Big Data and History:
Seeing the Past through a Macroscope

Danish Society for Research on Contemporary History, Copenhagen

---

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?**

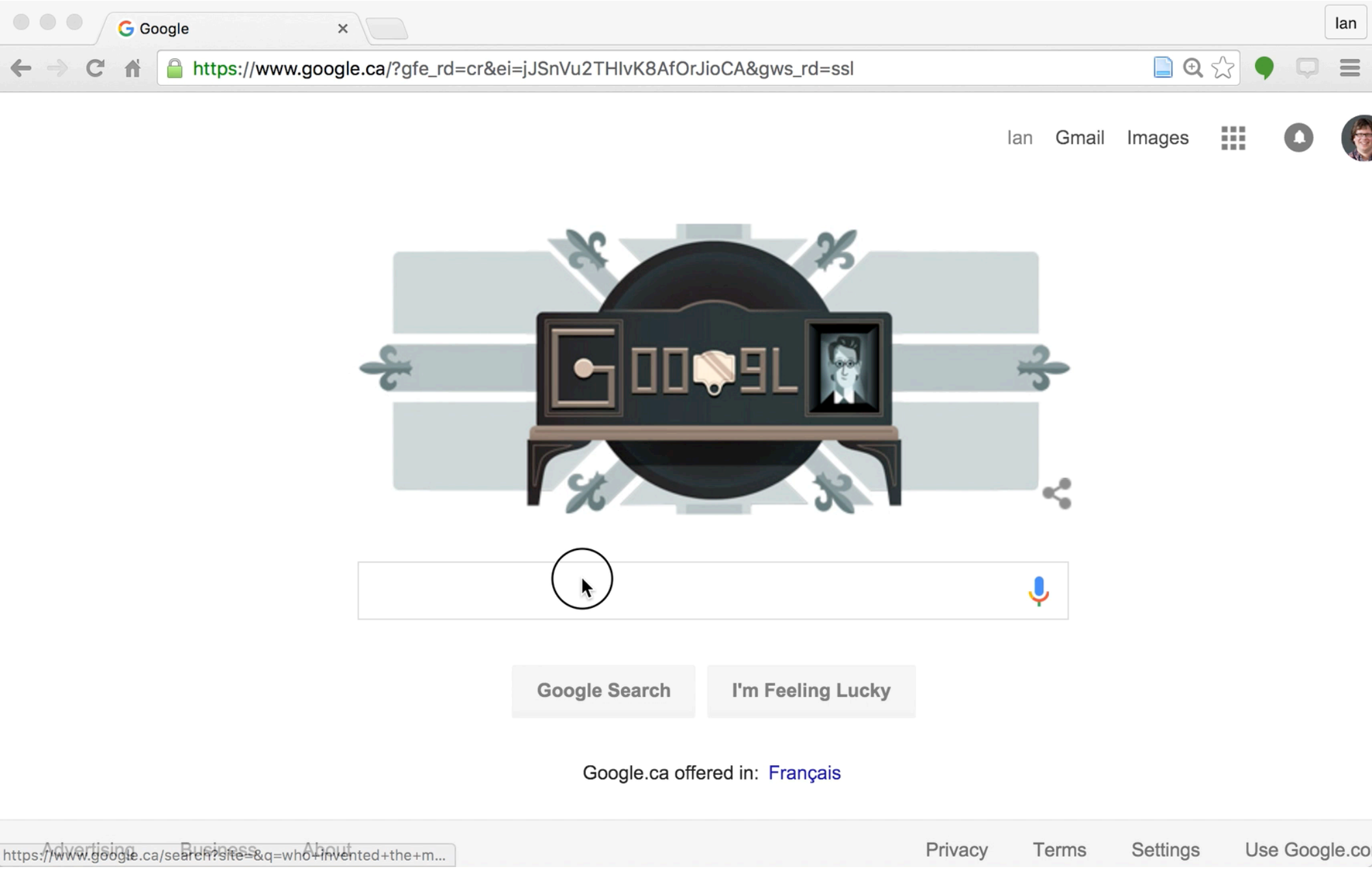
A Prologue: Big Data is Everywhere

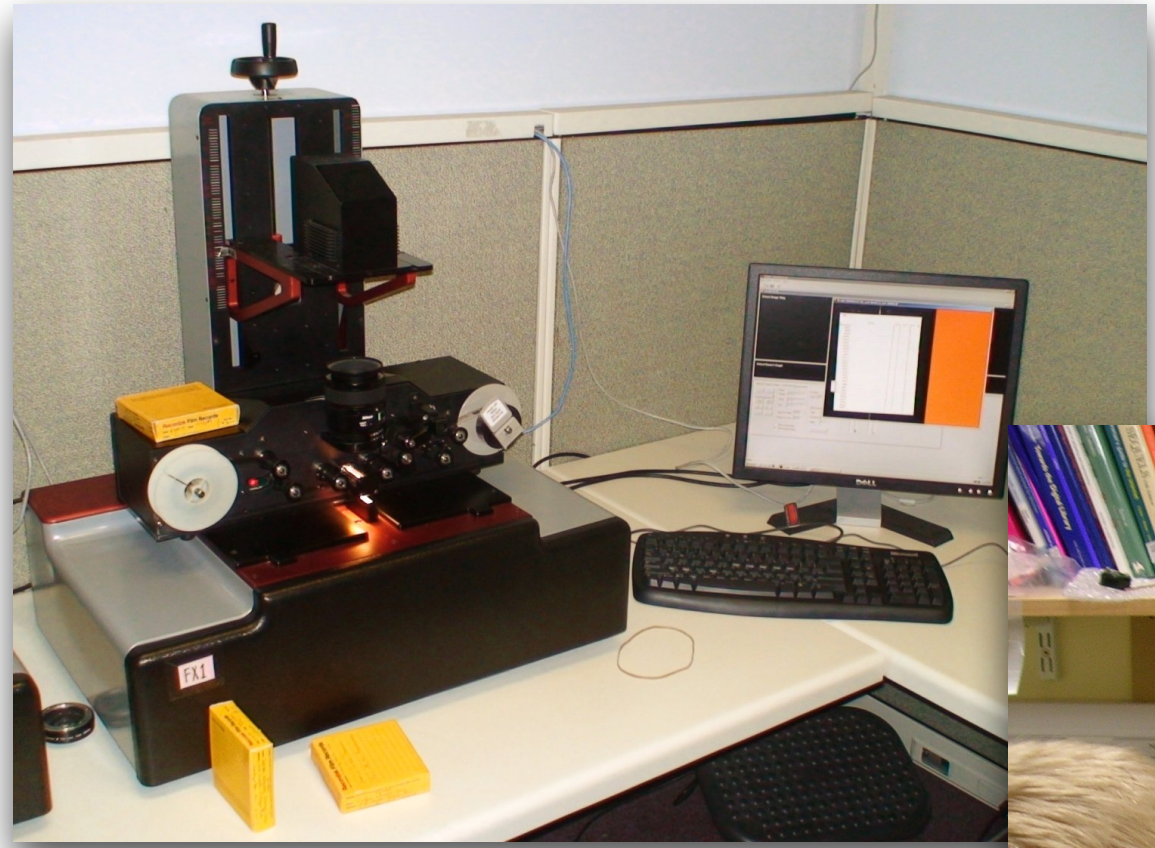
What do we mean by Big Data?

- 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=jJSnVu2THIvK8AfOrJioCA&gws_rd=ssl
Ian
Gmail
Images
Google Search
I'm Feeling Lucky
Google.ca offered in: Français
Advertising
Business
About
Privacy
Terms
Settings
Use Google.co
https://www.google.ca/search?site=&q=who+invented+the+m...


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

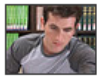
Search Clear form

Search subject areas

Use search forms customized for each subject.


The Arts


Business

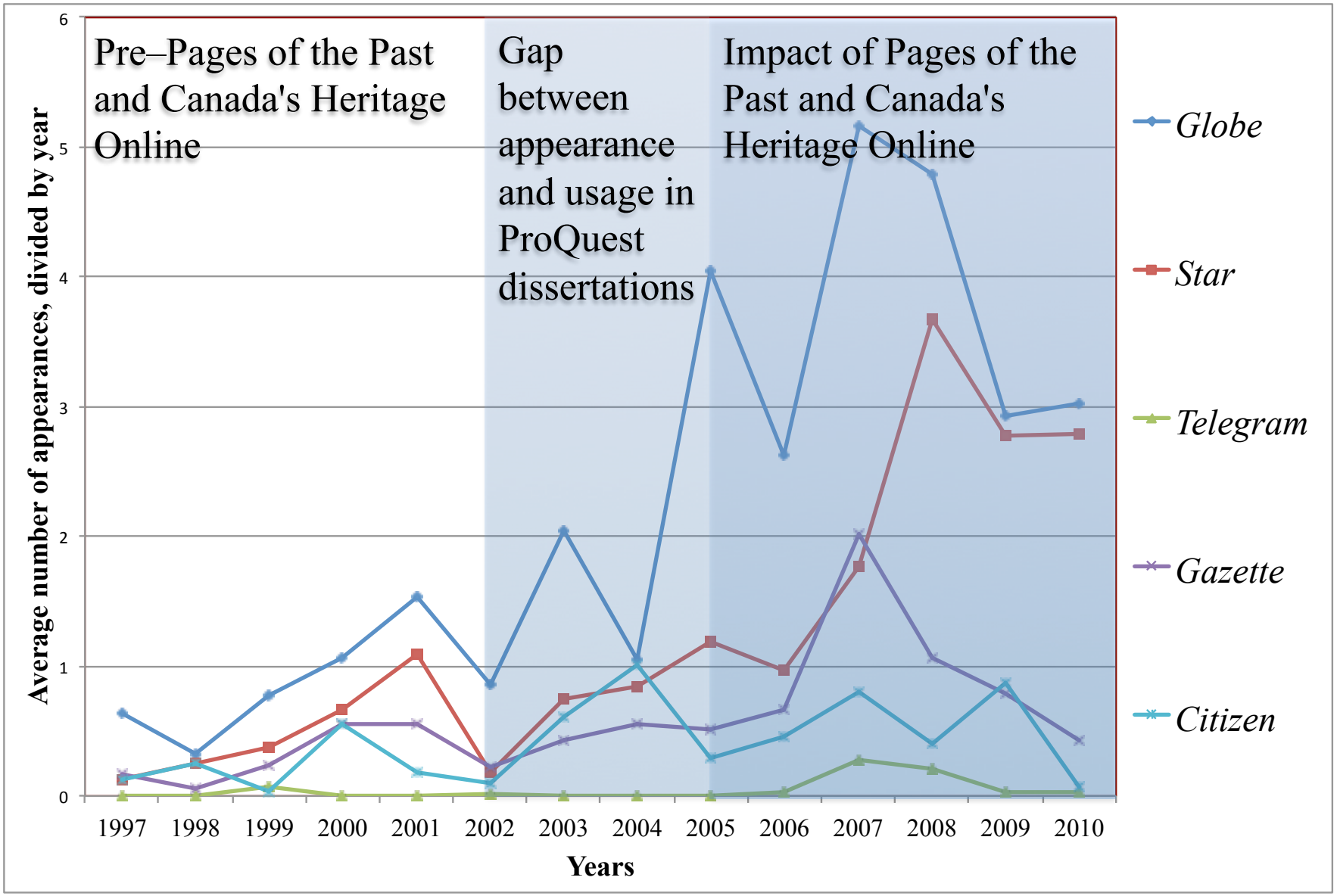

Dissertations & Theses


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

DR ONLINE

Om DR Online - English version - Tekst-version

Logo: DR Online

Radio på DR Online

Så er tiden endelig kommet, hvor DR begynder egentlige radiosendinger over Internettet. De sidste to måneder af 1996 kører vi et forsøg med nyheder og magasinstof sendt via RealAudio.

Logo: DR TV

Grevinden op tredje

Dokumentarprogrammet 'Grevinden på tredje' om Erna Hamilton blev modtaget med begejstring af seere og presse. Programmet genudsendes søndag 27.10. kl. 14.55. Læs manuskriptet, anmeldelser og instruktørens artikel fra Politiken på DR Online.

Logo: DR TV

Den afslørende hjemmeside

Rapporten - DR1's dybdeborende journalistiske program - har sin egen hjemmeside, hvor du kan finde mere information om programmets afsløringer.

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

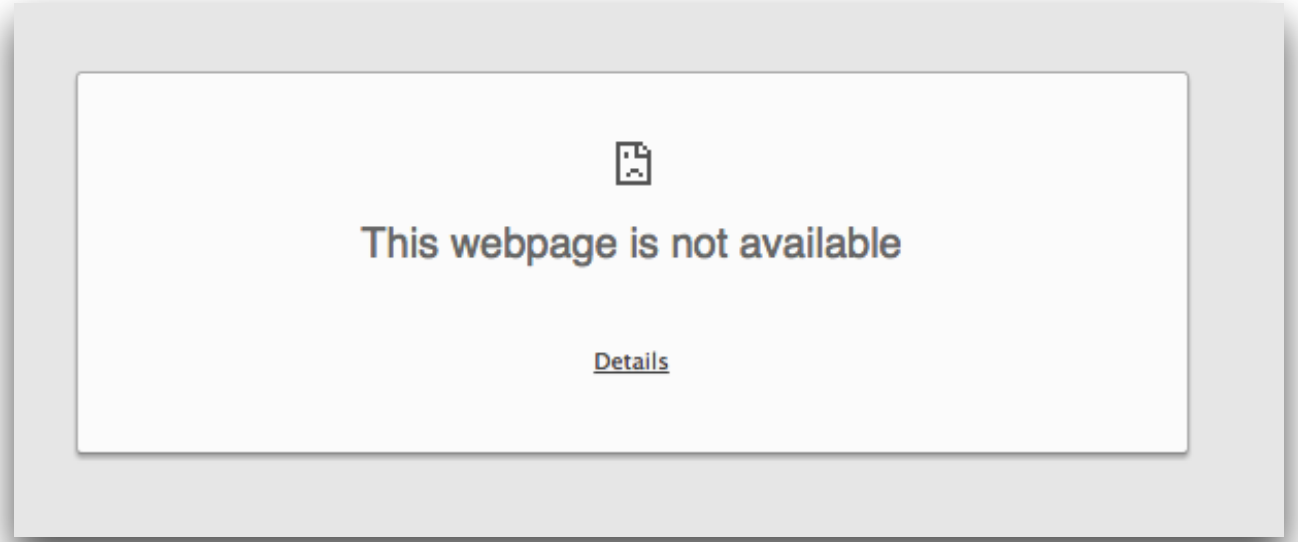
General Inquiries:
general@occupywallst.org
Press Inquiries:
press@occupywallst.org
Press Phone: +1 (347) 292-1444
Help & Directions: +1 (516) 708-4777
Watch: The world we're building
Read: This call to action
Liberty Square Eviction Defense: Text "@occupyalert" to 23559 to receive alerts in the event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

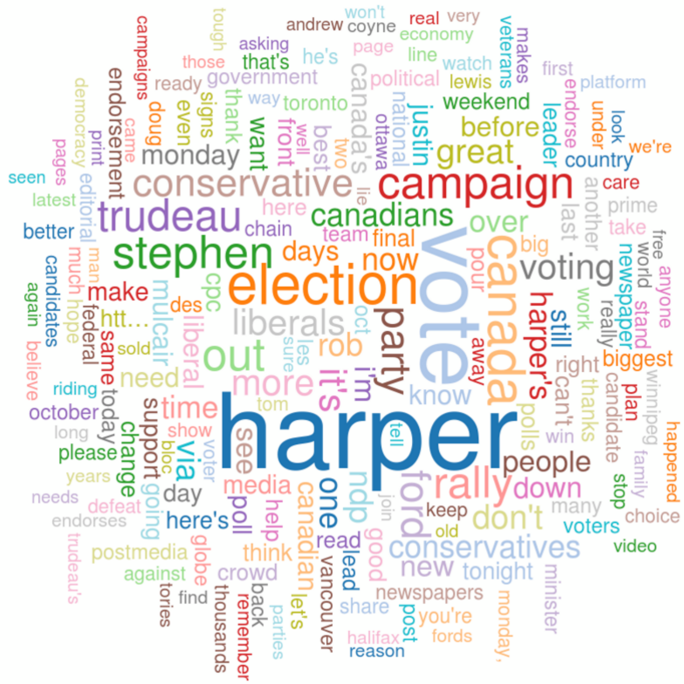
Click here for NYCGA committee meeting times.

