago
.gitmodules Initial "check-in" of the bootstrap submodule 2 years ago
.travis.yml Looks like it's simpler than I expected. a year ago
README.md Added Travis-CI status and a brief outline. 2 years ago
README.md
Shine
A prototype web archives exploration UI, based on a Solr back-end that has been populated using the warc-discovery indexer.
build passing
Code
Issues 40
Pull requests 0
Wiki
Pulse
Graphs
HTTPS clone URL
https://github.com
You can clone with HTTPS, SSH, or Subversion.
Clone in Desktop
Download ZIP

Great research question that our contemporary historians were studying (Canada changing)

+

Great collection (all the political parties + many interest groups)

+

Ope Source Software

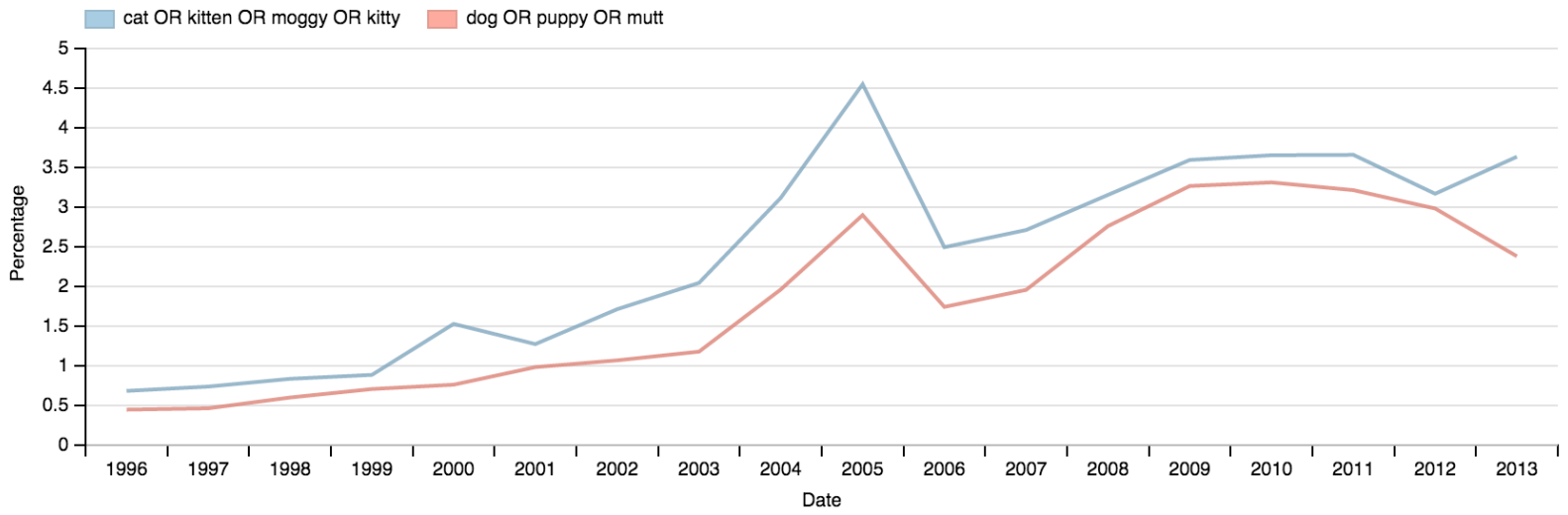
Trend results 1996-2013 fo

https://www.webarchive.org.uk/shine/graph?query=cat+OR+kitten+OR+moggy+OR+kitty%2C+dog+OR+pu...

UK Web Archive | Search | Trends | Login

Warning! This is a research prototype for a web archive search service, and may be taken down at any time.

cat OR kitten OR moggy OR kitty, dog OR | 1996 | 2013 | Update Graph | Reset All



Click on a point on the graph to show a random sample of the matching records from that year...

With Nick Ruest (York University), Nich Worby (University of Toronto), Jimmy Lin (University of Waterloo)

Welcome to the Web Archives for Historical Research political parties portal. Before diving in, we encourage you to visit our about page. ×

The Canadian Political Parties and Political Interest Groups Portal

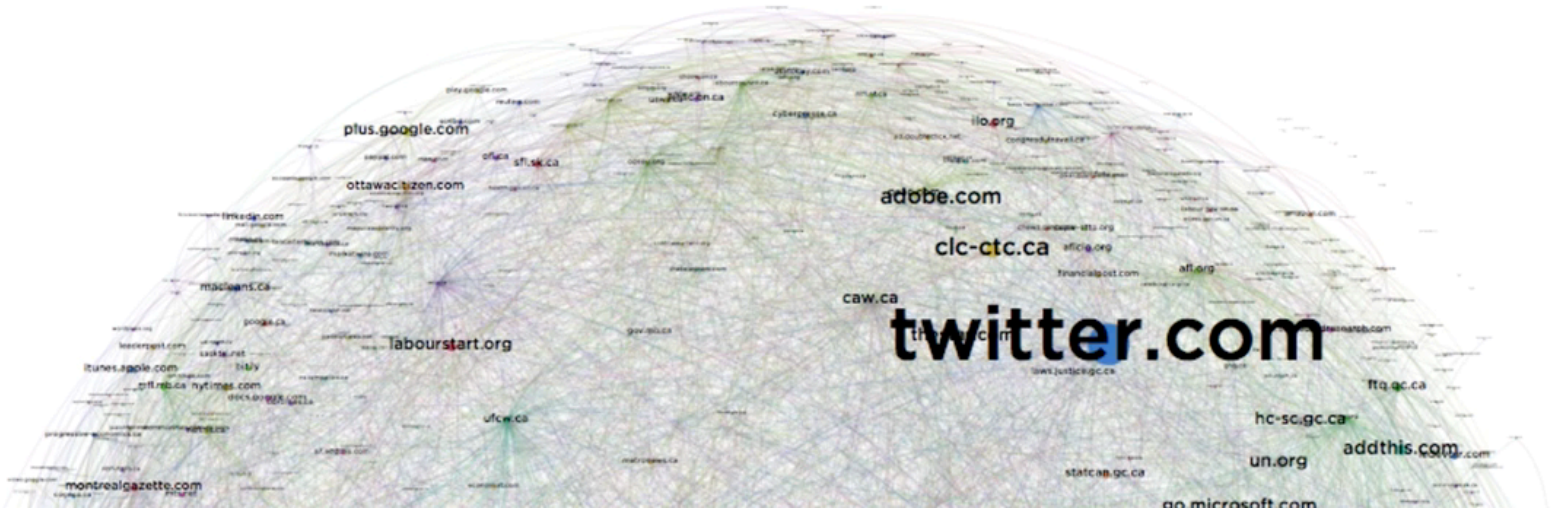
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

Curious how the Liberal Party of Canada responded to the 2008 financial crisis (a search for "recession" in 2008, liberal.ca)? How the Canadian Centre for Policy Alternatives reacted to Michael Ignatieff? Now you can check it all out.

Options include:

- **Basic keyword searching** [Example: "Rob Ford", only Liberal.ca]
- **Graphing trends over time** [Example: Liberal Opposition Leaders, 2005-2015]
- **Advanced search, including words in proximity to each other** [Example: environmental and tax within 25 words of each other]

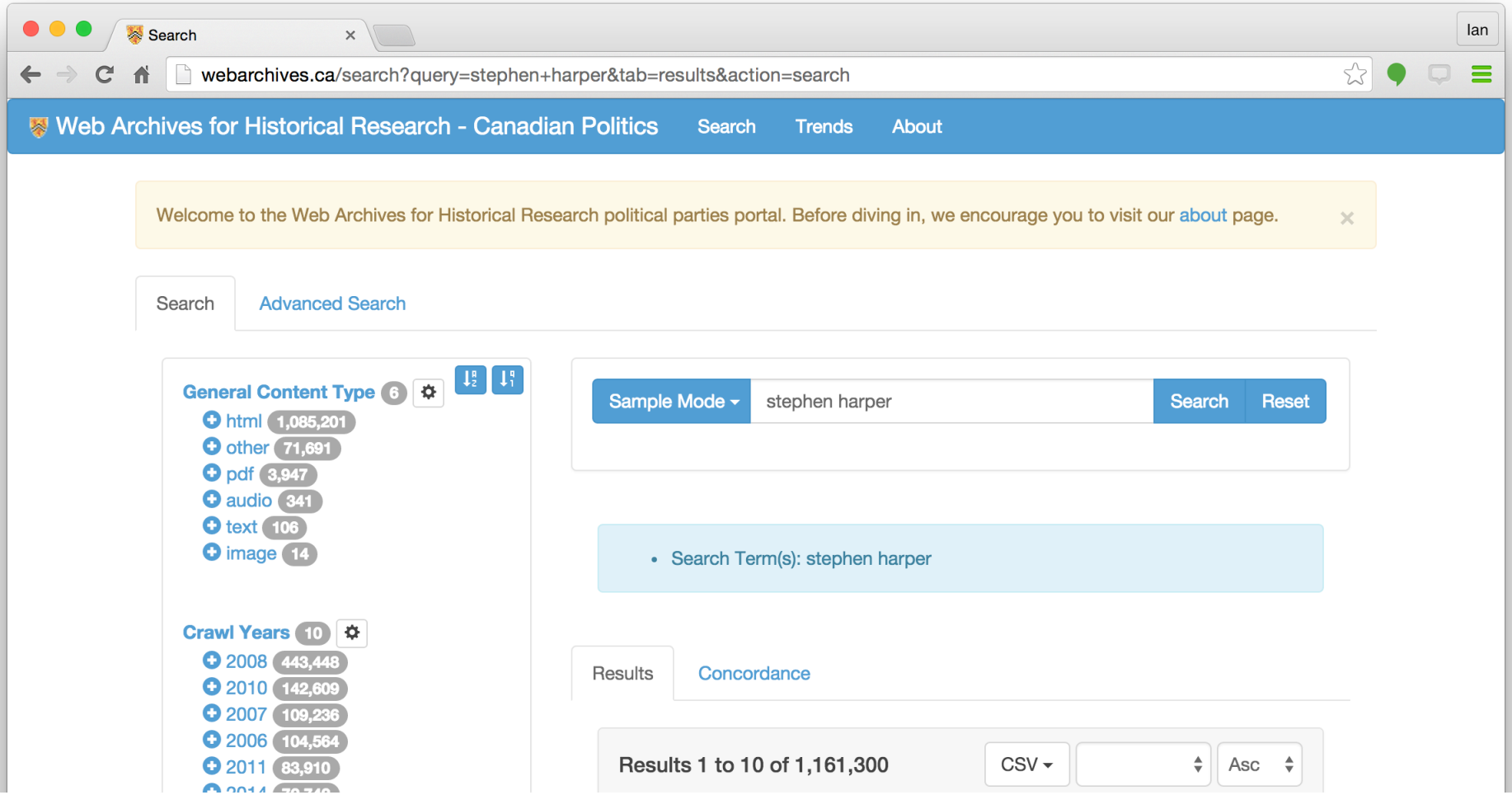
Below, here are all of the links for the entire time period, visualized below.



Five Things We Learned

- Political parties delete content
- User-generated comments were more common in political parties
- Absences can be more informative than presences
- We can see the rise/fall of prominent people
- Enabling user access is truly transformative

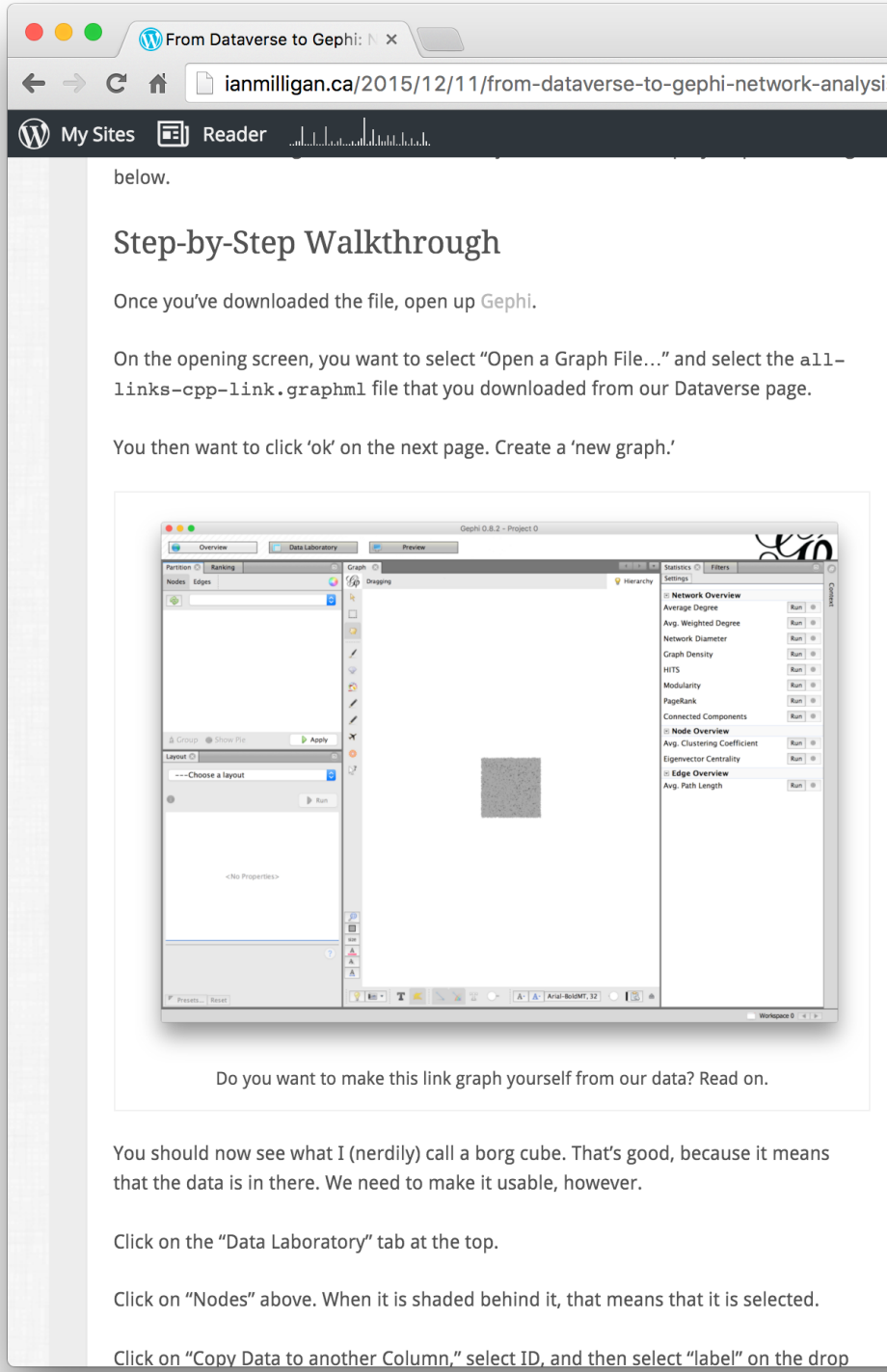
Good for public engagement - but limited for scholarship....



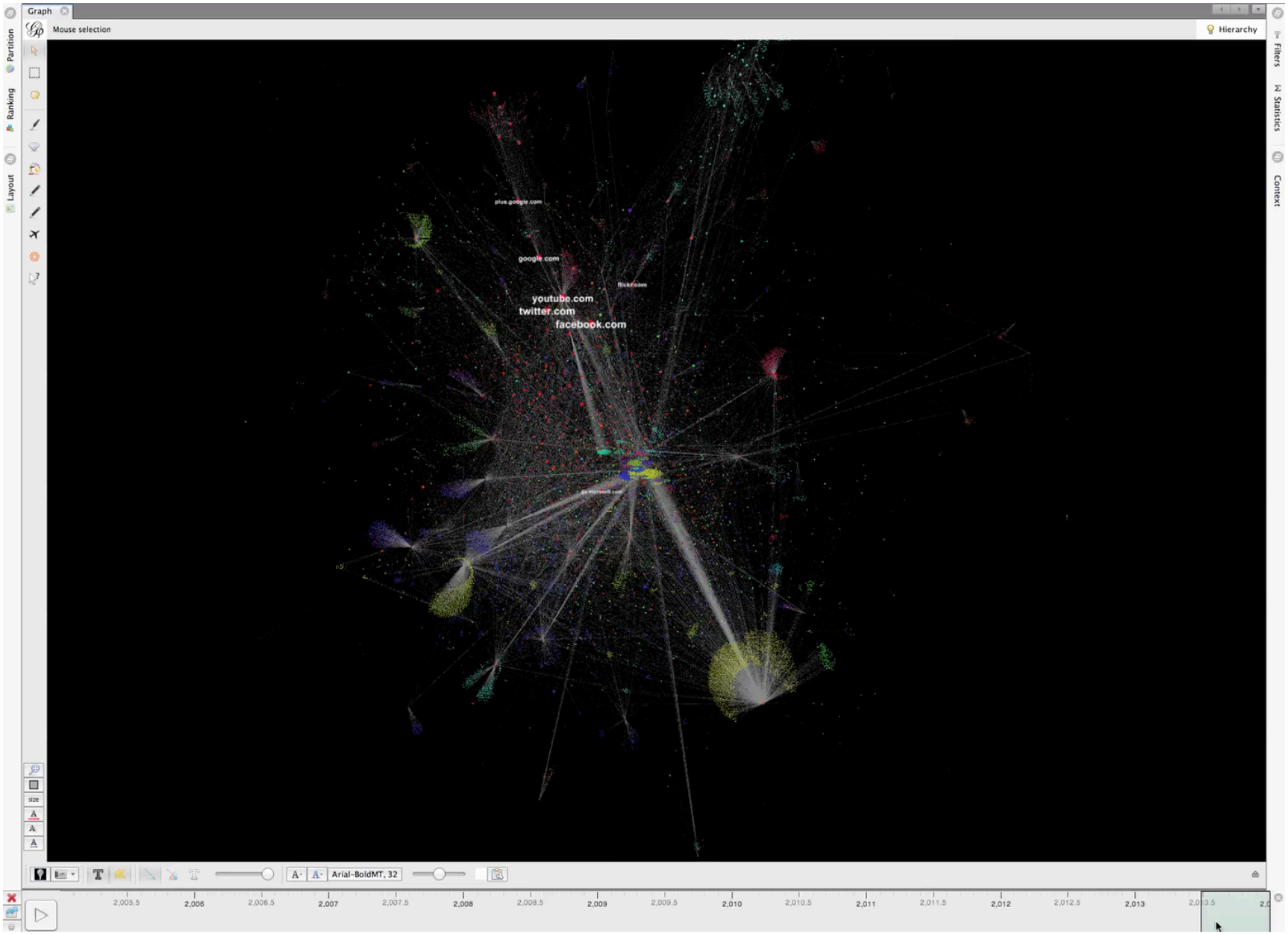
Getting over my bias towards content **and embracing metadata**

Gephi 0.8.2 (http://gephi.github.io/)

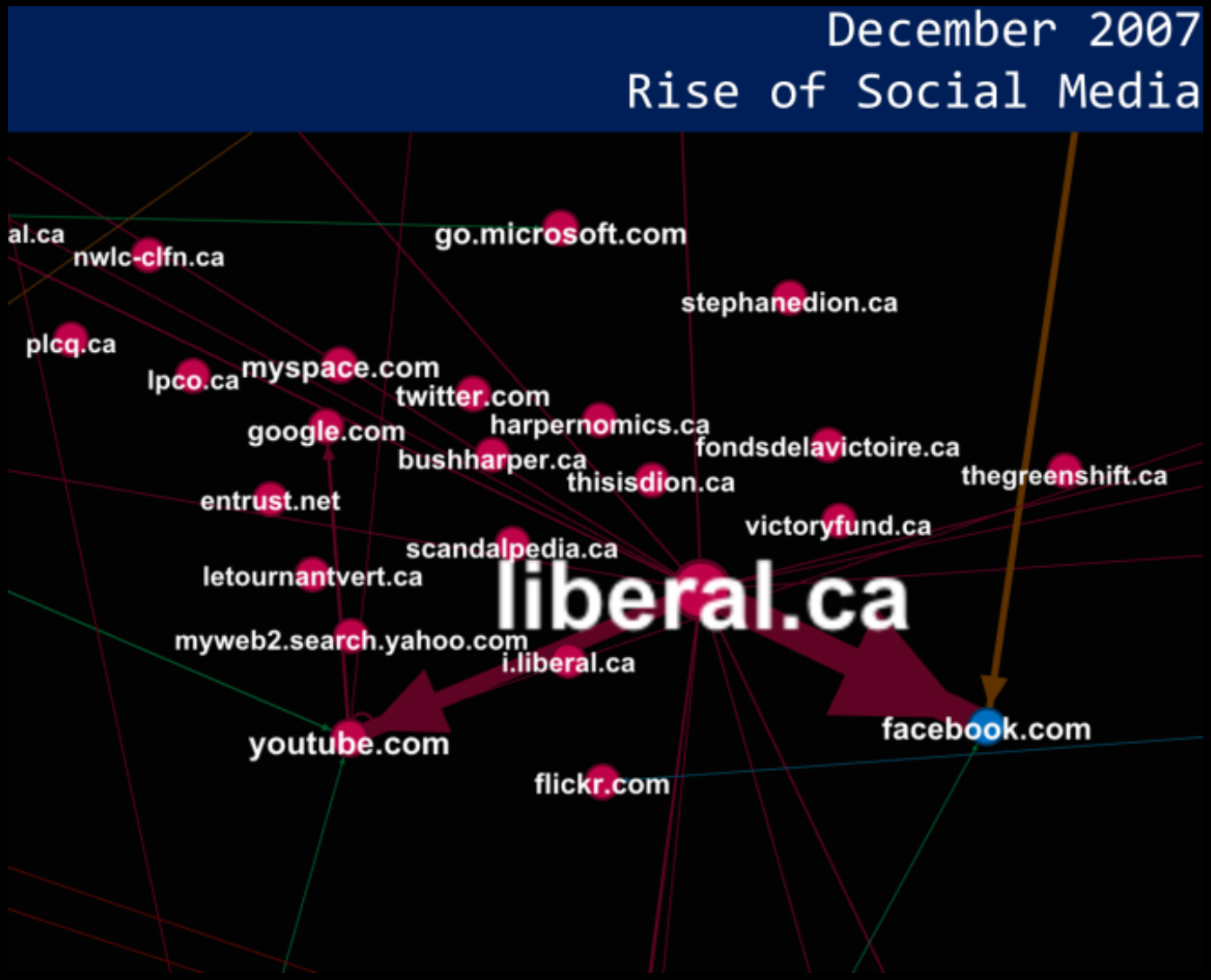
Walkthrough at ianmilligan.ca: "From Dataverse to Gephi" - try it on this data!