~~Scarcity~~

Abundance

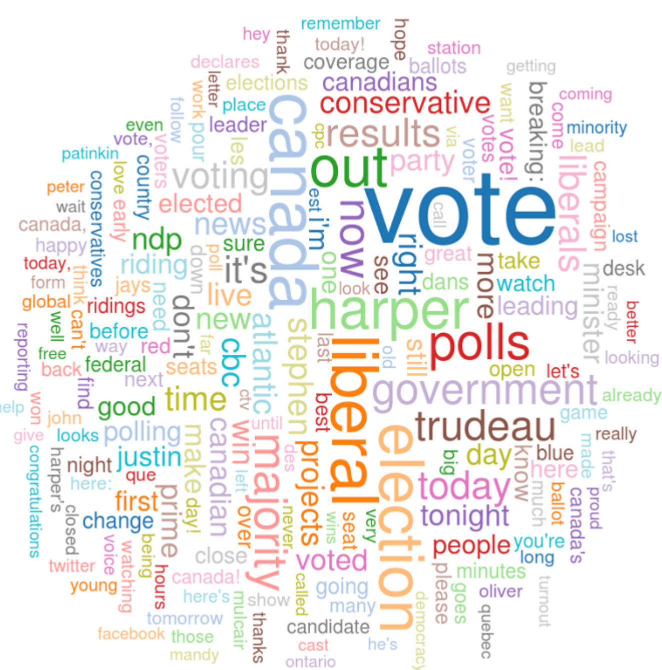

RIC'S GRILL
STEAK SEAFOOD & CHOP HOUSE
Wharf
IDLE NO MORE

October 17, 2015



October 18, 2015

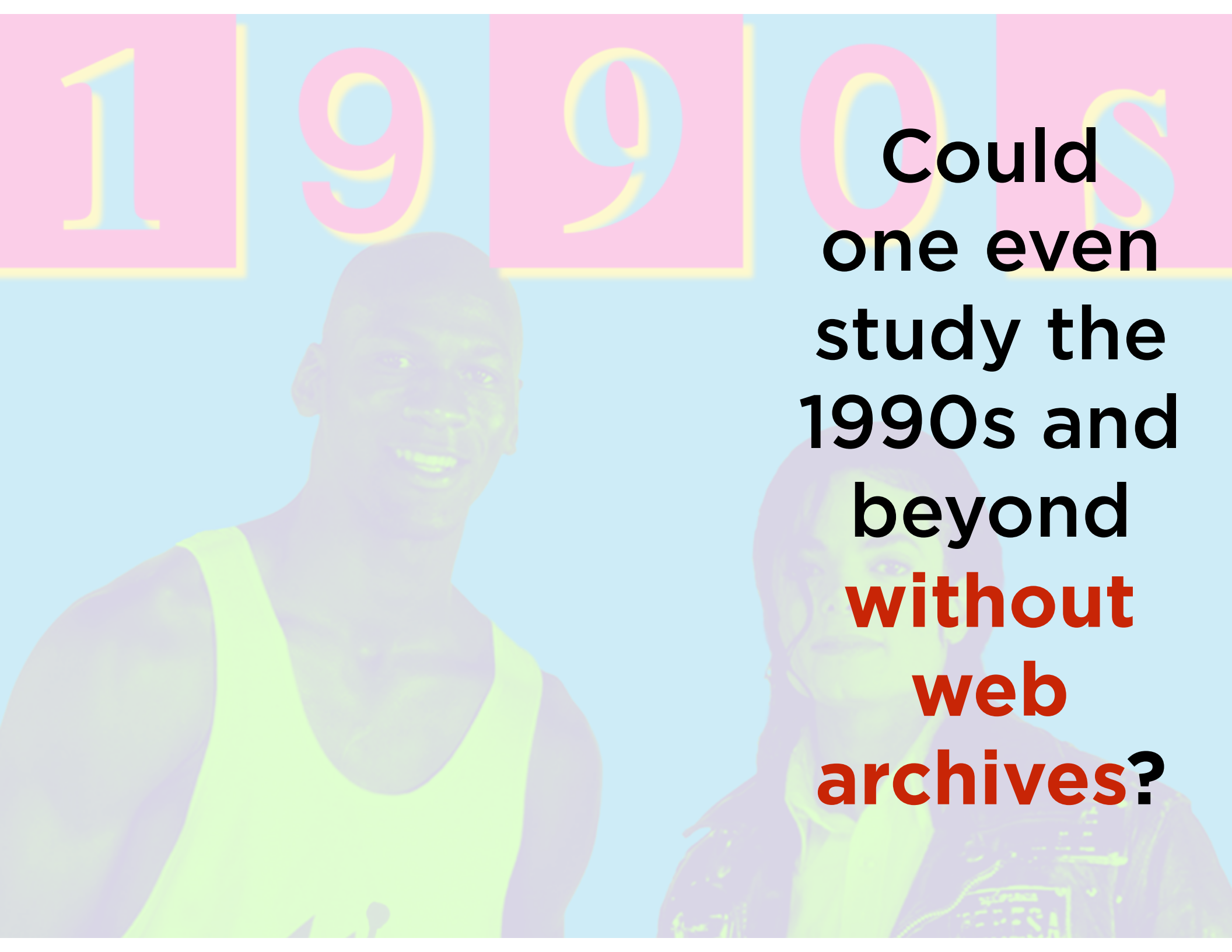
October 19, 2015



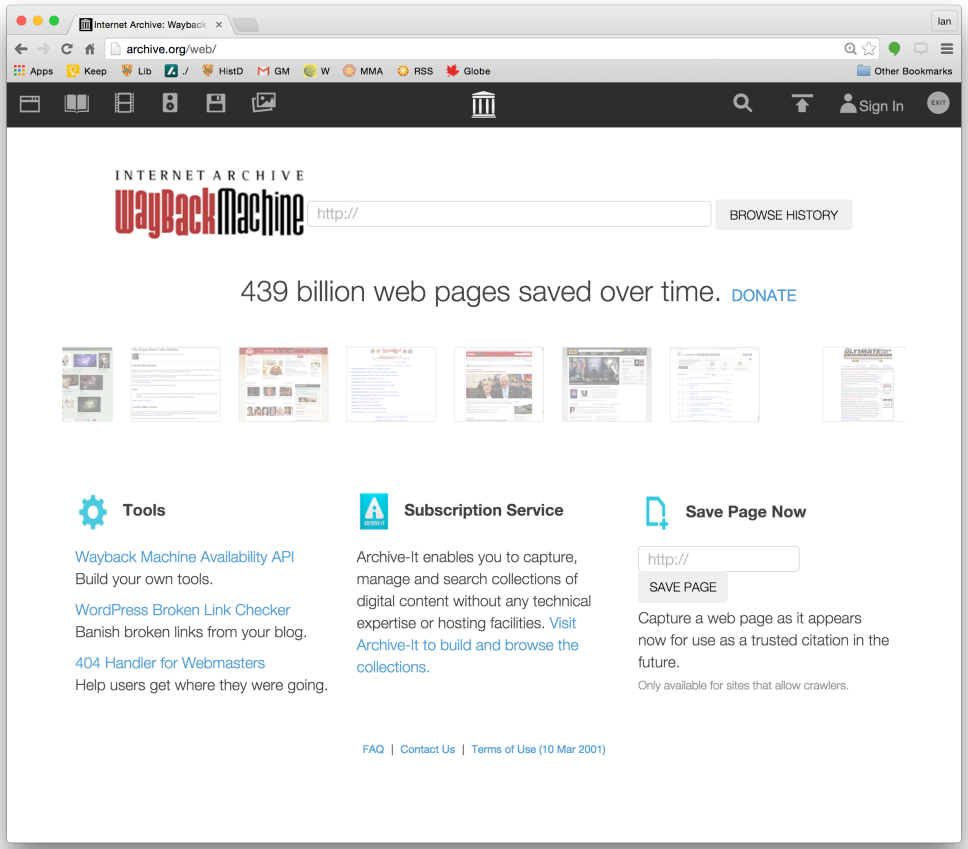
October 20, 2015

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

- James Gleick

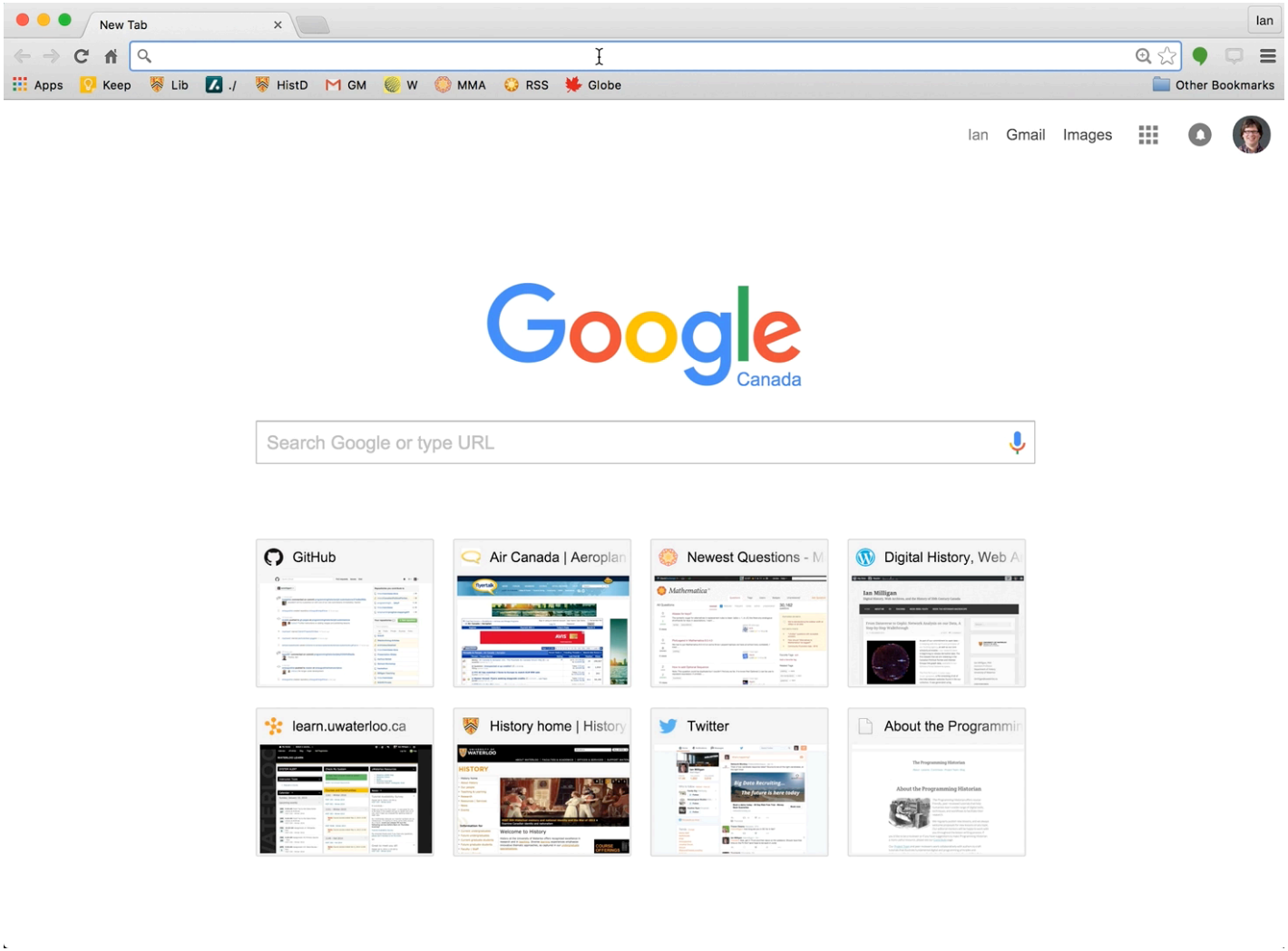

1990s
Could one even study the 1990s and beyond without web archives?

Nightmare Scenario



This won't be enough!

Nightmare Scenario



This won't be enough!

… but what will our search engines look like?