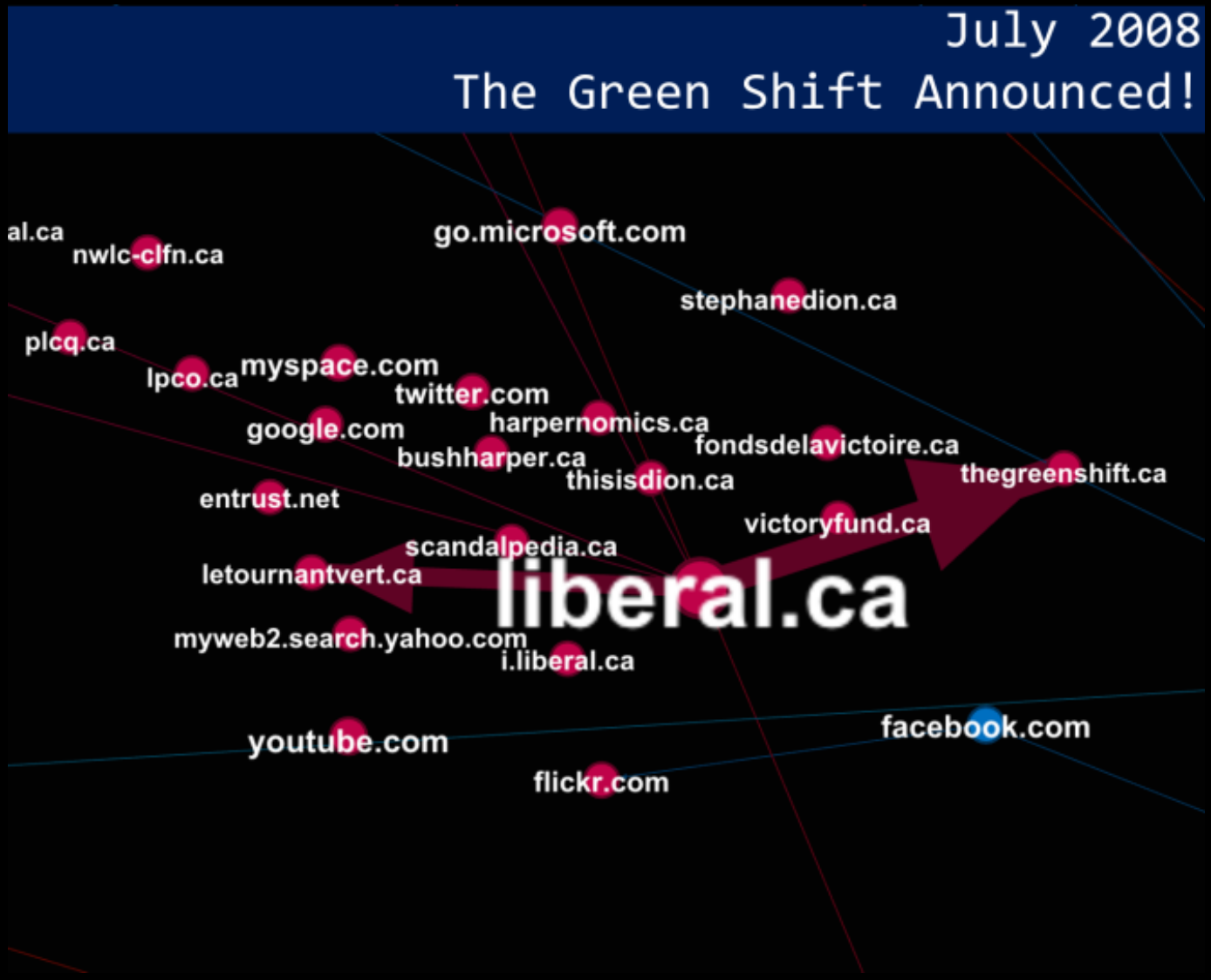
Metadata Extraction

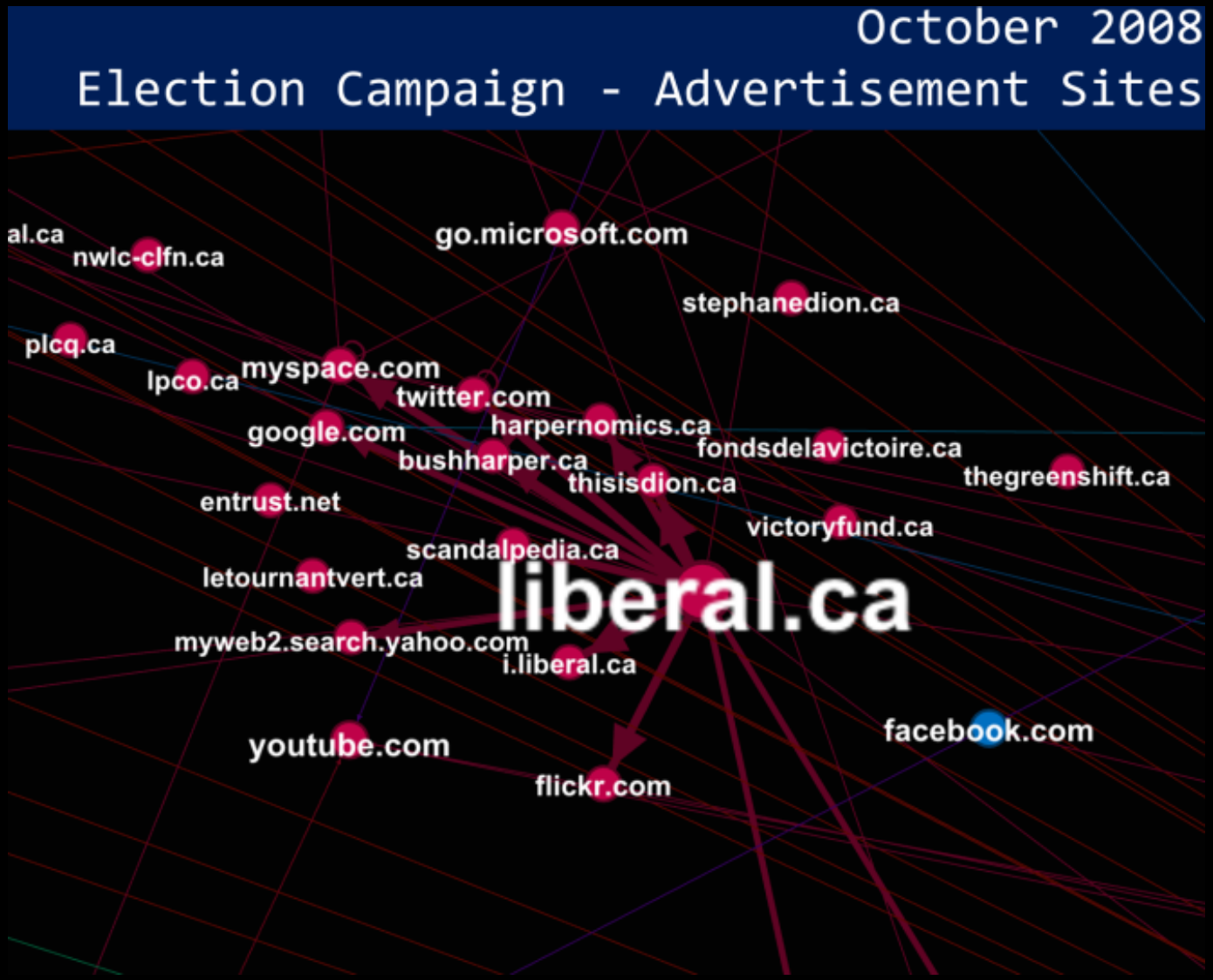



December 2006
Stephane Dion Elected Leader of Party
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
facebook.com
youtube.com
flickr.com


December 2007
Rise of Social Media
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
youtube.com
facebook.com
flickr.com

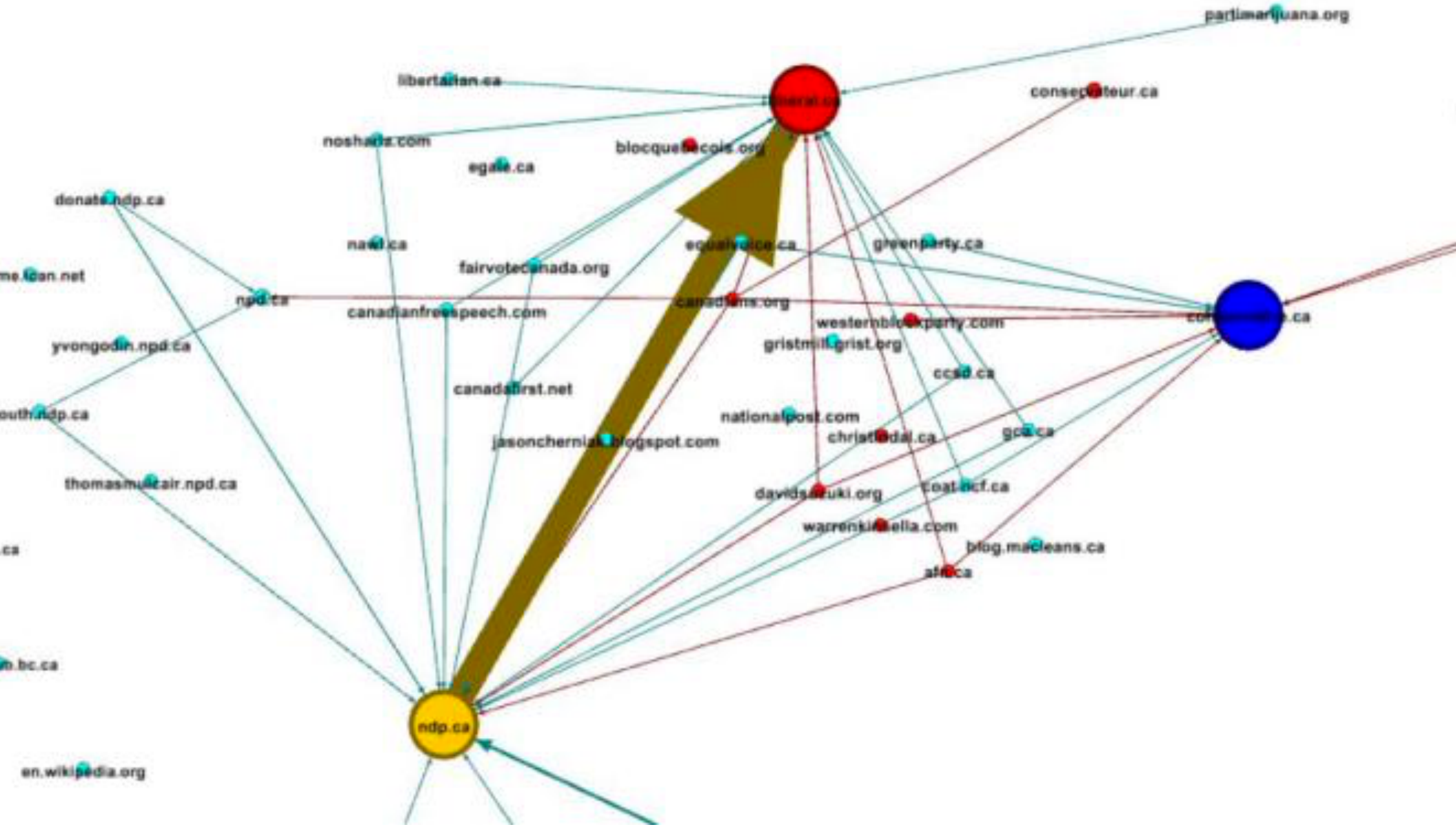

April 2008
Fundraising with the Victory Fund/
Fonds de la Victoire
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
facebook.com
youtube.com
flickr.com


July 2008
The Green Shift Announced!
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
youtube.com
facebook.com
flickr.com


October 2008
Election Campaign - Advertisement Sites
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
youtube.com
facebook.com
flickr.com


December 2008
Election campaign Ends; Attacking Harper
on Anti-American Grounds (bushharper)
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
facebook.com
youtube.com
flickr.com

2005 Canadian Federal Election

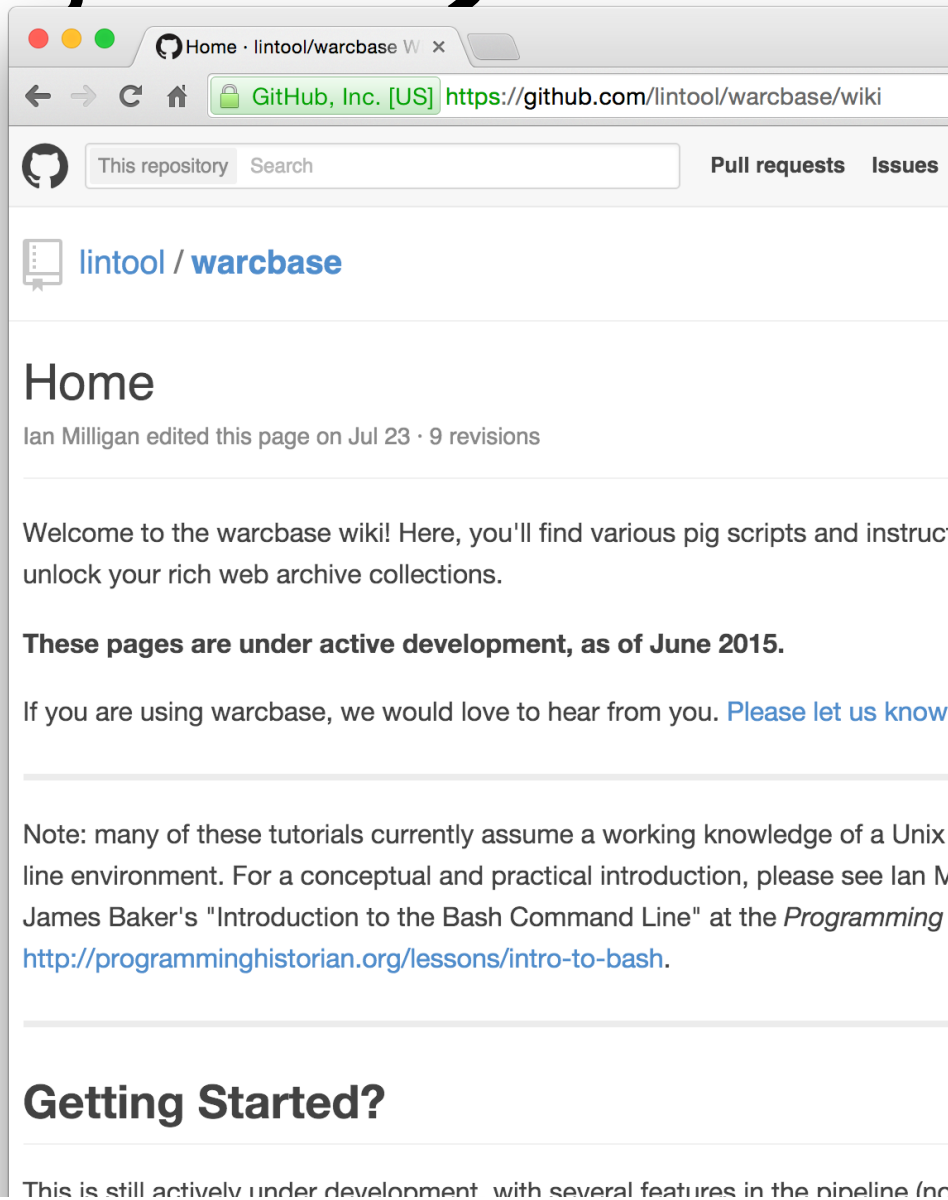


This requires
interdisciplinary
collaboration

Warcbase (soon to be twarcbase, too!)

- Jimmy Lin (main developer), Ian Milligan, Jeremy Wiebe, Alice Zhou, all University of Waterloo

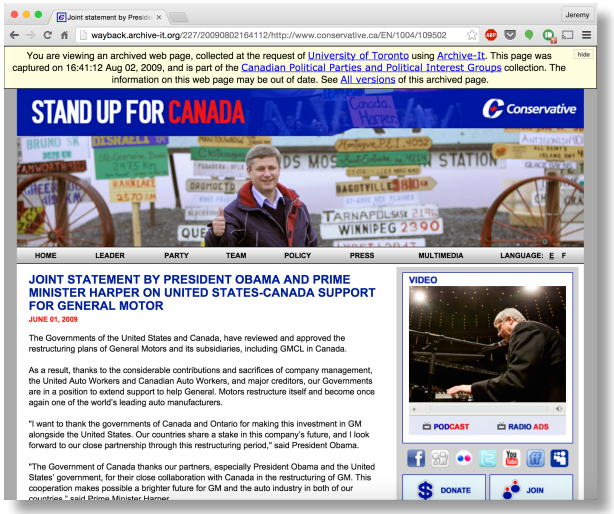
- Developing code, walkthroughs, and instructions for historians to be able to take a directory of WARCs and…



Warcbase

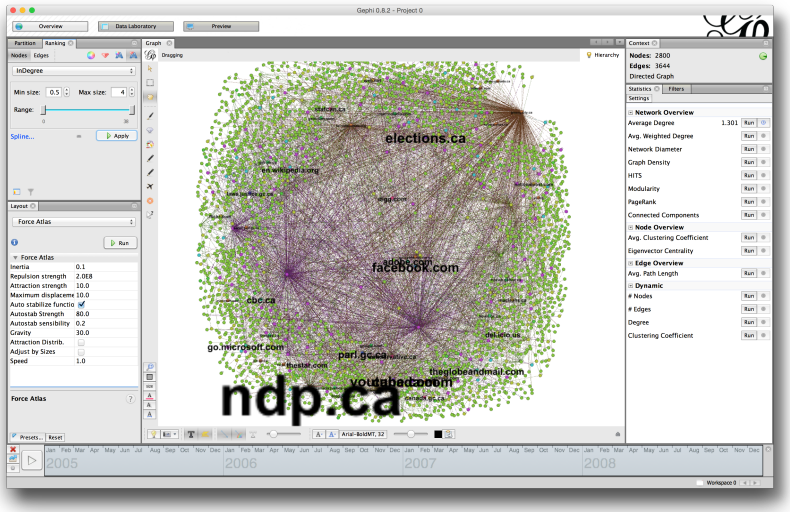
An open-source platform for managing web archives

http://warcbase.org



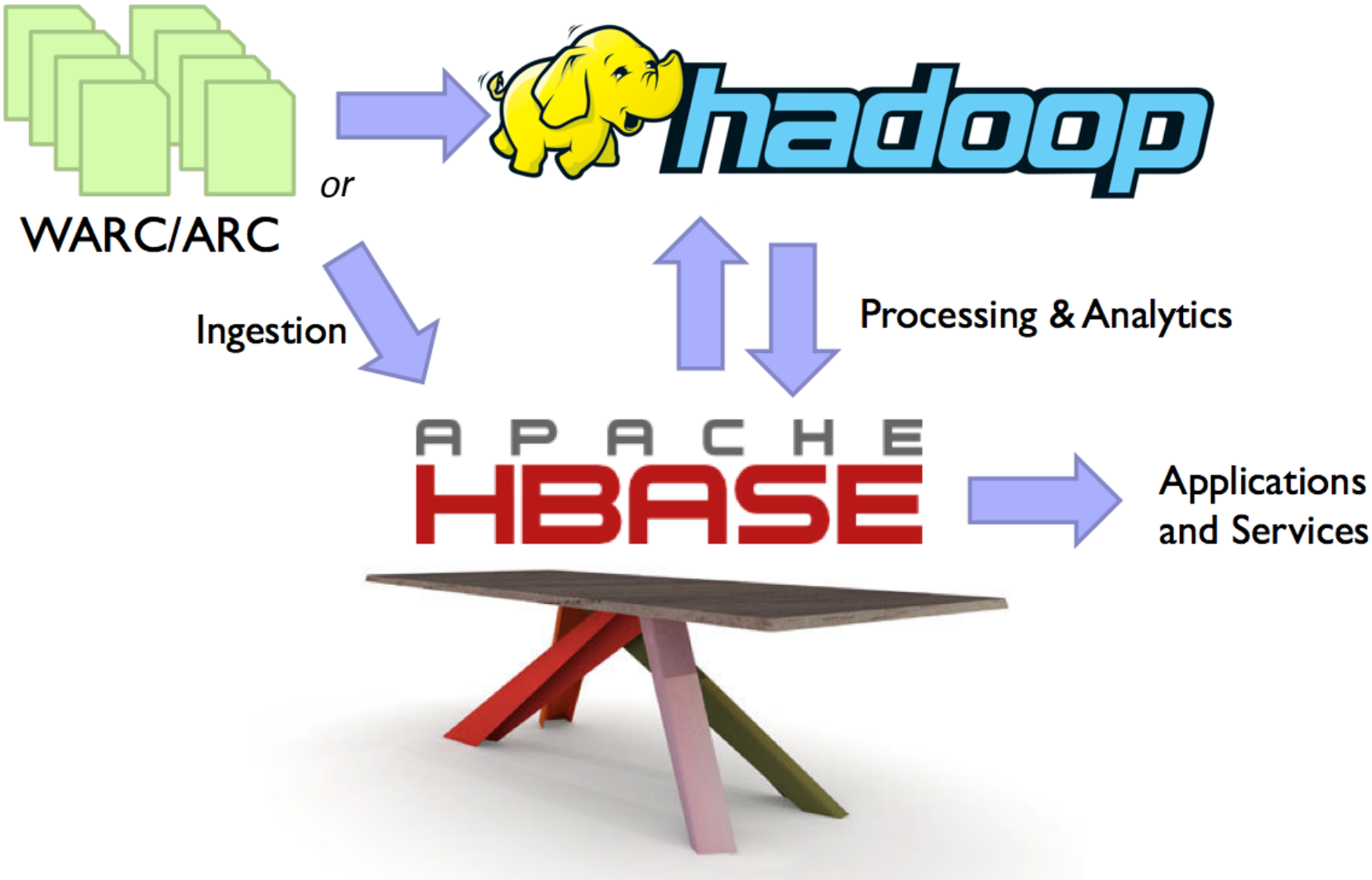
Two main facets

- A flexible data store: your own Wayback Machine
- **Scriptable analytics and data processing**



Funded by Mellon and (now) SSHRC.

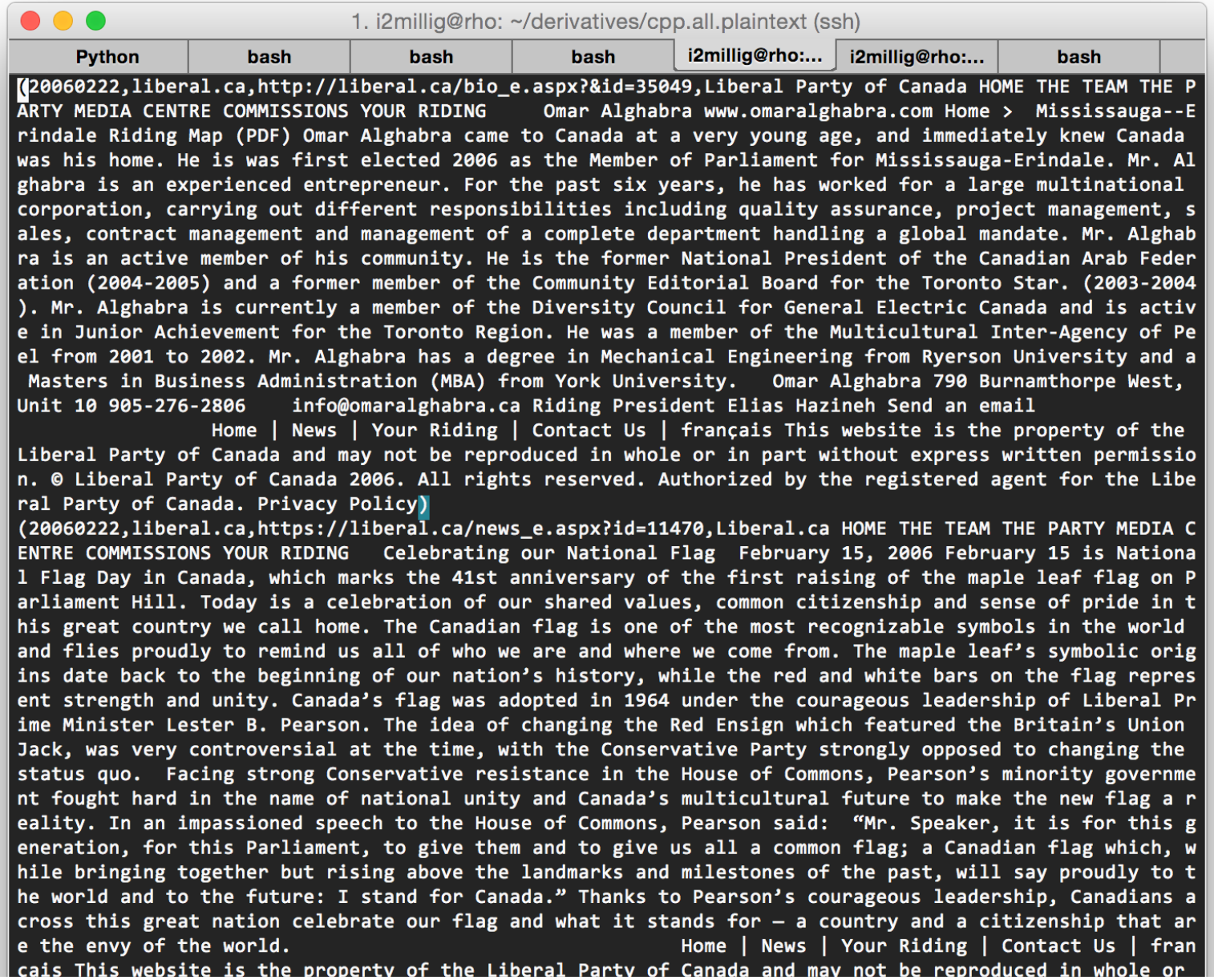
Warcbase


WARC/ARC
or
hadoop
Ingestion
Processing & Analytics
APACHE
HBASE
Applications
and Services