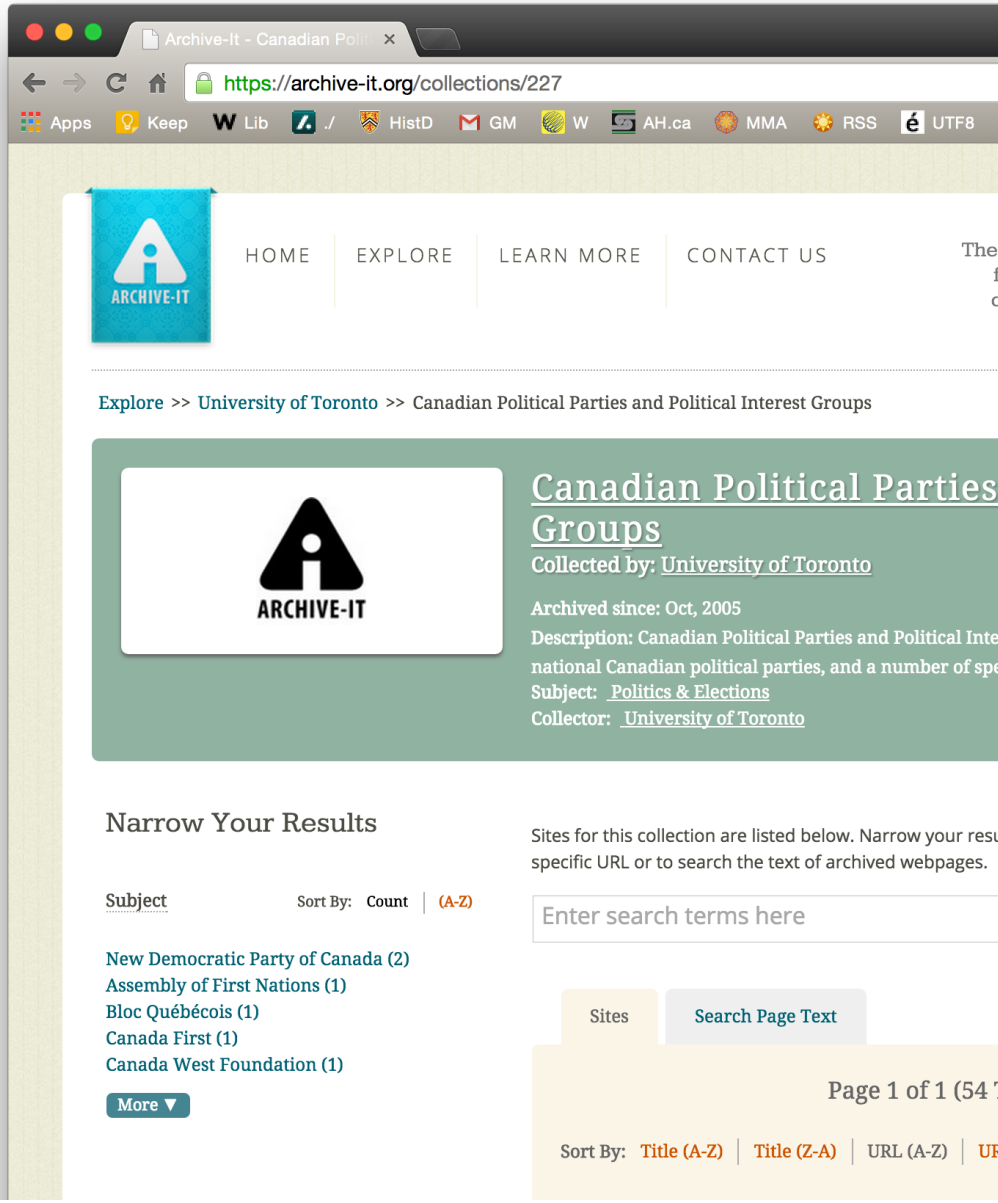
Nightmare Scenario

- Historians rely uncritically on **date-ordered keyword search results**, putting them at mercy of search algorithms they do not understand (similar to digitized newspapers);
- Historians are completely left out of post-1996 research, letting everybody else do the work (a la Culturomics project/*Nature* magazine article);
- Our profession gets left behind…

What can we do to access this information?

Building Portals

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

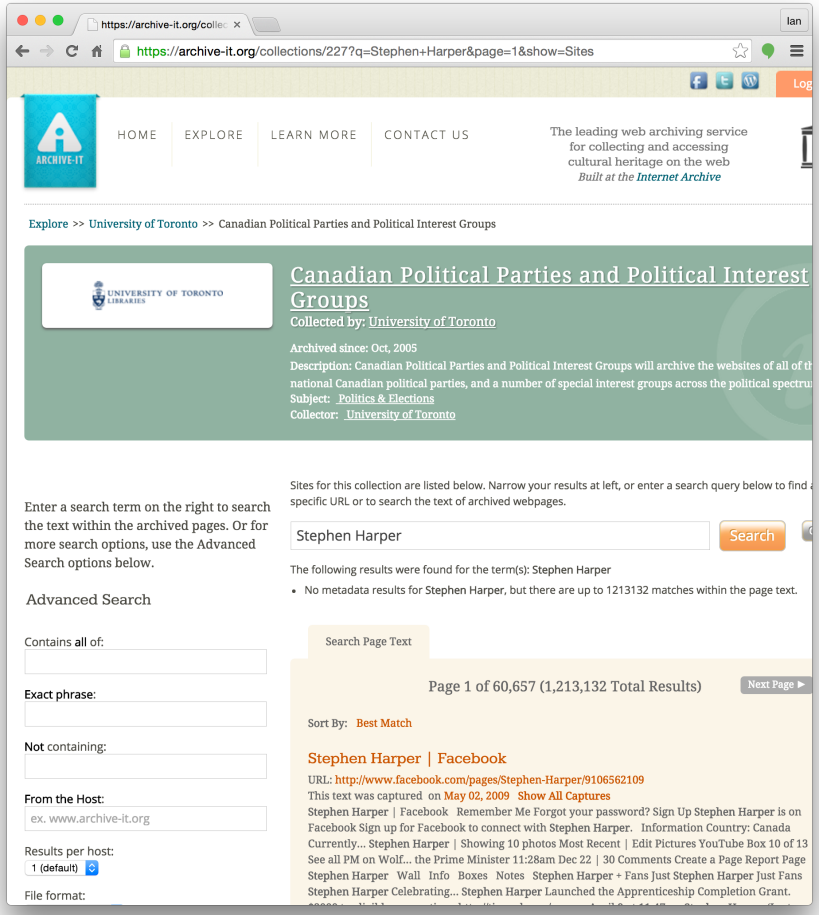


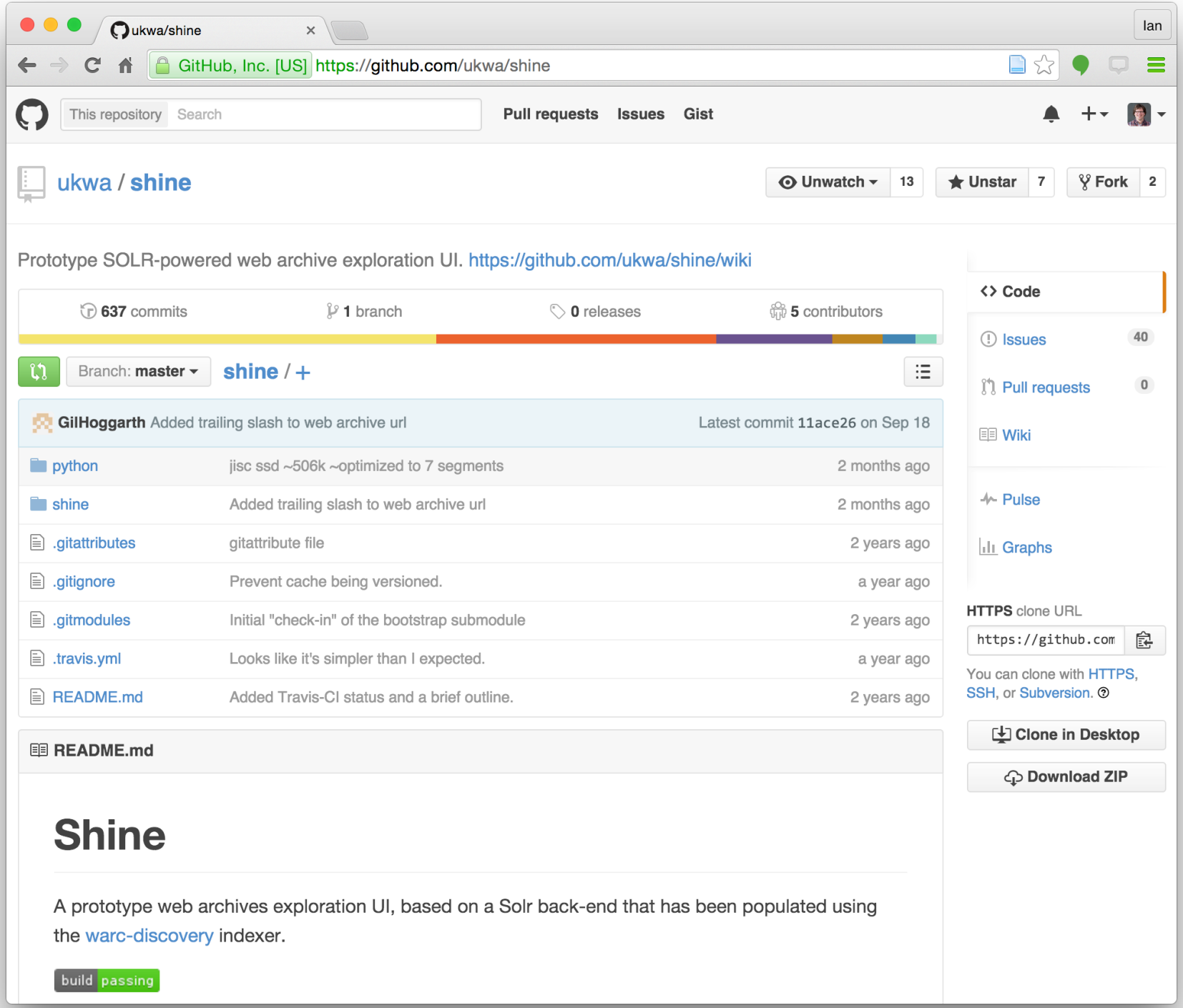
Pivotal Changes in Canadian Politics, 2005-2015

- Militarization of Canadian society?
- Change from ‘natural governing party’ of Liberals to Conservatives
- Major policy changes on foreign policy, environment, etc.
- How to measure?

Current Interface

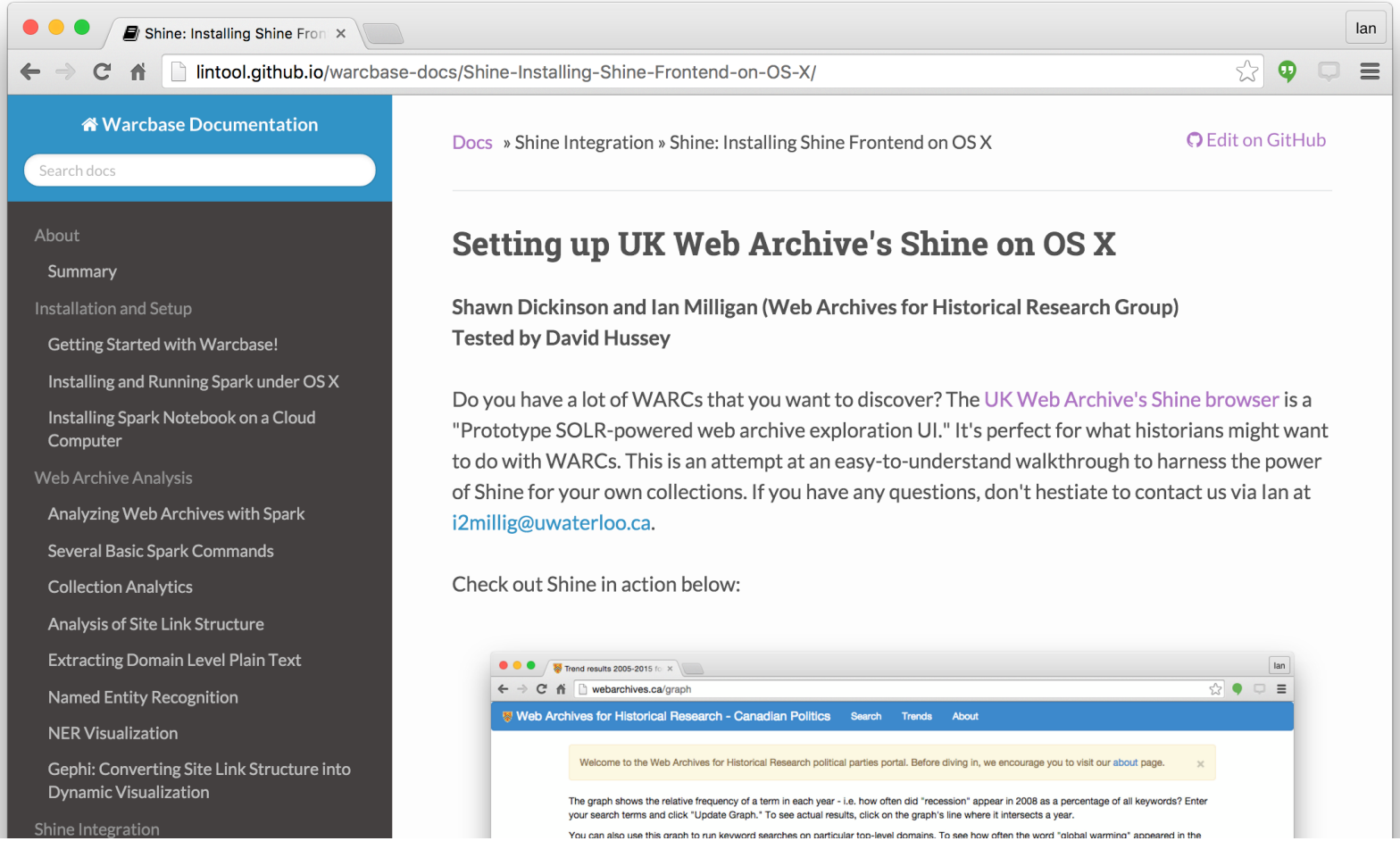
- Very limited - simple search engine, some advanced options; no facets
- Great collections.. but nobody uses them!