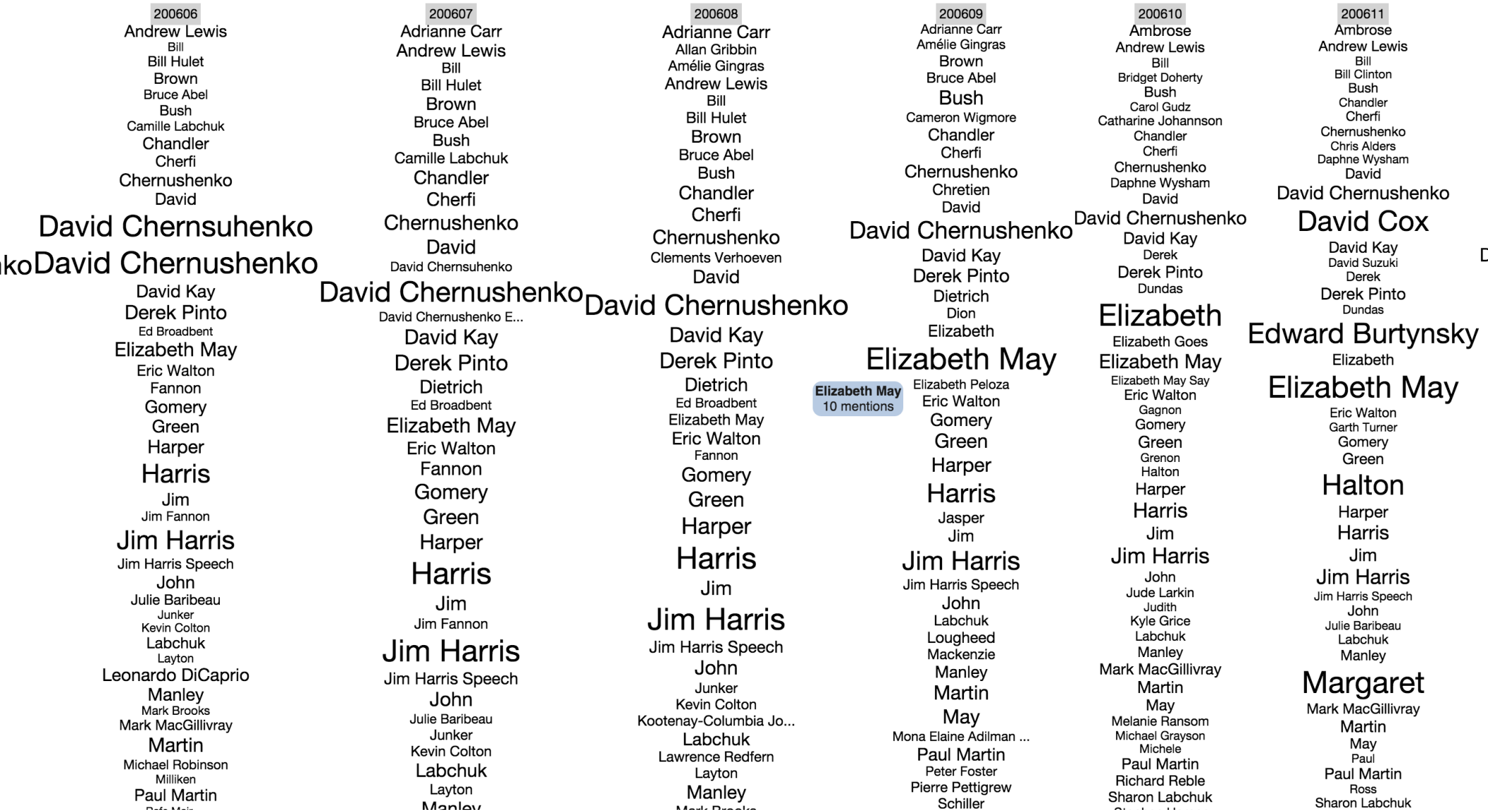
Warcbase

- Framework for distributed storage and distributed processing of very big data
- Scalable
  - From Raspberry Pi, to laptop, to powerful desktop, to single-node beefy server, to cluster
- Potentially very powerful
  - *Trantor*: 1.2PB of disk, 25 compute nodes (each w/ 128GB memory, 2×6-core Intel Xeon E5 v3 = 3.2TB memory and 300 current-generation Intel cores)