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

Walkthroughs at:

http://lintool.github.io/warcbase-docs/Shine-Installing-Shine-Frontend-on-OS-X/



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

+

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

+

Ope Source Software

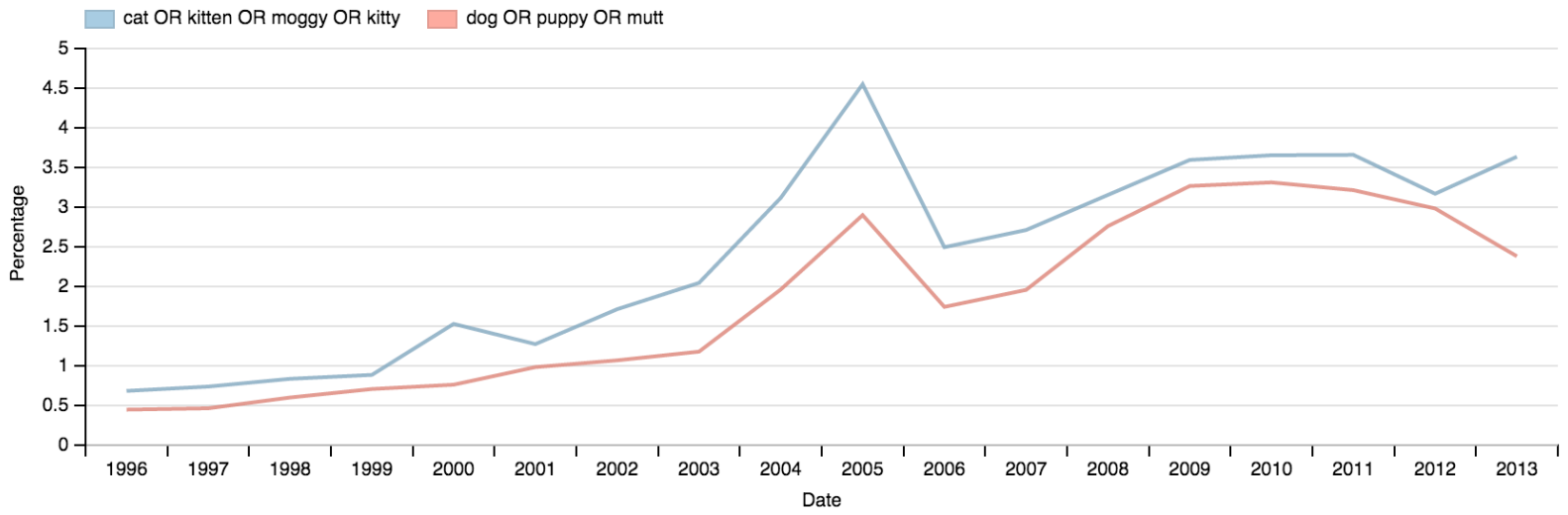
Trend results 1996-2013 fo

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

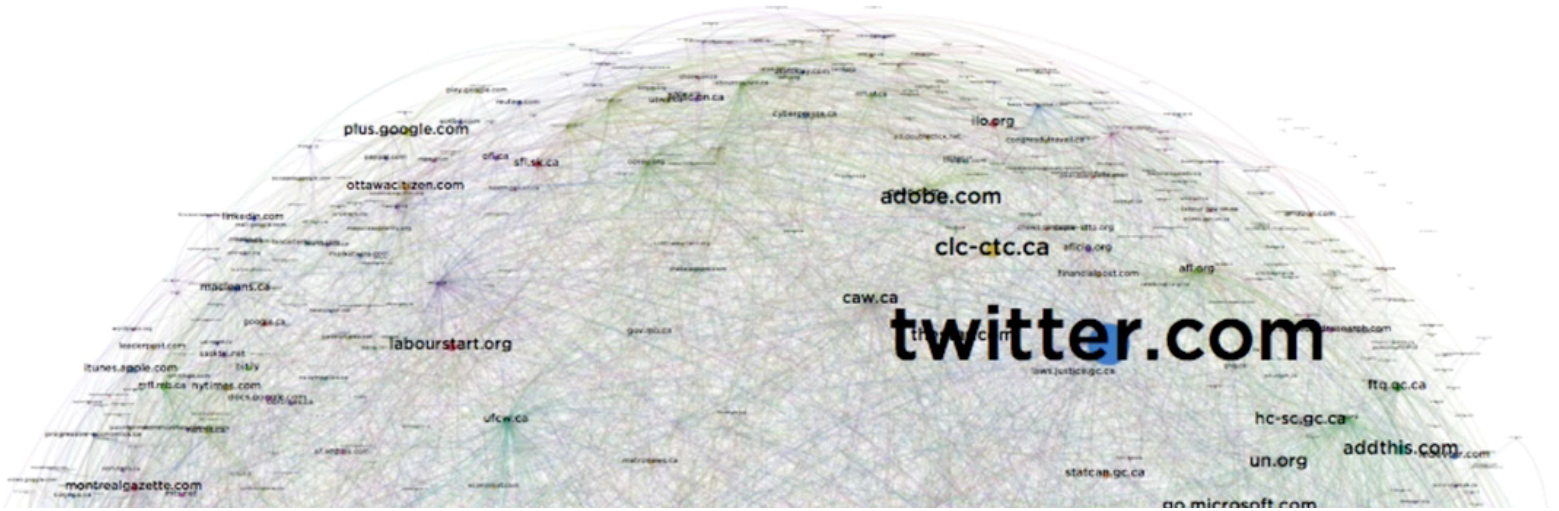
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

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]

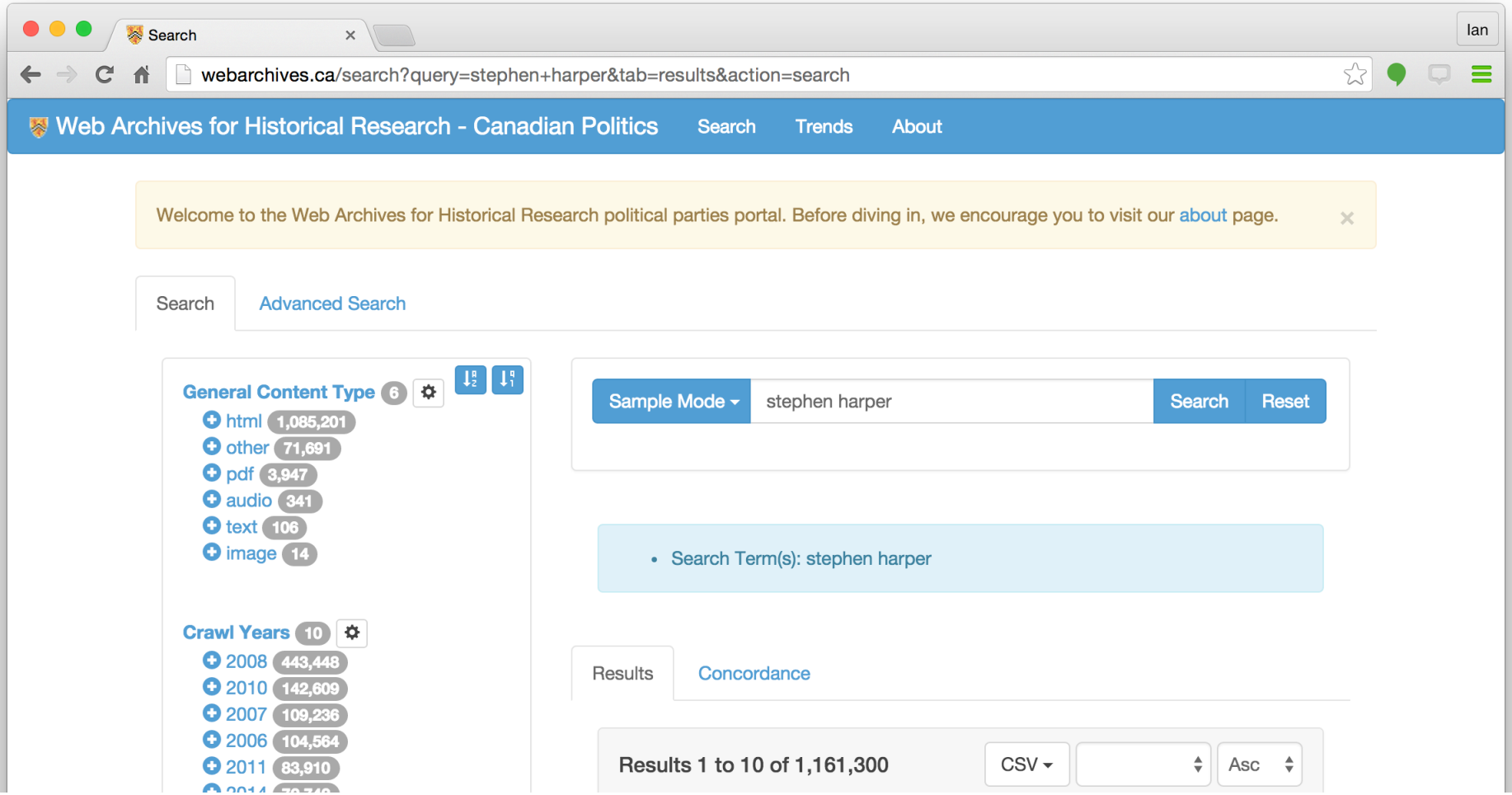
Below, here are all of the links for the entire time period, visualized below.



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

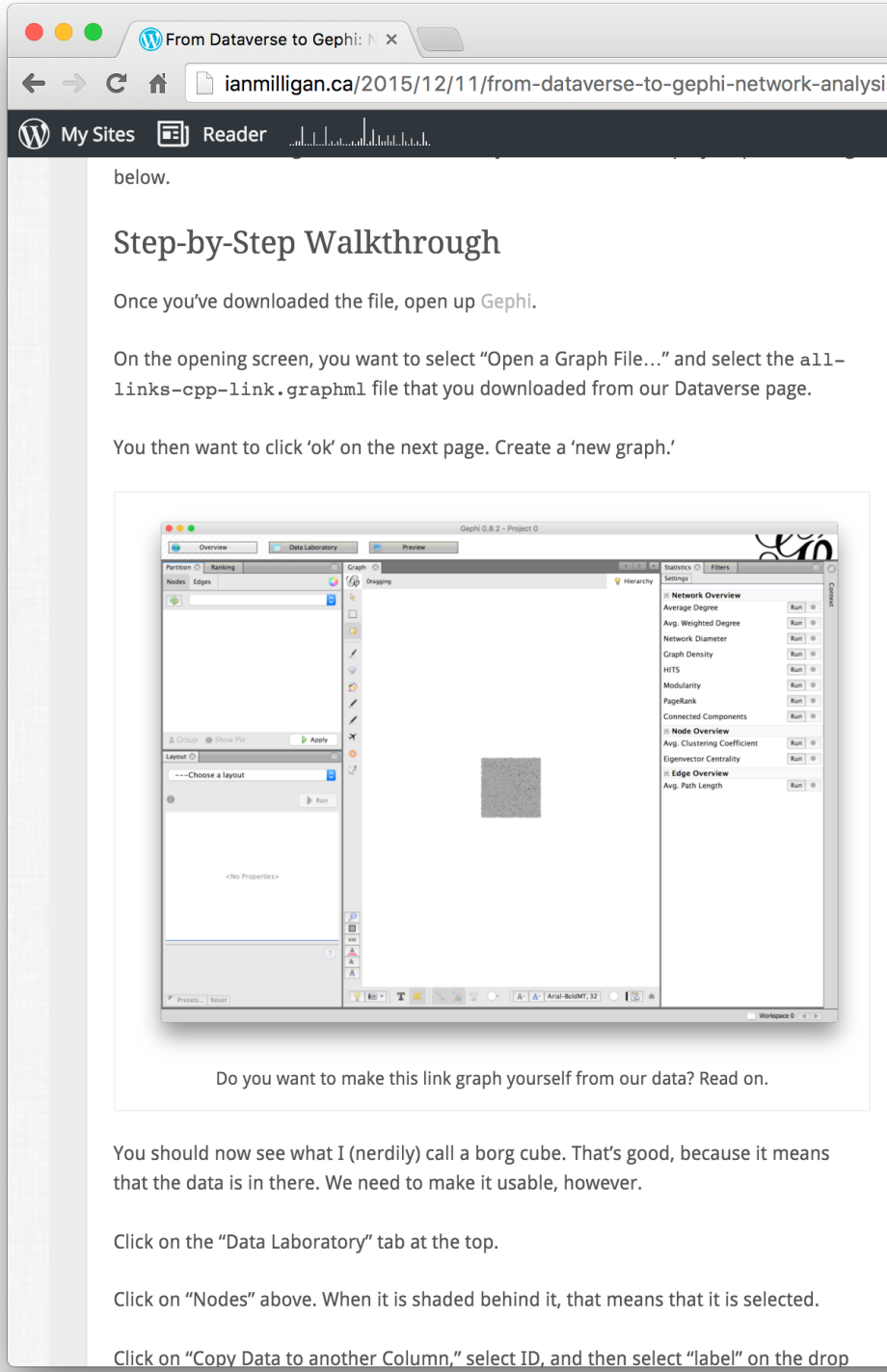
Good for public engagement - but limited for scholarship....