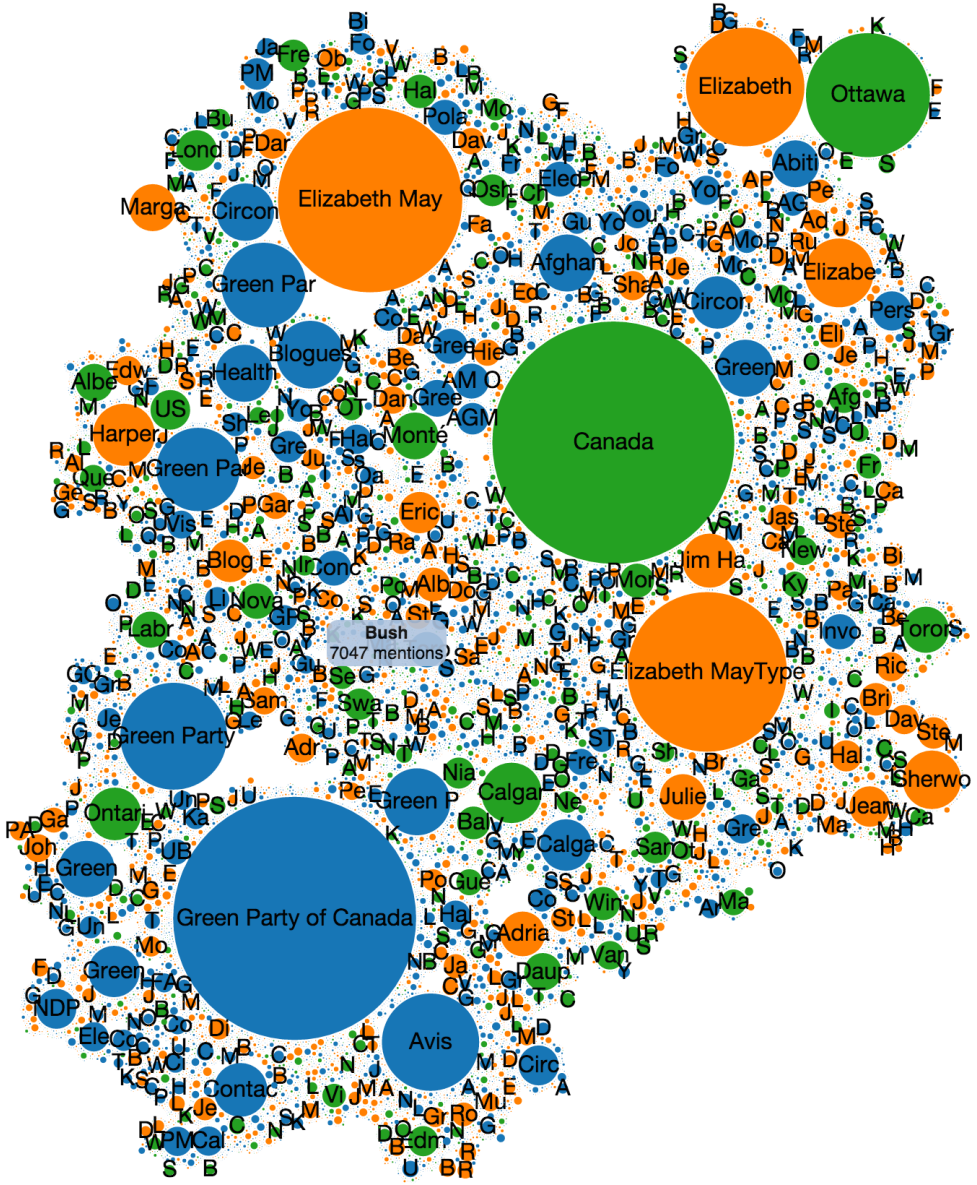
Extract all Plain Text



Extract Entities

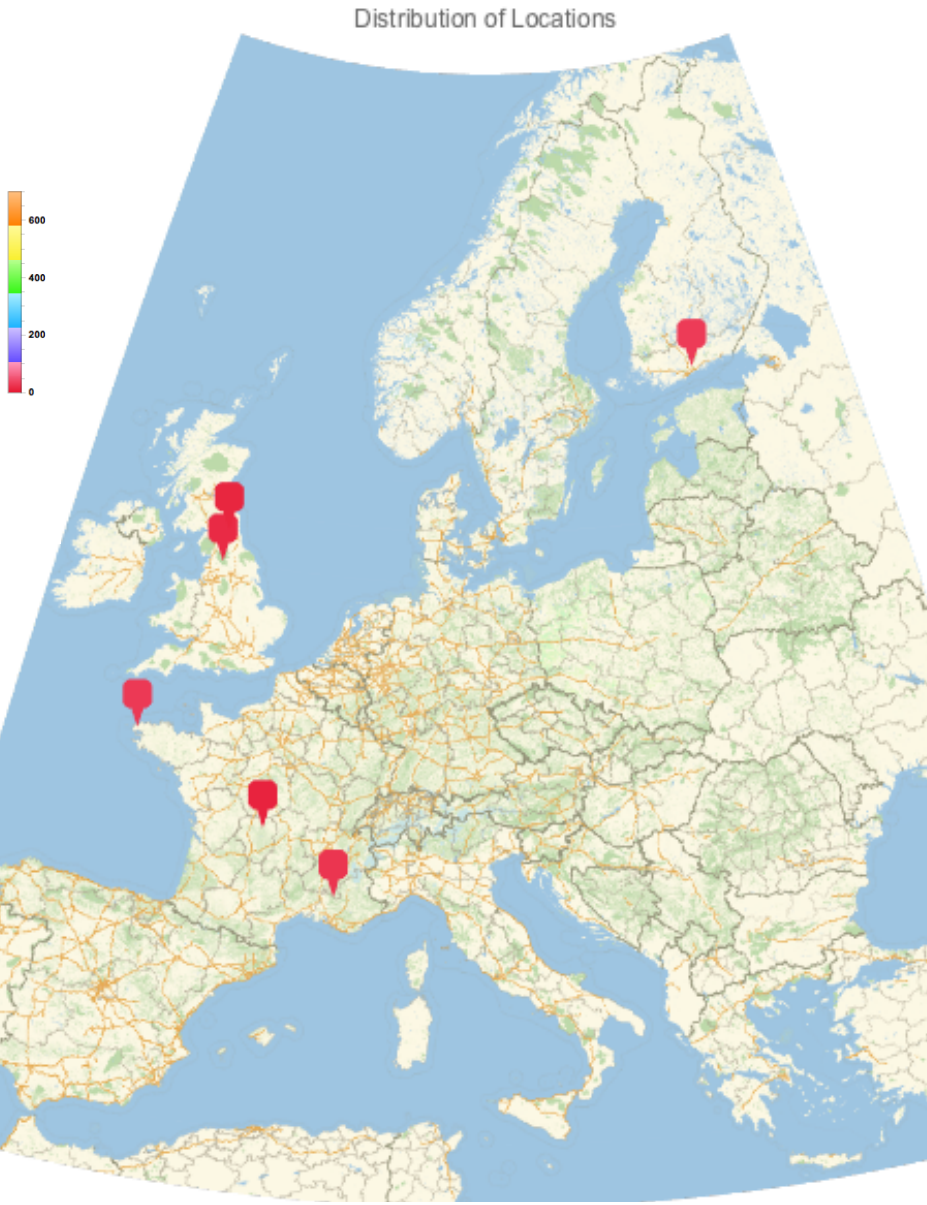


Extract Entities



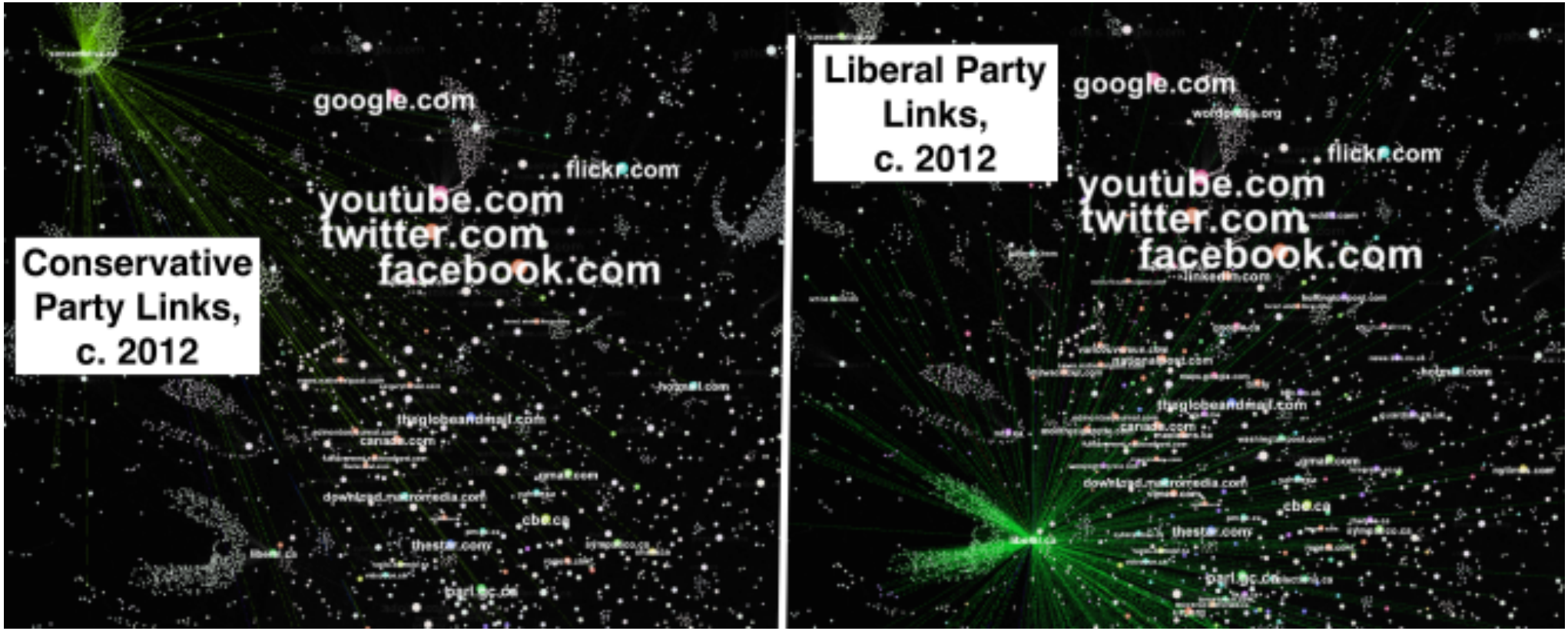
Extract Entities

Extract Entities

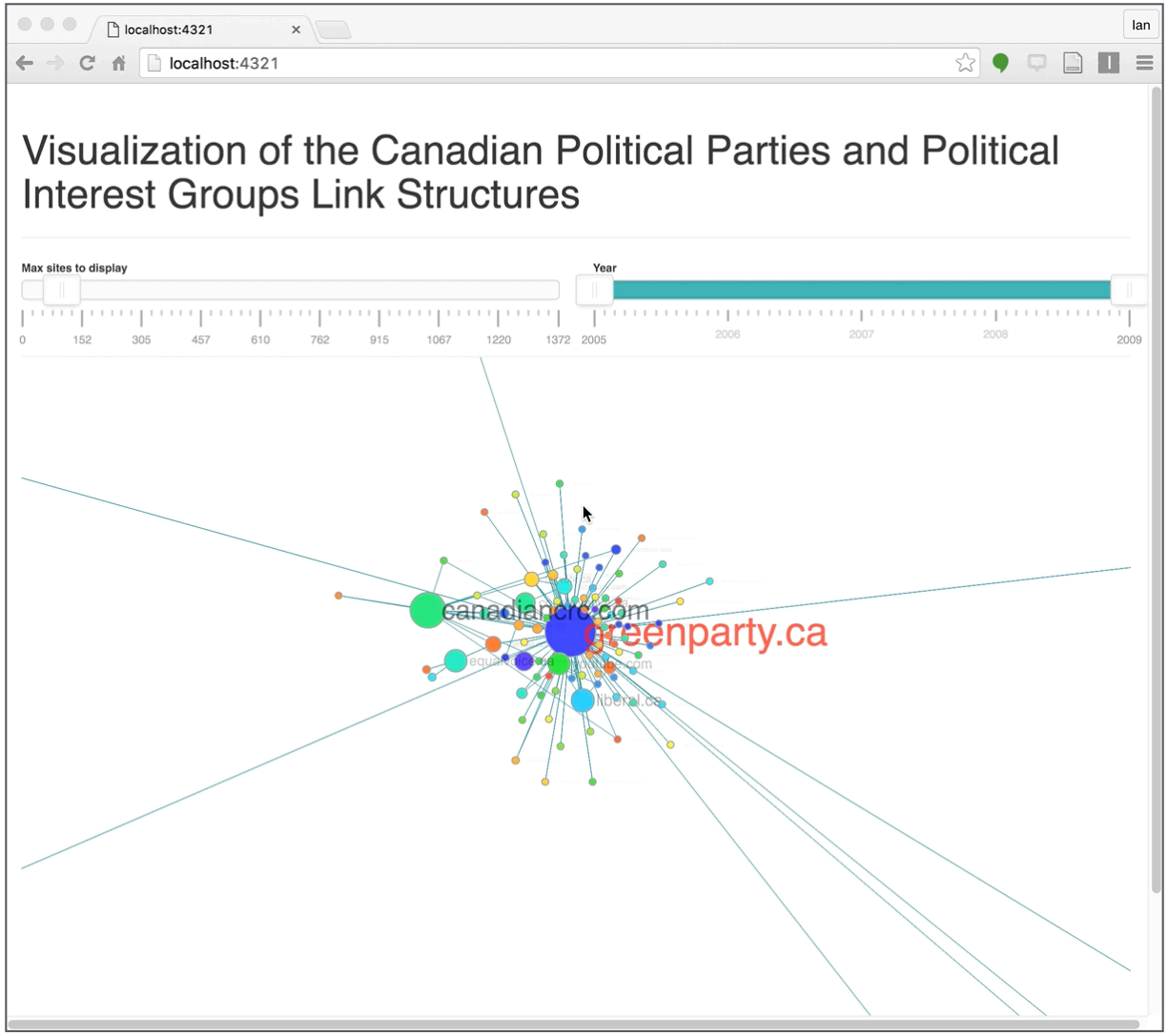



```
In[95]:= int = SemanticInterpretation[#] & /@ processedfreq[[All, 1]]
Out[95]= {Canada, Calgary, Colombia, Montreal, $Failed, Ontario, Canada, Afghanistan,
 Ottawa, Manitoba, Canada, British Columbia, Canada, Toronto, Nova Scotia, Canada,
 Saskatchewan, Canada, Quebec, Canada, Alberta, Canada, Winnipeg, Washington,
 Canada, Nunavut, Canada, White Horse, United States, Petawawa, Mumbai,
 Lima, The Americas, Saskatoon, Peru, Edmonton, Mexico, Chicoutimi-Jonquiere,
 New Brunswick, Canada, United States, Newfoundland and Labrador, Canada, Qandahar,
 $Failed, Asia-Pacific, Newfoundland and Labrador, Canada, Victoria, United States,
 Quebec City, India, $Failed, Atlantic Ocean, St. Germain, Durham, Battle of Waterloo,
```

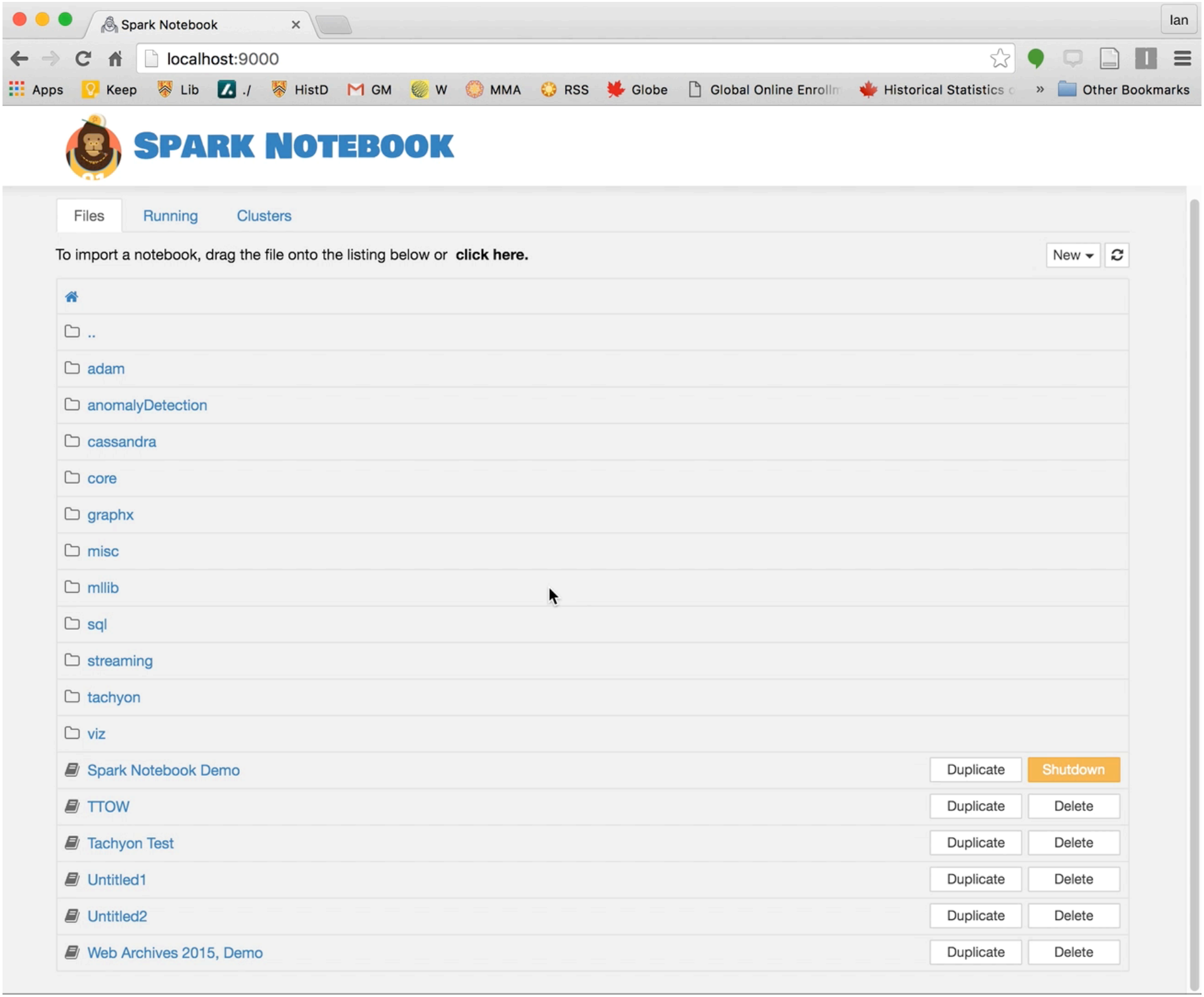
Extract Links/Gephi Connector



Or D3.js link networks in browser



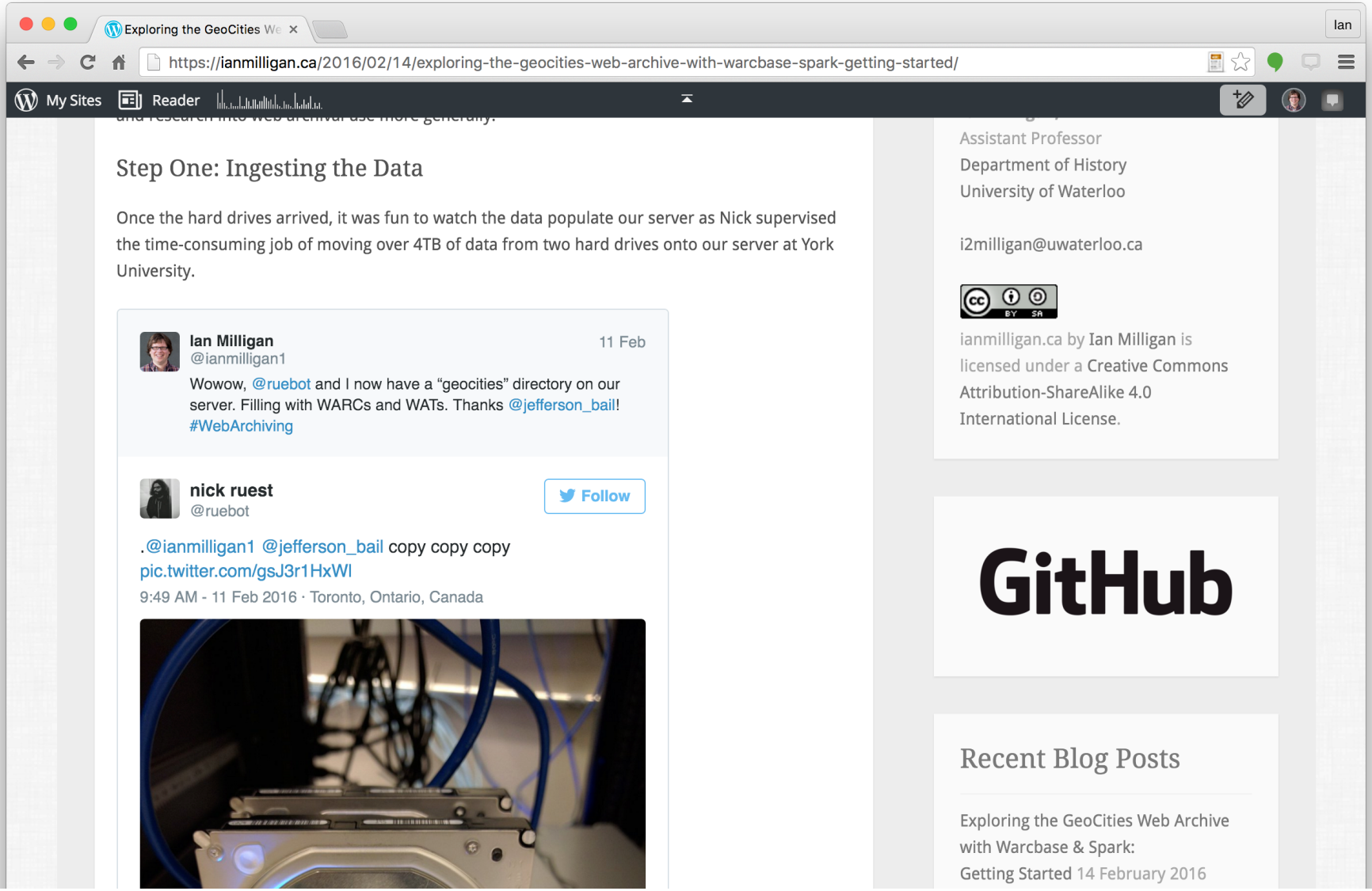
Bringing it all together in a notebook environment


Spark Notebook
Ian
localhost:9000
Apps
Keep
Lib
./
HistD
GM
W
MMA
RSS
Globe
Global Online Enrolln
Historical Statistics
Other Bookmarks
SPARK NOTEBOOK
Files
Running
Clusters
To import a notebook, drag the file onto the listing below or click here.
New
..
adam
anomalyDetection
cassandra
core
graphx
misc
mllib
sql
streaming
tachyon
viz
Spark Notebook Demo
Duplicate
Shutdown
TTOW
Duplicate
Delete
Tachyon Test
Duplicate
Delete
Untitled1
Duplicate
Delete
Untitled2
Duplicate
Delete
Web Archives 2015, Demo
Duplicate
Delete

All walkthroughs at:
docs.warcbase.org

Coming soon to Canadian web archive collections near you!