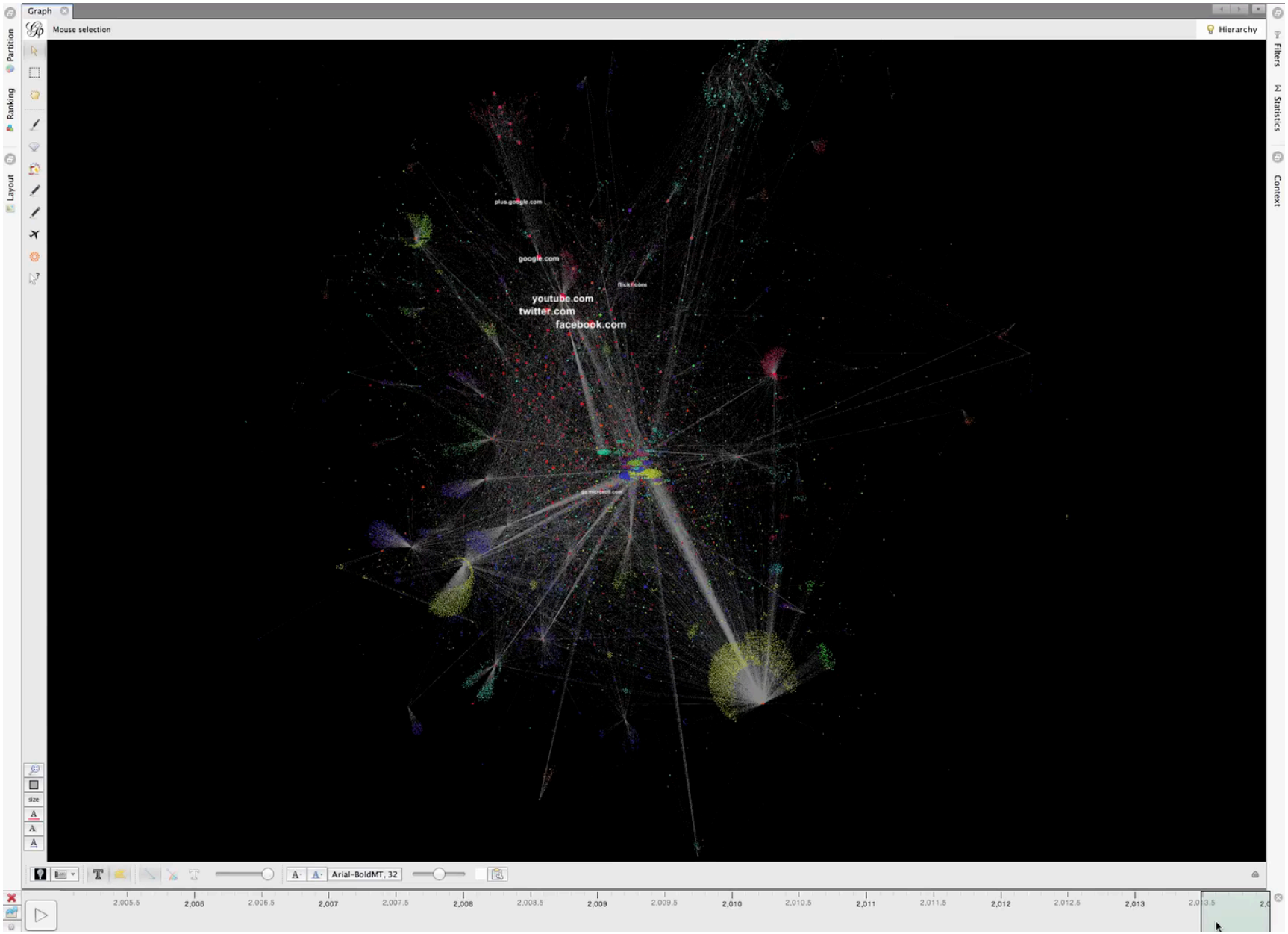
Getting over my bias towards content **and embracing metadata**

Gephi 0.9 (http://gephi.github.io/)

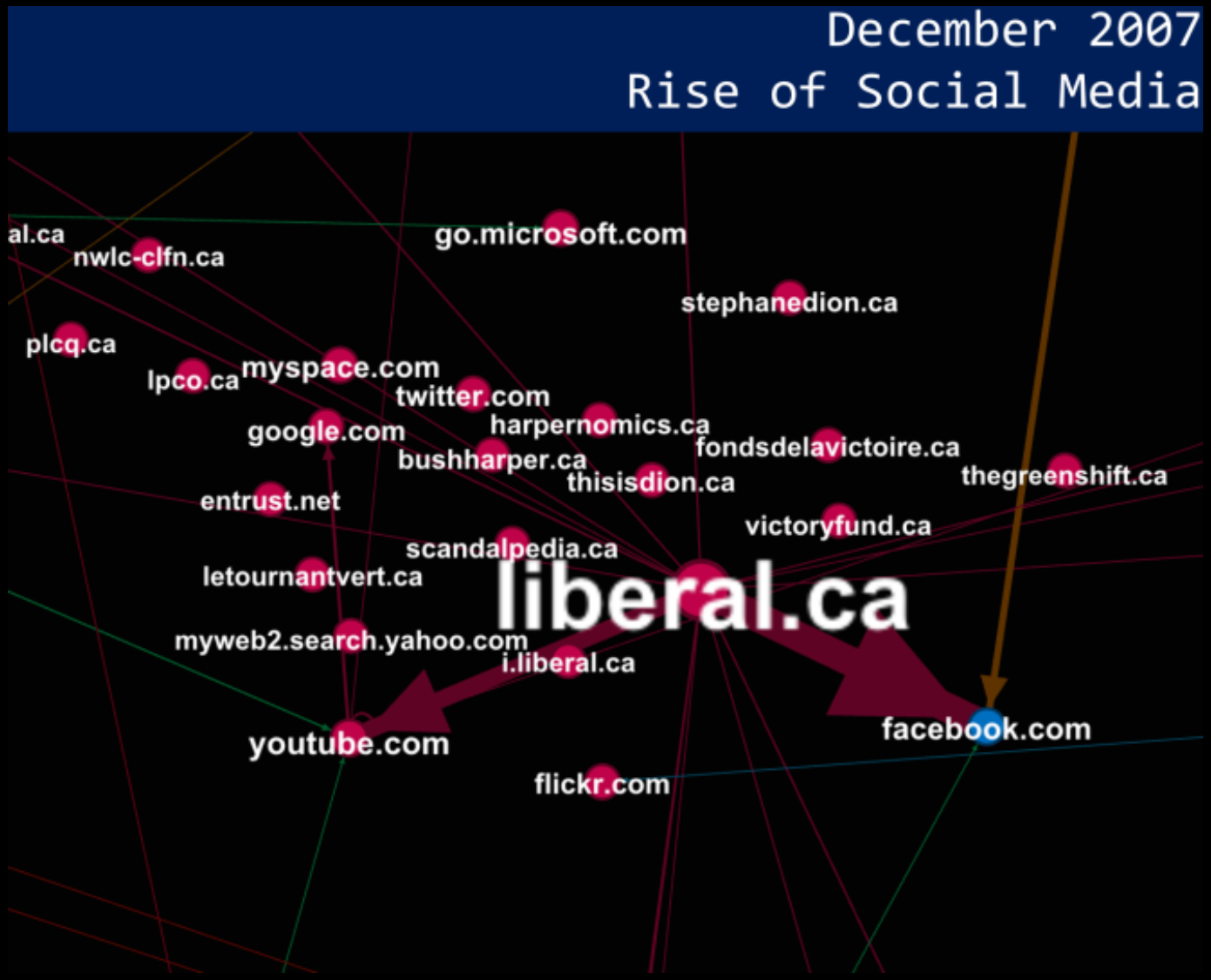
Walkthrough at ianmilligan.ca: "From Dataverse to Gephi" - try it on this data!



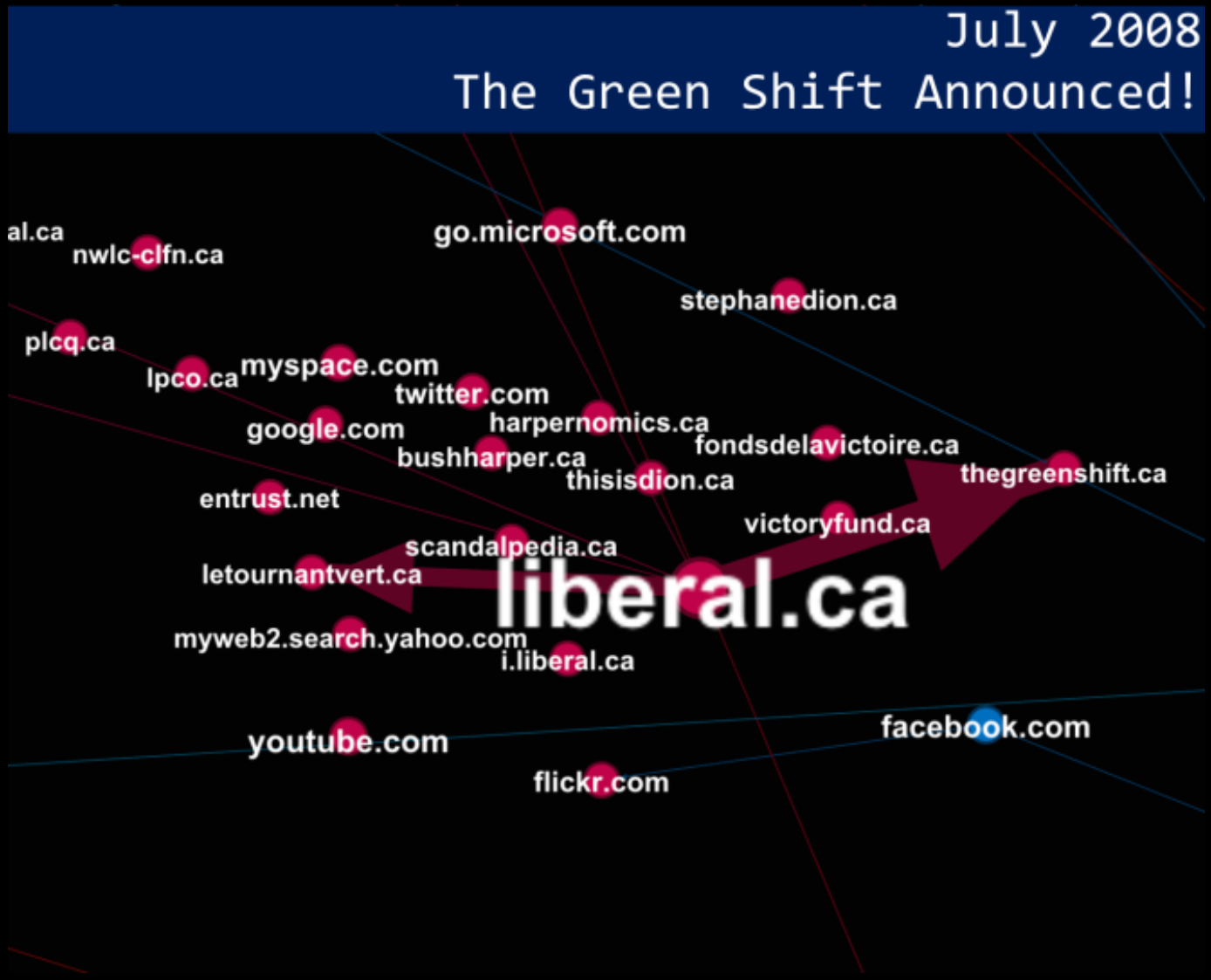
Metadata Extraction

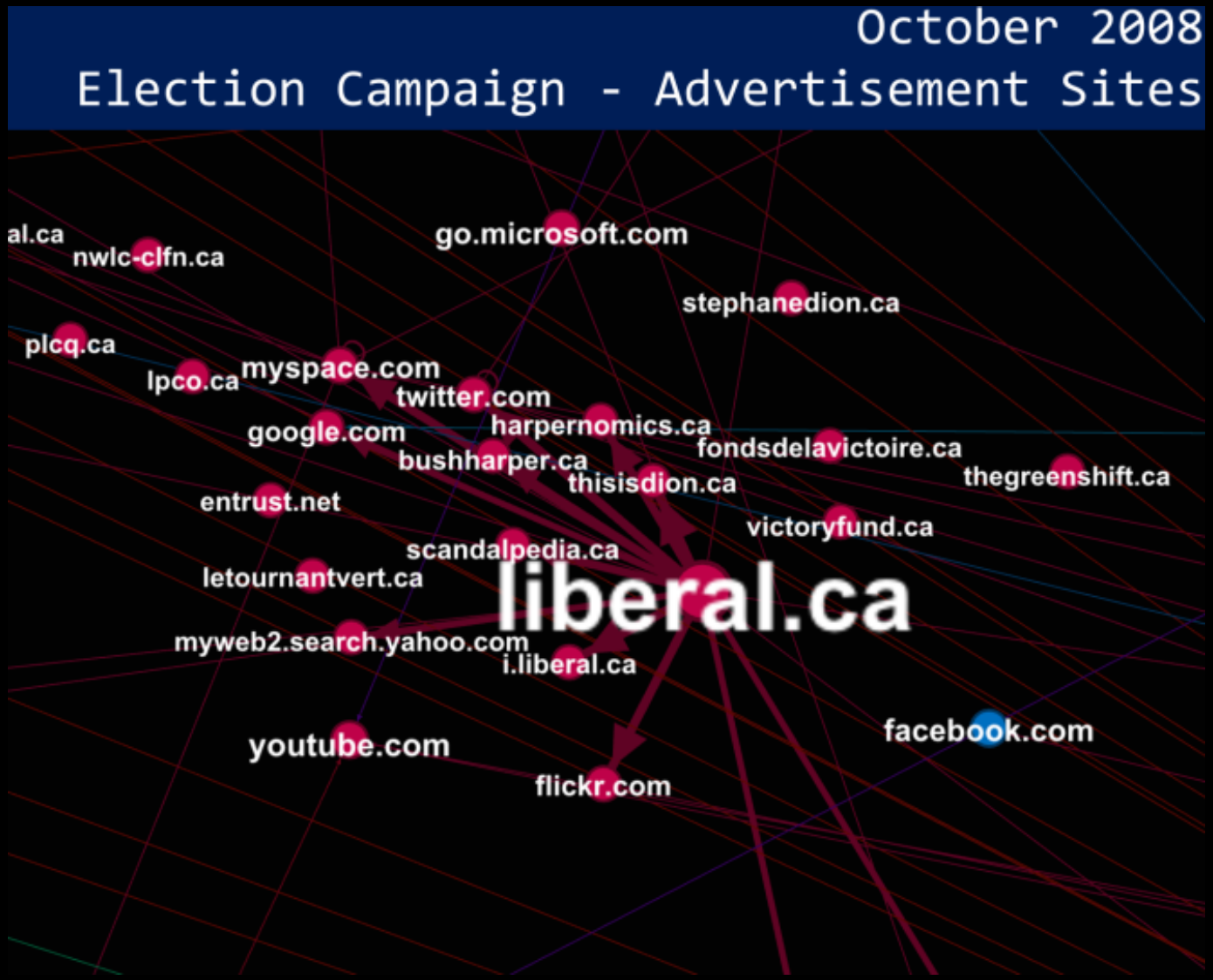



December 2006
Stephane Dion Elected Leader of Party
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca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

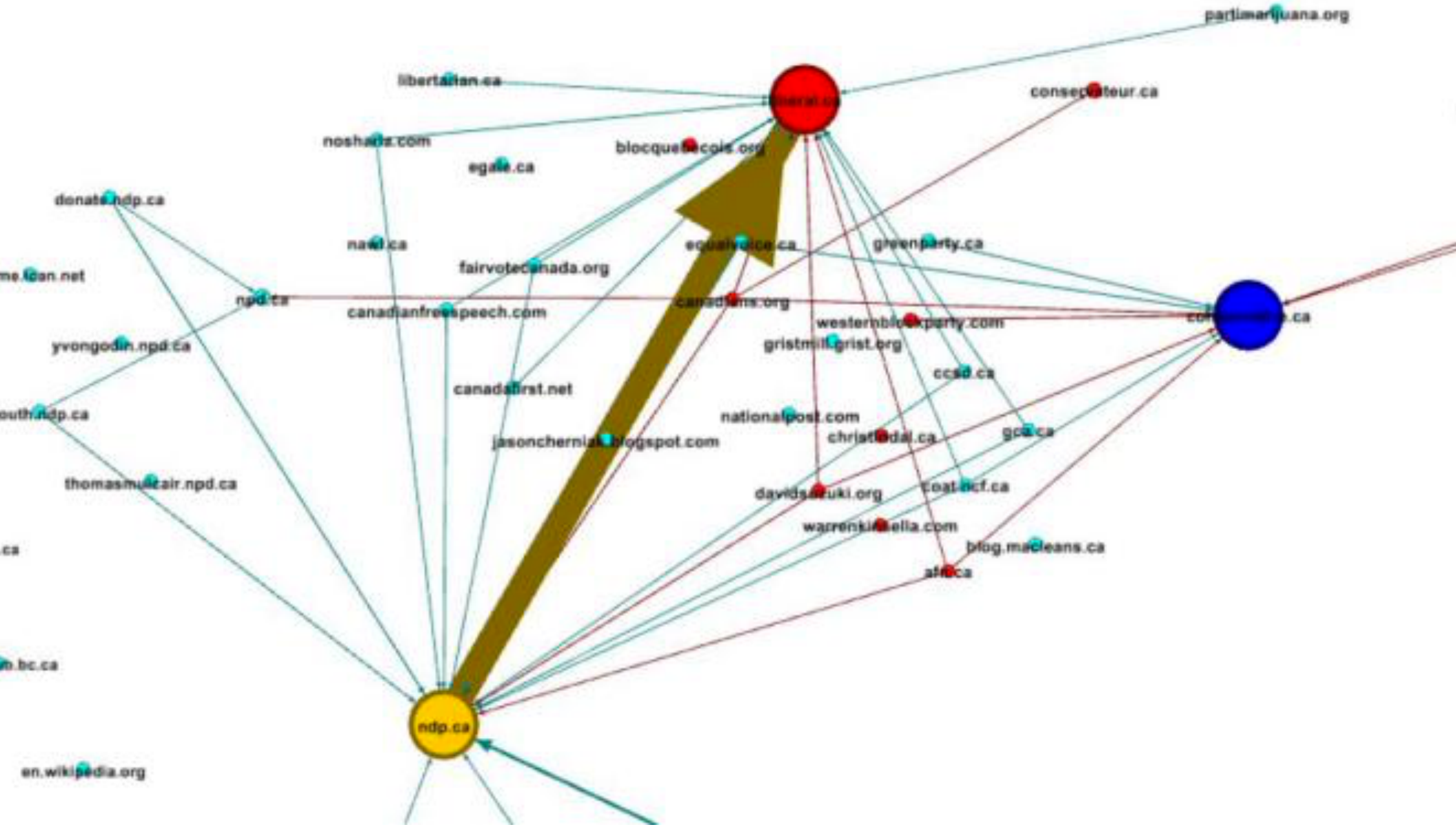

April 2008
Fundraising with the Victory Fund/
Fonds de la Victoire
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca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

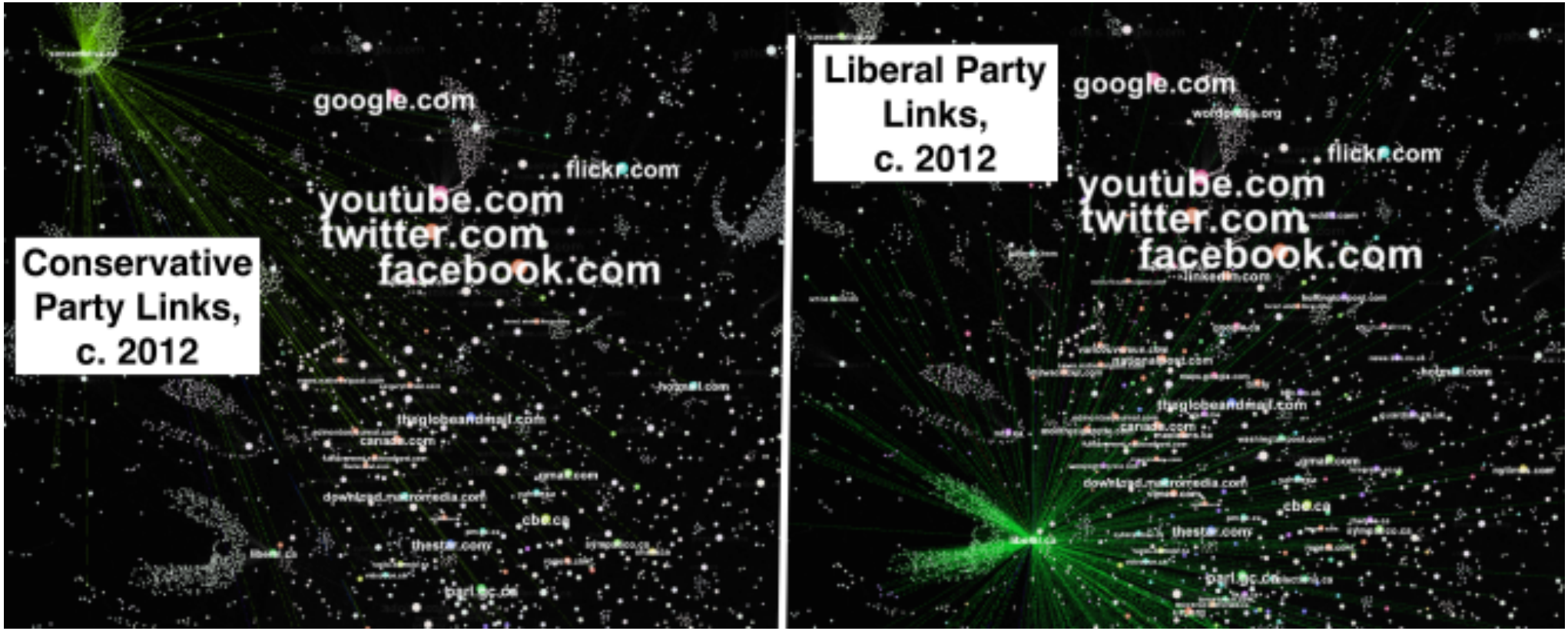

October 2008
Election Campaign - Advertisement Sites
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca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

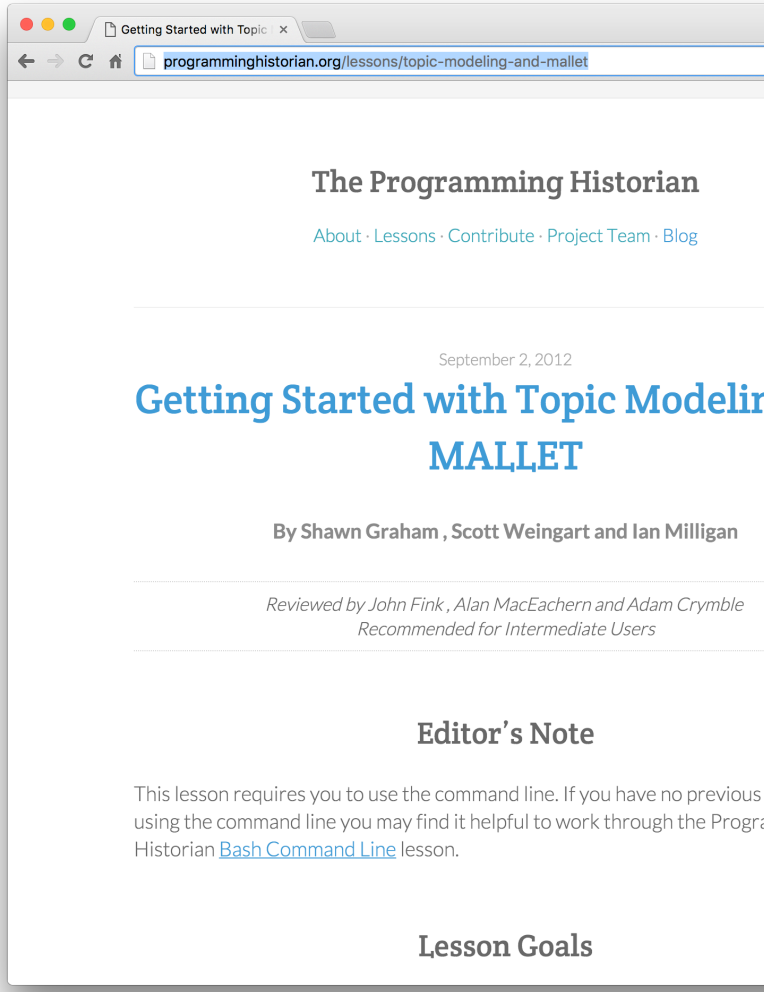


Metadata Extraction



Topic Modelling using MALLET (http://mallet.cs.umass.edu/)

Walkthrough: http://programminghistorian.org/lessons/topic-modeling-and-mallet



Metadata Extraction

liberal.ca	27
liberal.ola.org	27
liberal.us1.list-manage.com	27
liberal.us1.list-manage1.com	27
liberal.us1.list-manage2.com	27
liberaluniversity.liberal.ca	27
license.icopyright.net	27
live.cbc.ca	27
lpc.ca	27
macleans.ca	27
masses.tao.ca	27
mcss.gov.on.ca	27
mediaignite.com	27
mediasales.cbc.ca	27
membercentre.cbc.ca	27
mentalhealthcommission.ca	27
metrics.mmailhost.com	27
mondesdesfemmes.ca	27
music.cbc.ca	27
nawl.ca	27
newswire.ca	27
nowtoronto.com	27
npd.ca	27

colincarriemp.ca	12
colincarriemp.ca&lang=fr	12
colinmayes.ca	12
colinmayes.ca&lang=fr	12
congrespcc.ca	12
conservateur.ca	12
conservateur.us5.list-manage.com	12
conservative.ca	12
conservative.us5.list-manage.com	12
consumersfirst.ca	12
corneliuchisu.ca	12
corneliuchisu.ca&lang=fr	12
costasmenegakis.ca	12
costasmenegakis.ca&lang=fr	12
cpcconvention.ca	12

Metadata Extraction

- Conservative themes (2014): economic development, family, immigration, legislation, women's issues, senior issues, Ukrainians, constituency offices, some prominent (and not-so-prominent) MPs, and of course, our economic action plan.

- Liberal themes (2014): Justin Trudeau (the new leader), cuts to social programs, child poverty, mental health, municipal issues, labour, workers, Stop the Cuts, and housing.

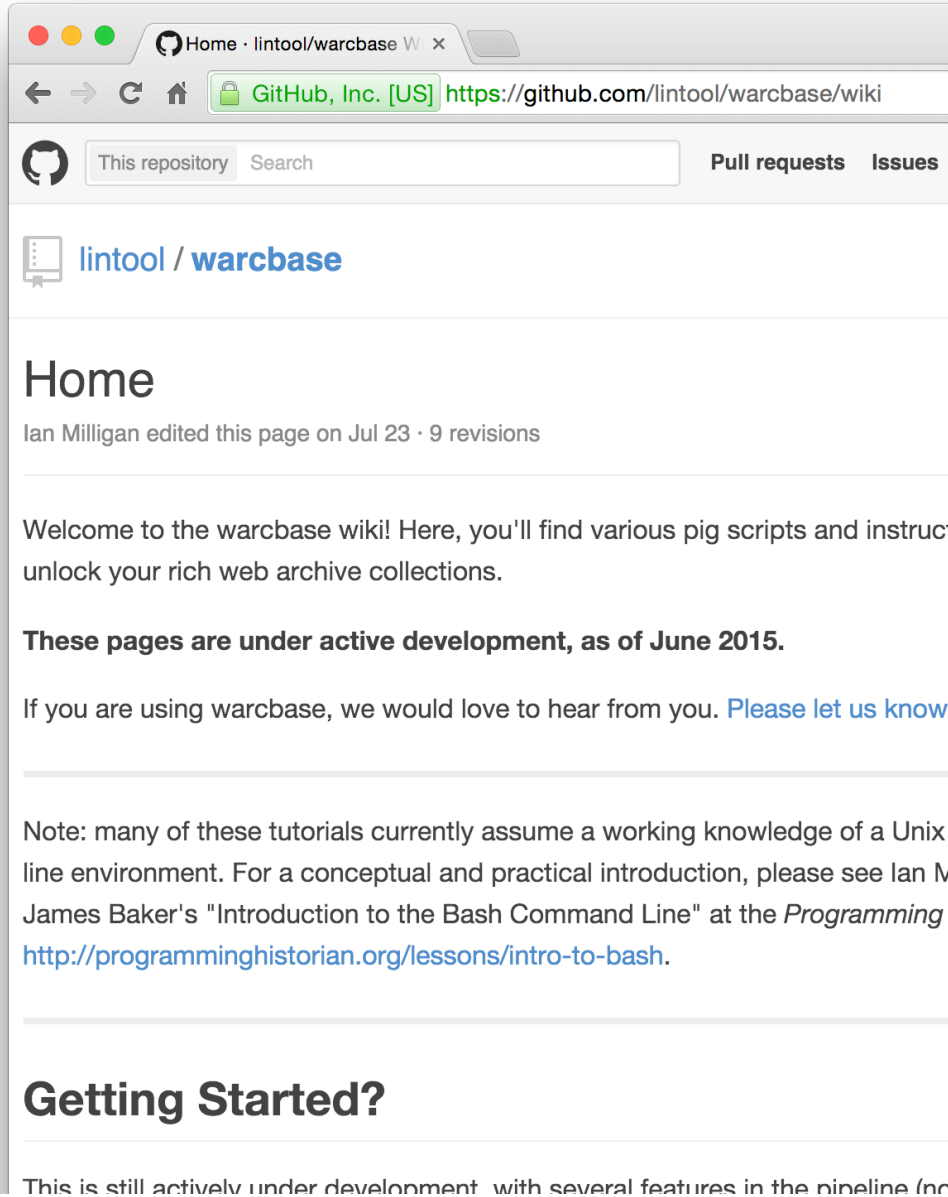
Metadata Extraction

- Conservative themes (2006): education, university, but **tons of information on Aboriginal issues**;
- Liberal themes (2006): community questions, electoral topics, universities, human rights, child care support.