Follow along w/ GeoCities?



1. i2millig@rho: /mnt/vol1/data_sets/geocities/warcs (ssh)

bash | bash | i2millig@rho: /mnt/vol1/data...

```
GEOCITIES-20091029114236-00191-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029115416-00171-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029123034-00172-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029130439-00173-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029134536-00174-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029140344-00192-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029141553-00193-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029141726-00175-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029144445-00176-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029152151-00177-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029160824-00178-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029164941-00179-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029165037-00194-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029170431-00195-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029171605-00180-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029174154-00181-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029180818-00182-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029182725-00183-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029185858-00184-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029193728-00185-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029194541-00196-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029195911-00197-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029202041-00186-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029221340-00198-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029222459-00199-ia400110.us.archive.org.warc.gz
GEOCITIES-20091030021147-00197-ia400103.us.archive.org.warc.gz
GEOCITIES-20091030021444-00198-ia400103.us.archive.org.warc.gz
GEOCITIES-20091030022413-00171-ia400104.us.archive.org.warc.gz
i2millig@rho:/mnt/vol1/data_sets/geocities/warcs$ du -h
4.1T    .
i2millig@rho:/mnt/vol1/data_sets/geocities/warcs$
```


2. ssh
ianmilligan1@Ians-MacBook-Pro:~$ rho
i2millig@rho.library.yorku.ca's password:
Welcome to Ubuntu 14.04.2 LTS (GNU/Linux 3.13.0-32-generic x86_64)
* Documentation: https://help.ubuntu.com/
System information as of Mon Mar 7 13:43:20 EST 2016
System load: 0.99 Users logged in: 1
Usage of /: 34.7% of 744.67GB IP address for em1: 130.63.180.18
Memory usage: 16% IP address for em2: 10.0.0.18
Swap usage: 6% IP address for docker0: 172.17.0.1
Processes: 359
Graph this data and manage this system at:
https://landscape.canonical.com/
242 packages can be updated.
130 updates are security updates.
Last login: Mon Mar 7 13:43:21 2016 from 38.123.136.254
i2millig@rho:~$ ./spark-1.5.1/bin/spark-shell --jars ~/warcbase/target/warcbase-0.1.0-SNAPSHOT-fatjar.jar
WARN NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Welcome to
version 1.5.1
Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_45)
Type in expressions to have them evaluated.
Type :help for more information.
WARN MetricsSystem - Using default name DAGScheduler for source because spark.app.id is not set.
Spark context available as sc.
SQL context available as sqlContext.
scala> :paste
// Entering paste mode (ctrl-D to finish)
import org.warcbase.spark.matchbox._
import org.warcbase.spark.rdd.RecordRDD._
val r =
RecordLoader.loadWarc("/mnt/vol1/data_sets/geocities/warcs/GEOCITIES-20090808133634-04399-crawling08.us.archive.org.warc.gz", sc)
.keepValidPages()
.map(r => ExtractTopLevelDomain(r.getUrl))
.countItems()
.take(10)
// Exiting paste mode, now interpreting.
INFO WacWarcInputFormat - Loading file:/mnt/vol1/data_sets/geocities/warcs/GEOCITIES-20090808133634-04399-crawling08.us.archive.org.warc.g
z
import org.warcbase.spark.matchbox._
import org.warcbase.spark.rdd.RecordRDD._
r: Array[(String, Int)] = Array((geocities.com,3748), (www.geocities.com,240), (www.myfilehut.com,12), (asiarooms.com,7), (us.geocities.com
,6), (www.theginge.com,3), (www.angelfire.com,3), (images.quizilla.com,3), (pub28.bravenet.com,3), (ss.webring.yahoo.com,2))
scala>

Extract all URLs

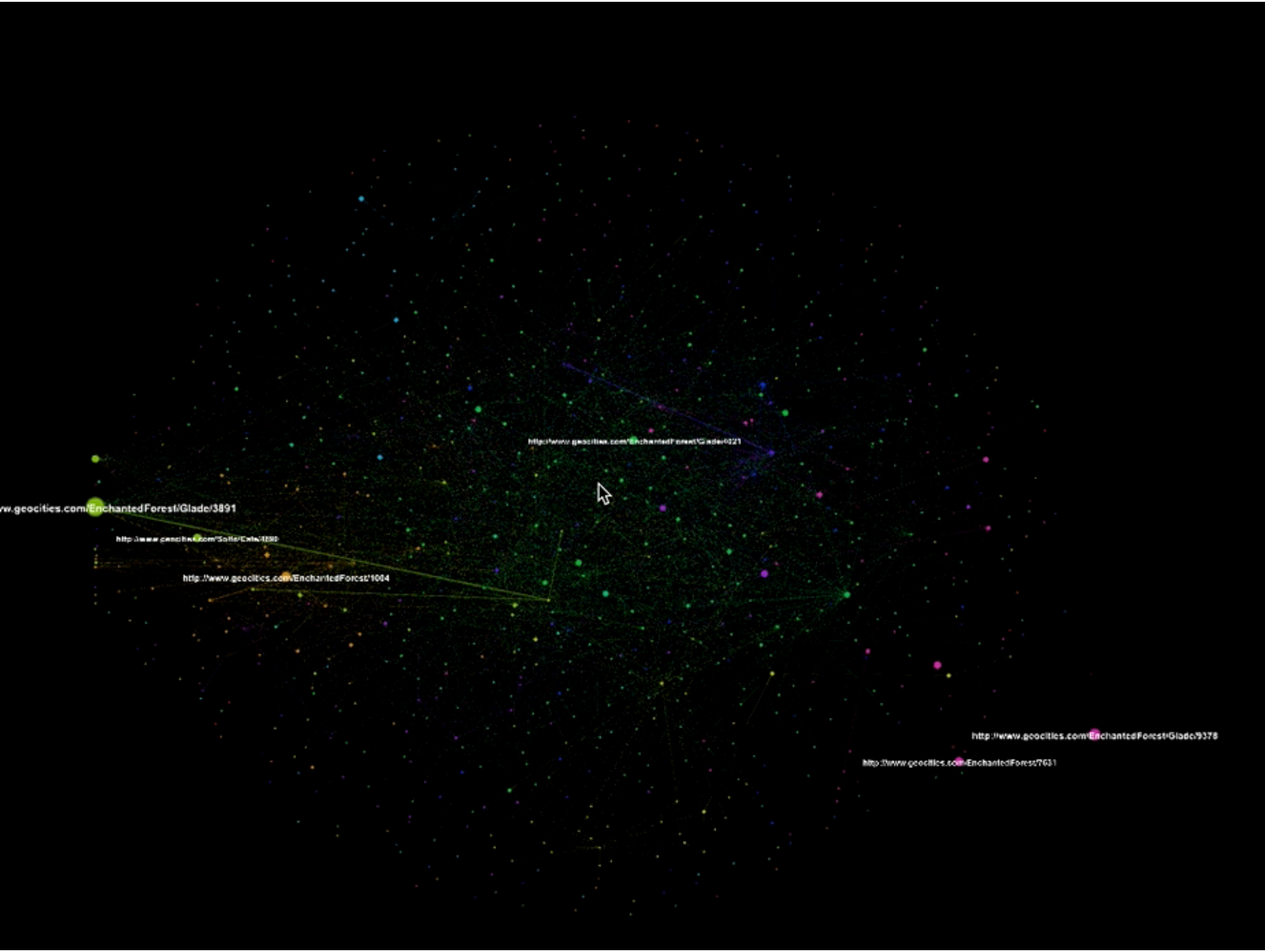
```
1  import org.warcbase.spark.matchbox._
2  import org.warcbase.spark.rdd.RecordRDD._
3
4  val r = RecordLoader.loadWarc("/mnt/vol1/data_sets/geocities/
       warcs", sc)
5  .keepValidPages()
6  .map(r => r.getUrl)
7  .saveAsTextFile("/mnt/vol1/derivative_data/geocities/url-list")
```