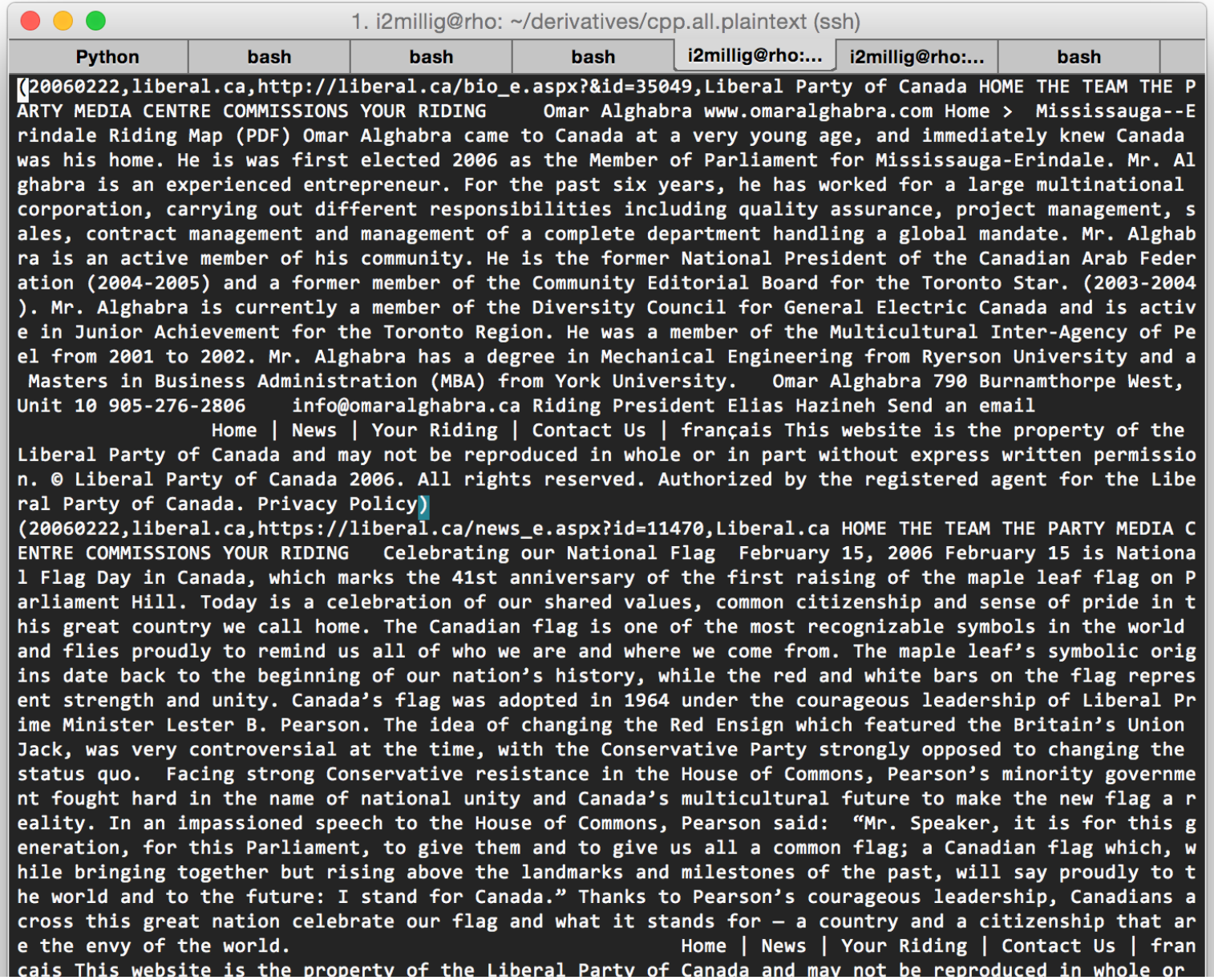
Interdisciplinary

Warcbase

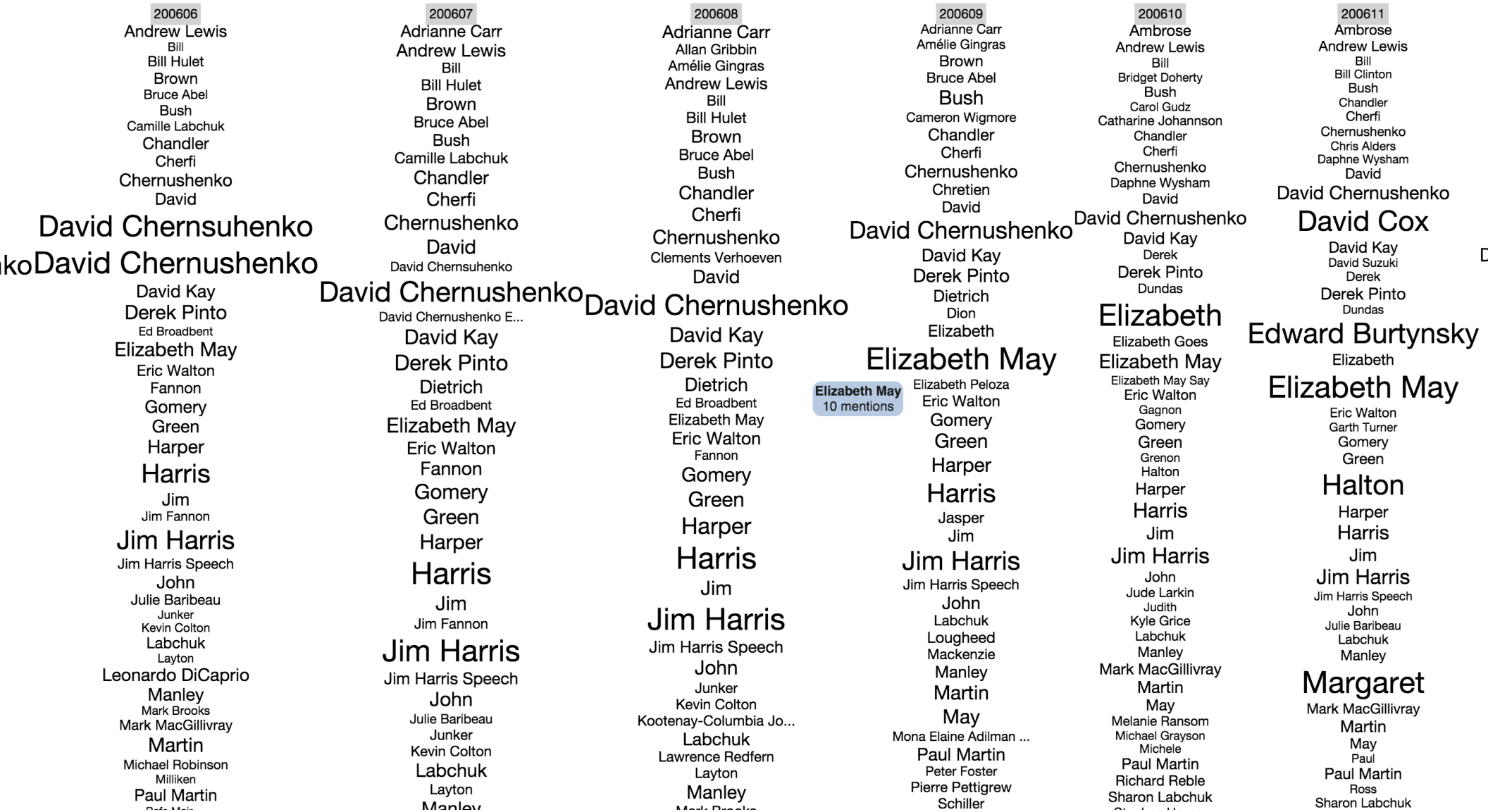
- Jimmy Lin (main developer), Ian Milligan, Jeremy Wiebe, Alice Zhou, all University of Waterloo
- Developing code, walkthroughs, and instructions for historians to be able to take a directory of WARCs and…