9.9 GB text file, 186,761,346 URLs

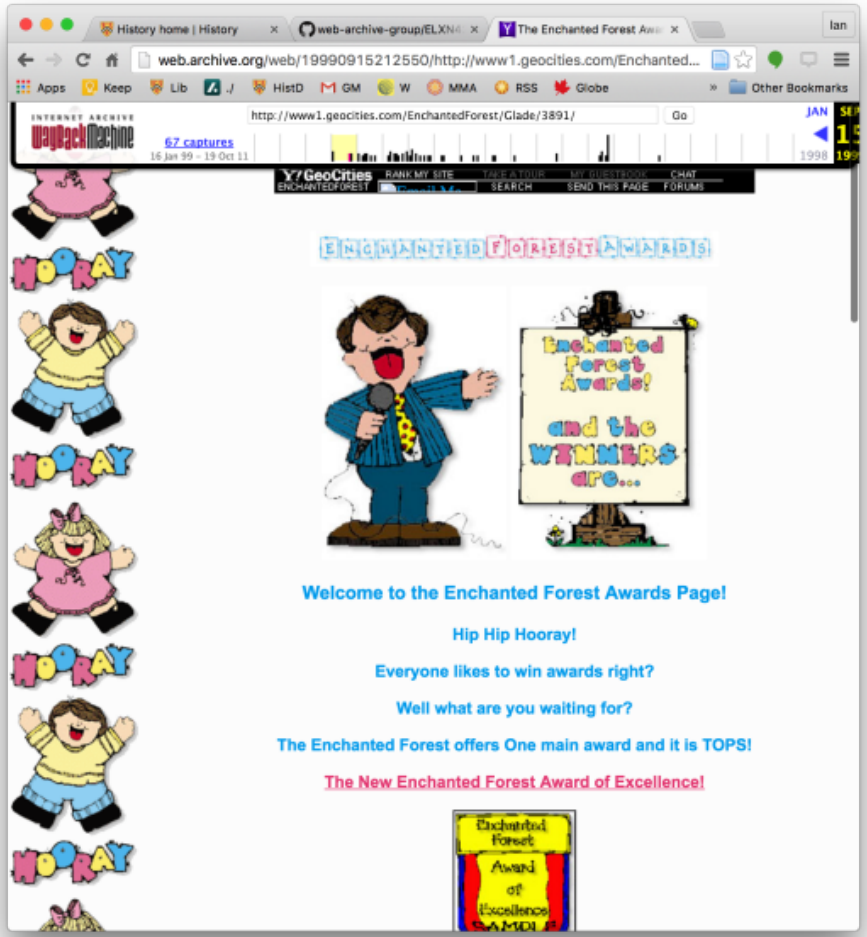
Extract a Link Graph

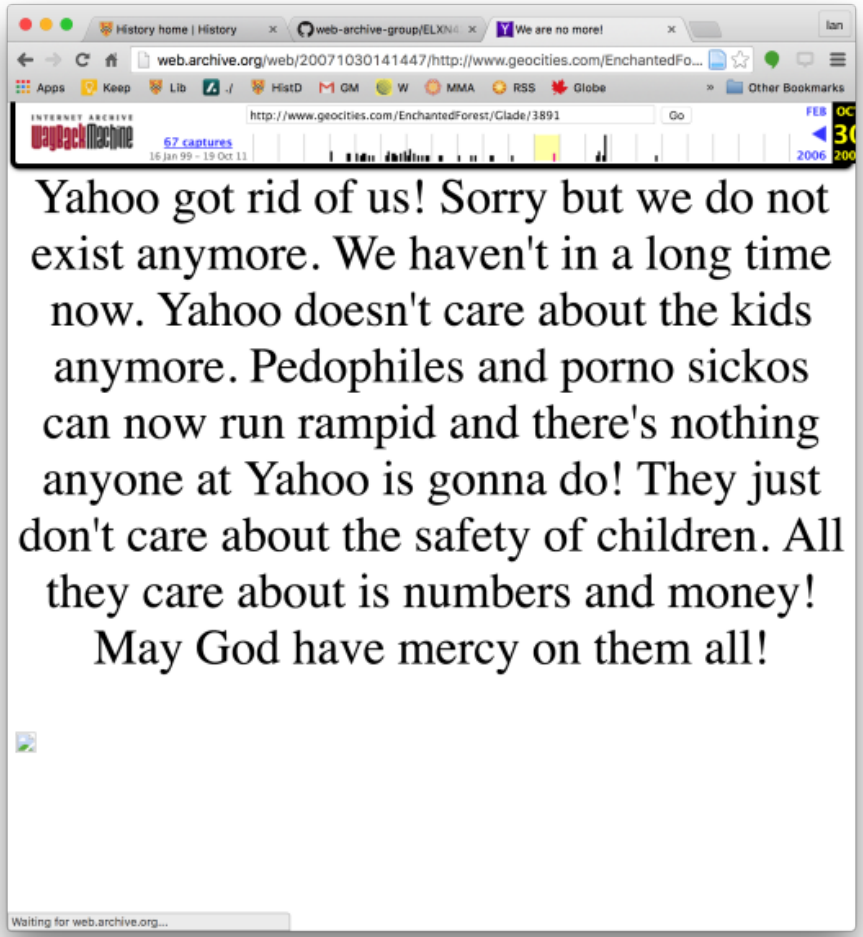
```
1  import org.warcbase.spark.matchbox.{ExtractTopLevelDomain,
       ExtractLinks, RecordLoader}
2  import org.warcbase.spark.rdd.RecordRDD._
3
4  RecordLoader.loadArc("/mnt/vol1/data_sets/geocities/warcs/*", sc)
5  .keepValidPages()
6  .map(r => (r.getCrawldate, ExtractLinks(r.getUrl, r.
       getContentString)))
7  .flatMap(r => r._2.map(f => (r._1, ExtractTopLevelDomain(f._1).
       replaceAll("^\\s*www\\.", ""), ExtractTopLevelDomain(f._2).
       replaceAll("^\\s*www\\.", ""))))
8  .filter(r => r._2 != "" && r._3 != "")
9  .countItems()
10 .filter(r => r._2 > 5)
11 .saveAsTextFile("/mnt/vol1/data_sets/geocities/geocities.
       sitelinks")
```


es.com/WestHollywood/Castro/6902

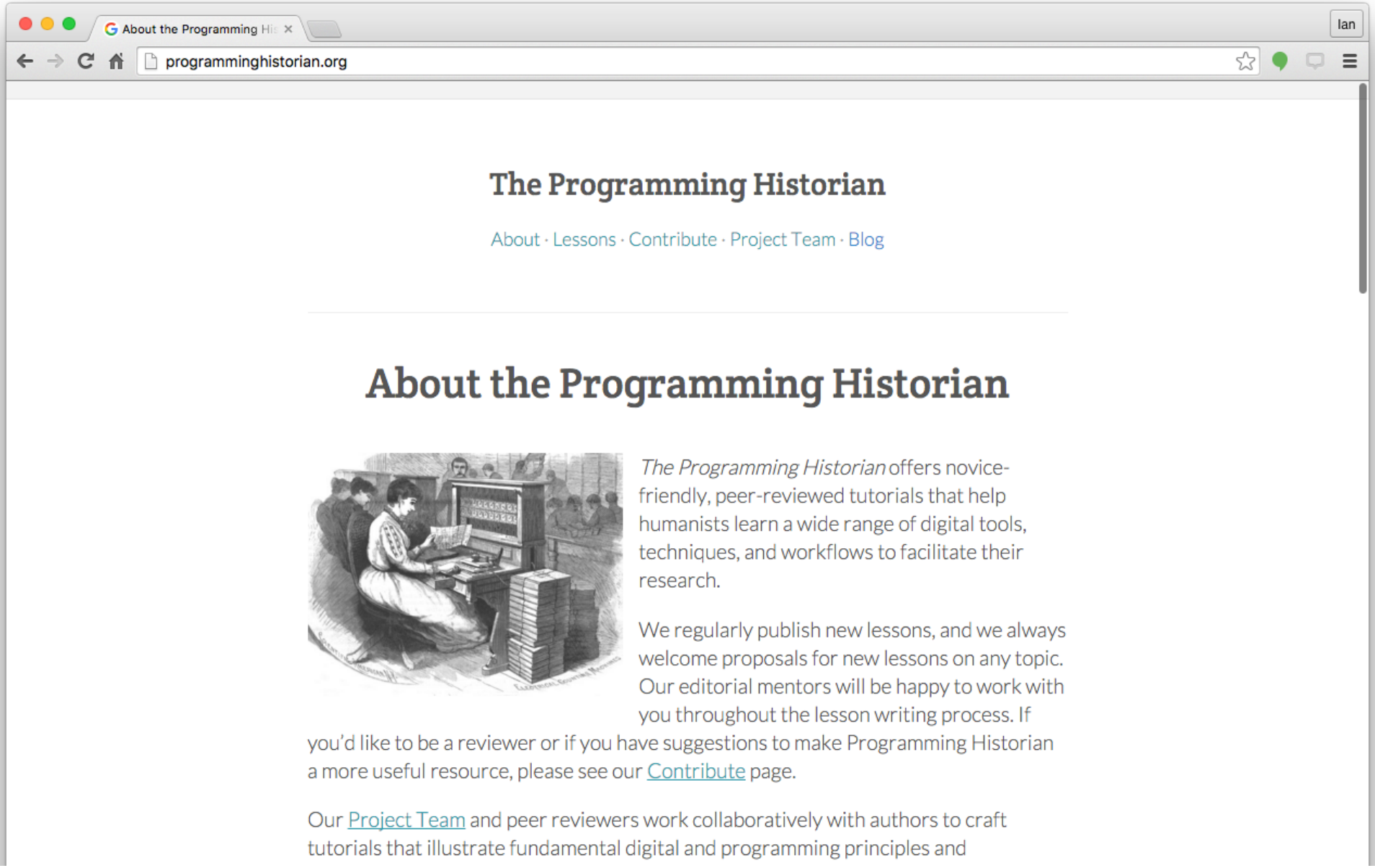

http://www.geocities.com/EnchantedForest/Glade/9221
ww.geocities.com/EnchantedForest/Glade/3891
http://www.geocities.com/EnchantedForest/1004
http://www.geocities.com/EnchantedForest/Glade/9378
http://www.geocities.com/EnchantedForest/7631

Label	PageRank	In-De...	Out-D...
http://www.geocities.com/EnchantedForest/Glade/3891	0.008	62	0
http://www.geocities.com/EnchantedForest/Glade/9378	0.005	2	2
http://www.geocities.com/EnchantedForest/1004	0.005	23	14
http://www.geocities.com/EnchantedForest/7631	0.004	2	1
http://www.geocities.com/SoHo/Cafe/4690	0.004	32	0
http://www.geocities.com/EnchantedForest/Glade/4021	0.004	26	0
http://www.geocities.com/TheTropics/Paradise/4079	0.003	31	0
http://www.geocities.com/RainForest/Vines/4892	0.003	3	3
http://www.geocities.com/EnchantedForest/3278	0.003	15	0
http://www.geocities.com/EnchantedForest/3696	0.003	15	0
http://www.geocities.com/EnchantedForest/Dell/5914	0.003	13	0
http://www.geocities.com/EnchantedForest/1469	0.003	5	20
http://www.geocities.com/EnchantedForest/Tower/9644	0.003	3	13


web.archive.org/web/19990915212550/http://www1.geocities.com/Enchanted...
http://www1.geocities.com/EnchantedForest/Glade/3891/
67 captures
16 Jan 99 - 19 Oct 11
GeoCities
RANK MY SITE
SEARCH
SEND THIS PAGE
FORUMS
ENCHANTED FOREST AWARDS
HOORAY
HOORAY
HOORAY
HOORAY
enchanted forest awards!
and the WINNERS are...
Welcome to the Enchanted Forest Awards Page!
Hip Hip Hooray!
Everyone likes to win awards right?
Well what are you waiting for?
The Enchanted Forest offers One main award and it is TOPS!
The New Enchanted Forest Award of Excellence!
Enchanted Forest
Award
of
Excellence


web.archive.org/web/20071030141447/http://www.geocities.com/EnchantedFo...
http://www.geocities.com/EnchantedForest/Glade/3891
67 captures
16 Jan 99 - 19 Oct 11
Yahoo got rid of us! Sorry but we do not exist anymore. We haven't in a long time now. Yahoo doesn't care about the kids anymore. Pedophiles and porno sickos can now run rampid and there's nothing anyone at Yahoo is gonna do! They just don't care about the safety of children. All they care about is numbers and money! May God have mercy on them all!
Waiting for web.archive.org...

Where to learn?

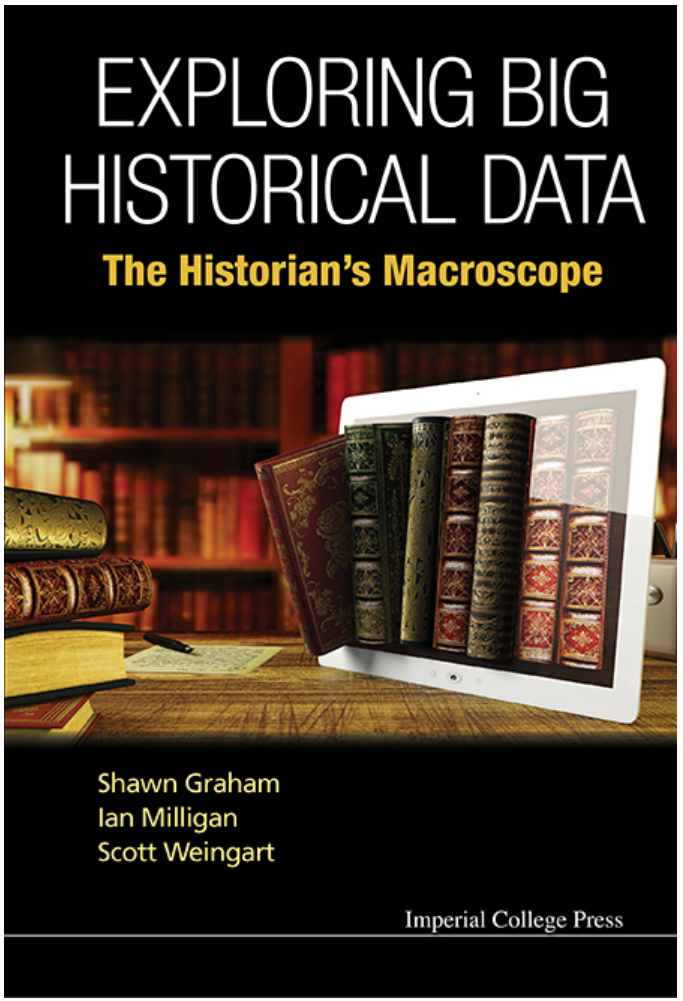


Programming Historian

- Network Analysis Lessons
- Topic Modeling Lessons
- Command Line Lessons
- etc.

Exploring Big Historical Data

- Check out our draft at macroscope.org
  - Conceptual introduction to topic modelling
  - Network analysis
  - Visualizations
  - Field of digital humanities



In-Person Events

- Hackathons!
- **Toronto, ON**. March 2016 (SSHRC/NSF supported).
- **Hannover, Germany**. May 2016 (WebSci ’16).
- **Washington DC**, June 2016 (SSHRC/NSF travel funding for graduate students/PDFs).

... but most of all, a willingness to learn and fail.

Because, as I hope I have shown today..

it's worth it.

More voices, more people, the promise of social history achieved.

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Ontario

compute • calcul CANADA

UNIVERSITY OF WATERLOO

Thanks very much!

Questions?

Ian Milligan
Assistant Professor
@ianmilligan1

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History