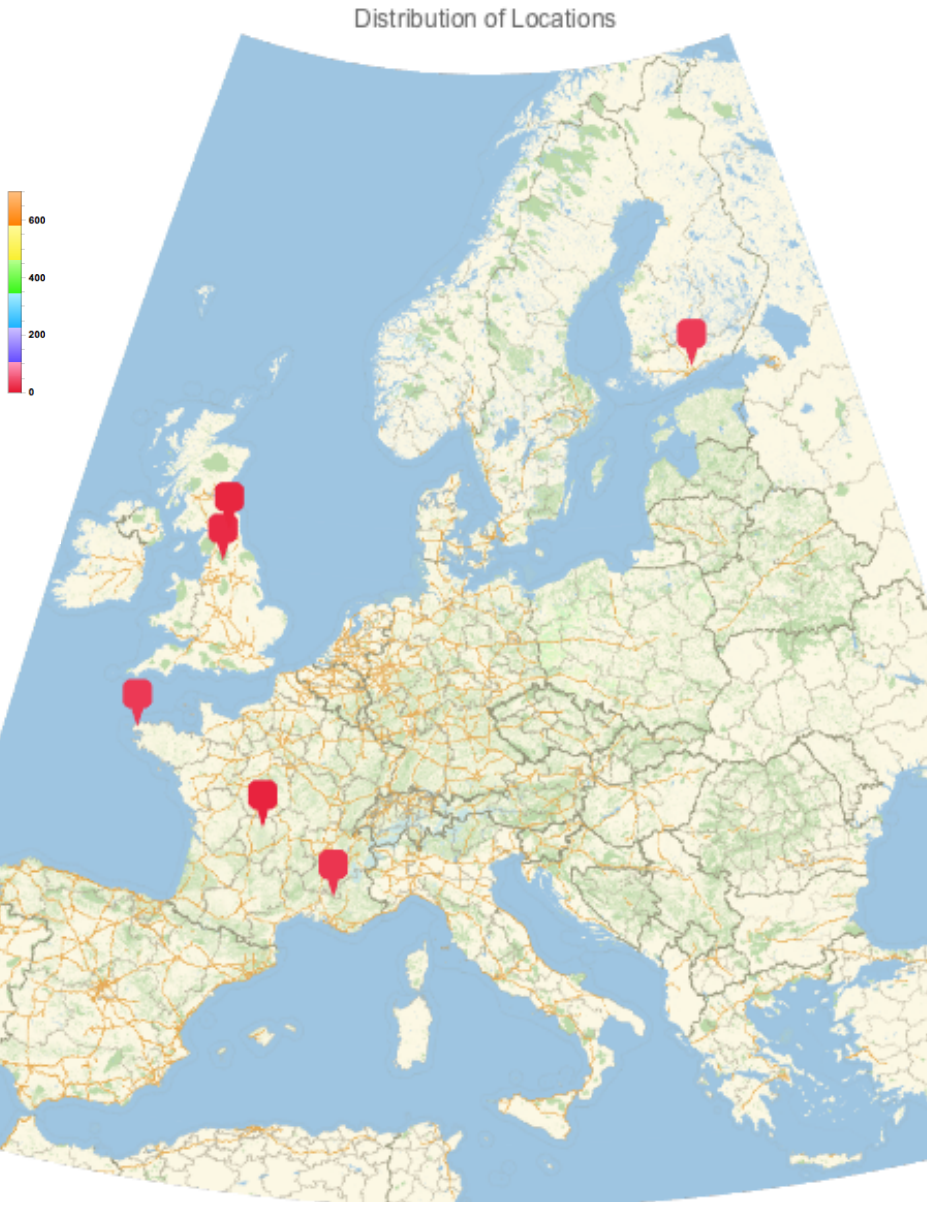
Extract all Plain Text



Extract Entities



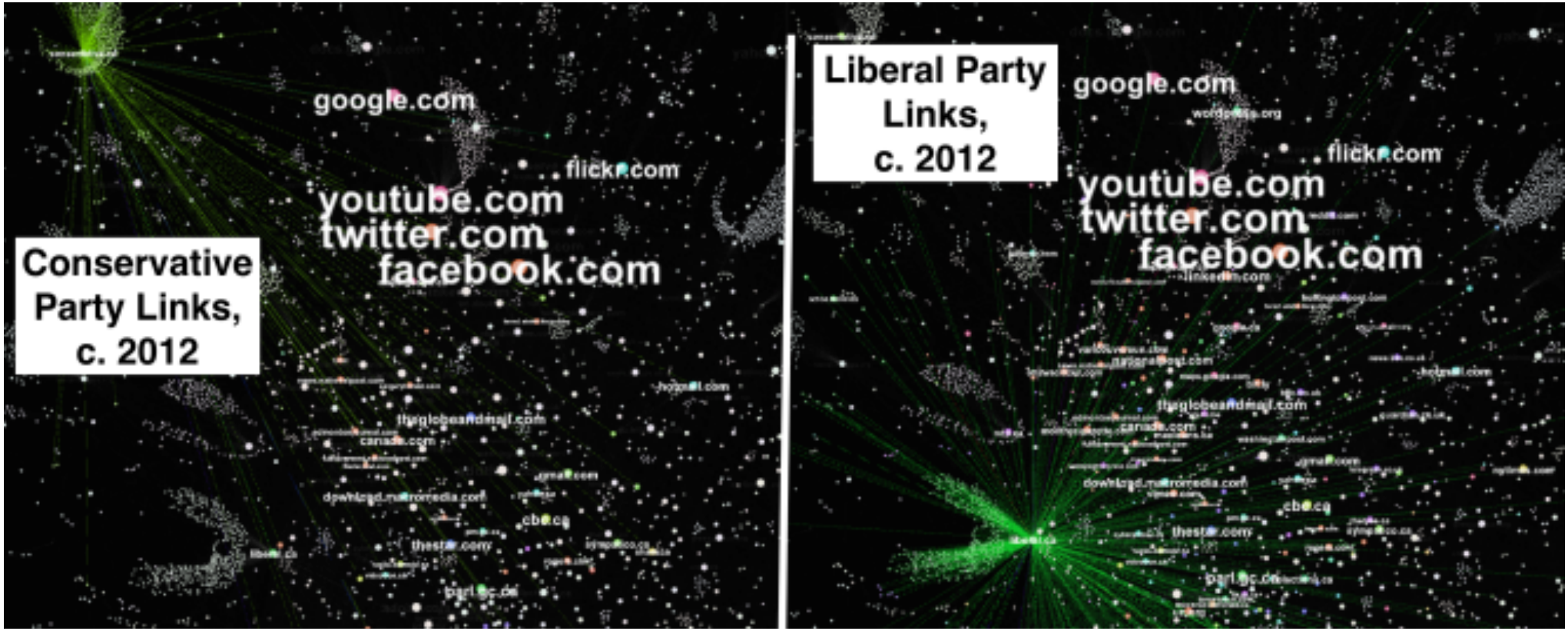
Extract Entities



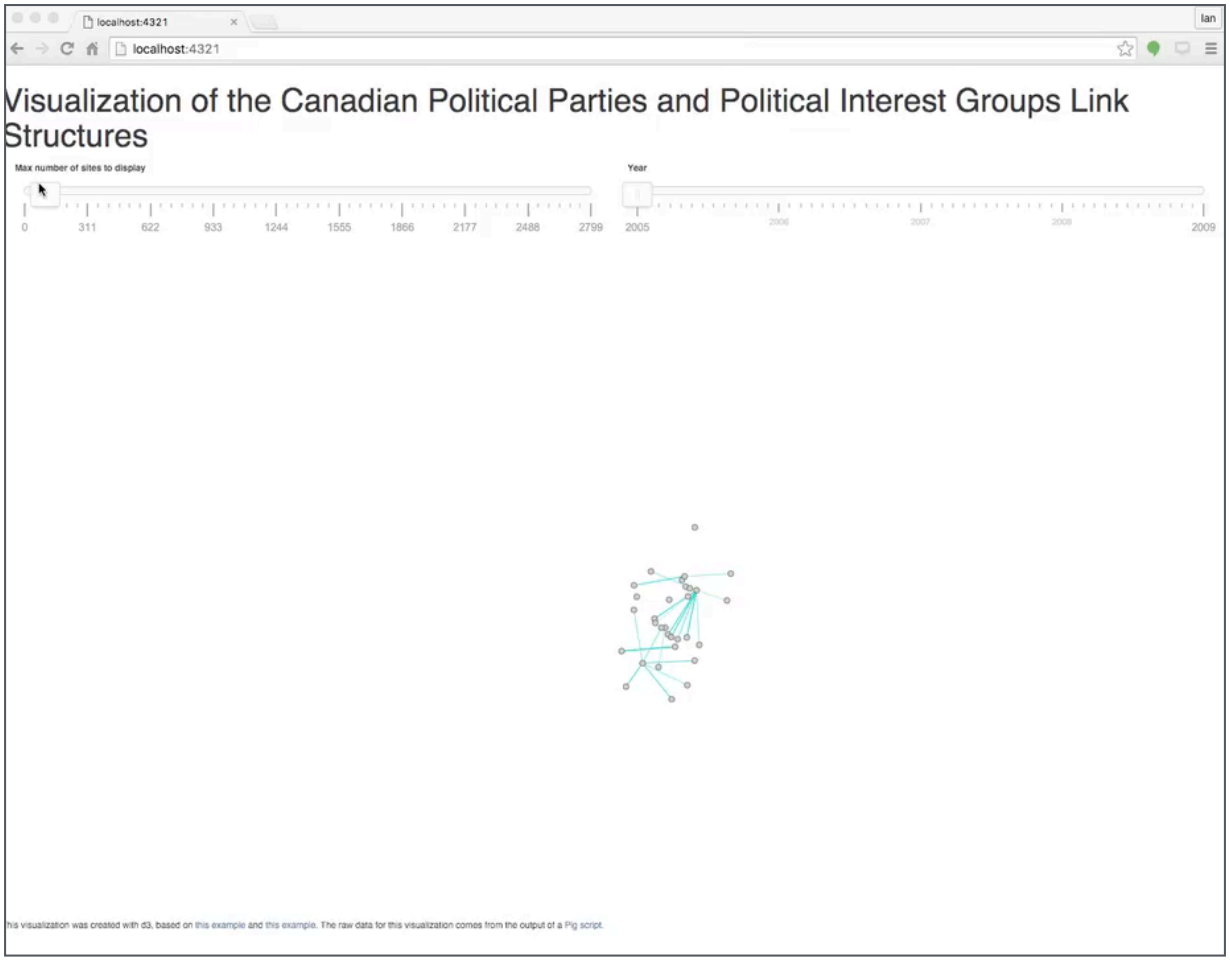

```
In[95]:= int = SemanticInterpretation[#] & /@ processedfreq[[All, 1]]
Out[95]= {Canada, Calgary, Colombia, Montreal, $Failed, Ontario, Canada, Afghanistan,
 Ottawa, Manitoba, Canada, British Columbia, Canada, Toronto, Nova Scotia, Canada,
 Saskatchewan, Canada, Quebec, Canada, Alberta, Canada, Winnipeg, 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 Entities

Extract Links/Gephi Connector

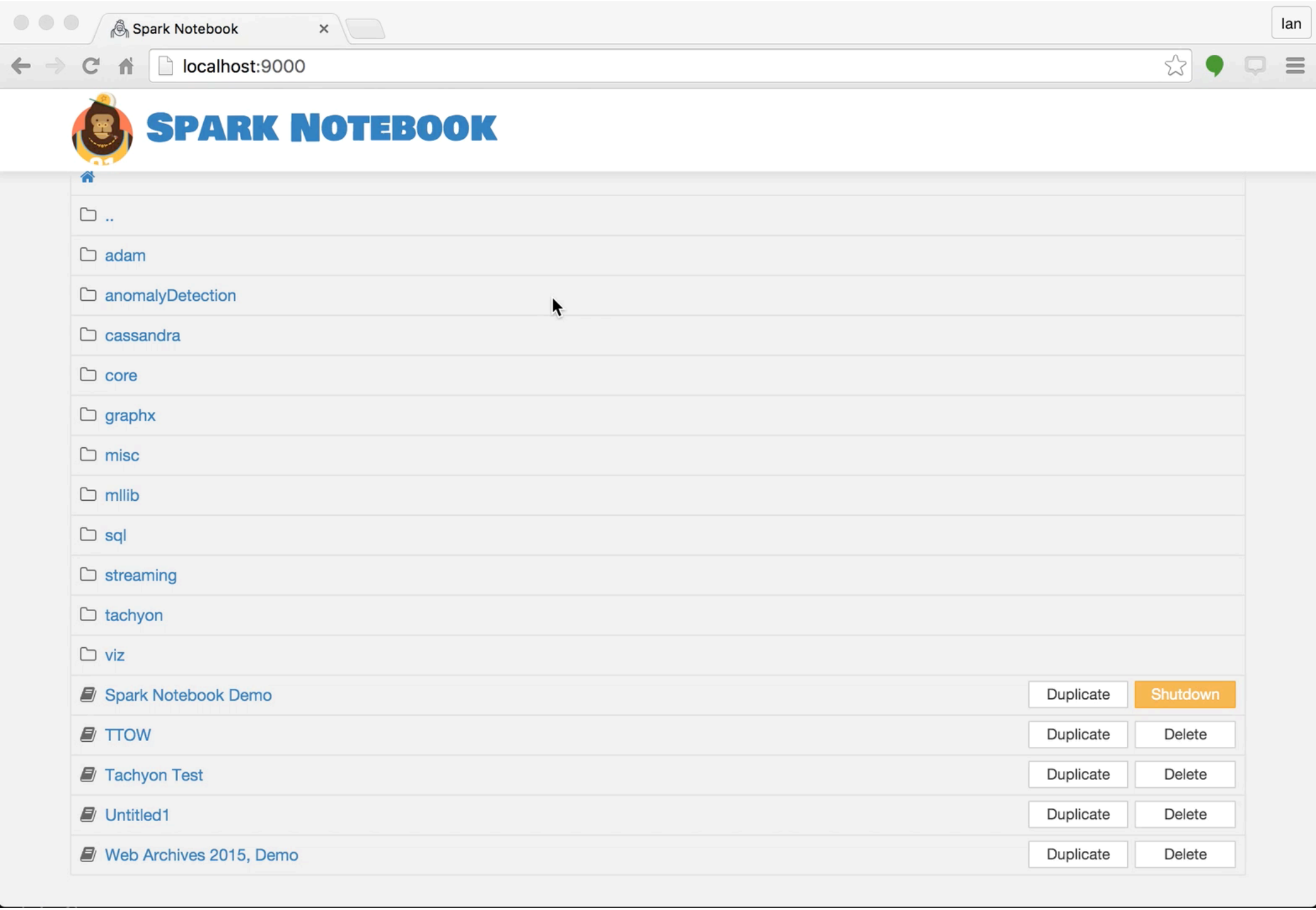


Or D3.js link networks in browser

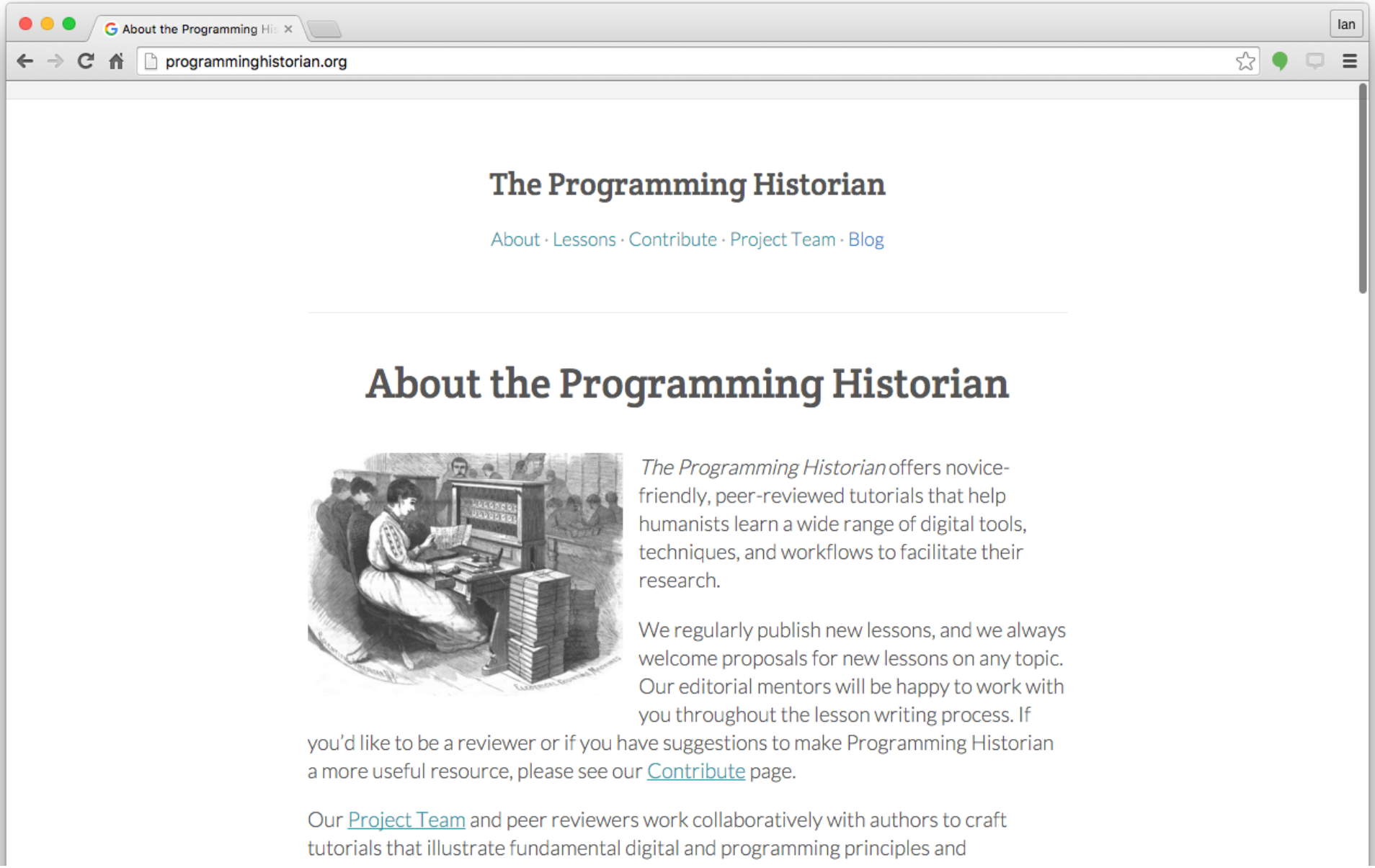


All walkthroughs at:
docs.warcbase.org

Bringing it all together in a notebook environment


Spark Notebook
Ian
localhost:9000
SPARK NOTEBOOK
..
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
Web Archives 2015, Demo
Duplicate
Delete

Where to learn?

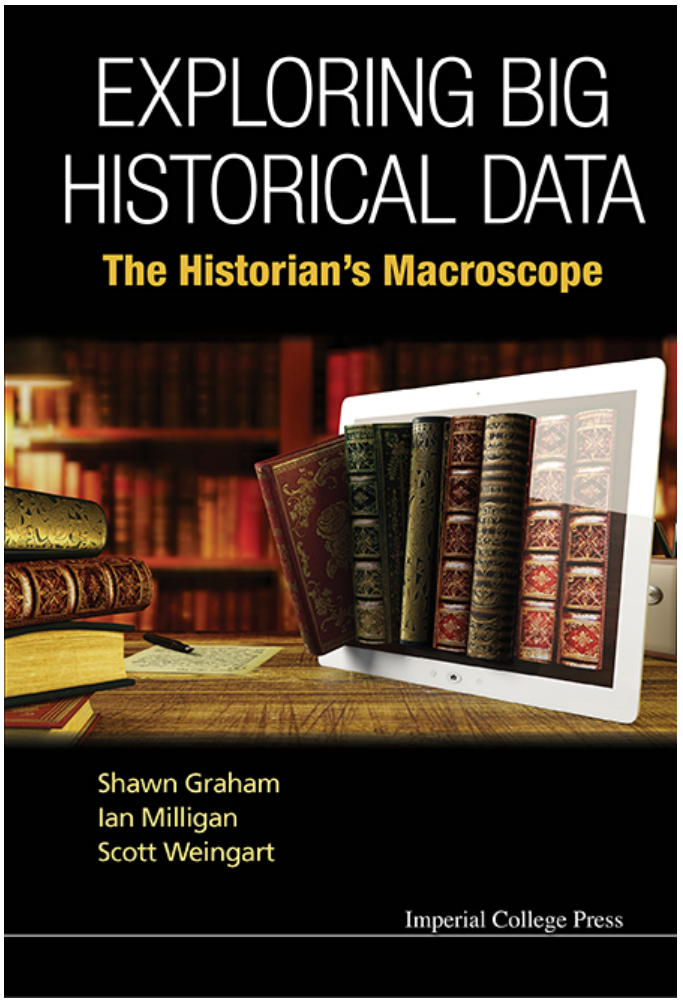


Programming Historian

- 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



Events

- **Software Carpentry** – in-person events, looking into building connections with *Programming Historian*
- **Interdisciplinary hackathons** - *Archives Unleashed* (Toronto, March 2016; Washington, June 2016 - TBA)
- **Conferences** - Like this one, or others

... 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.

Thanks very much!

Questions?

Ian Milligan
Assistant Professor
@ianmilligan1

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History