Big Data and History:
Seeing the Past through a Macroscope

Danish Association for Contemporary History Annual Meeting, Copenhagen Denmark

---

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 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?**

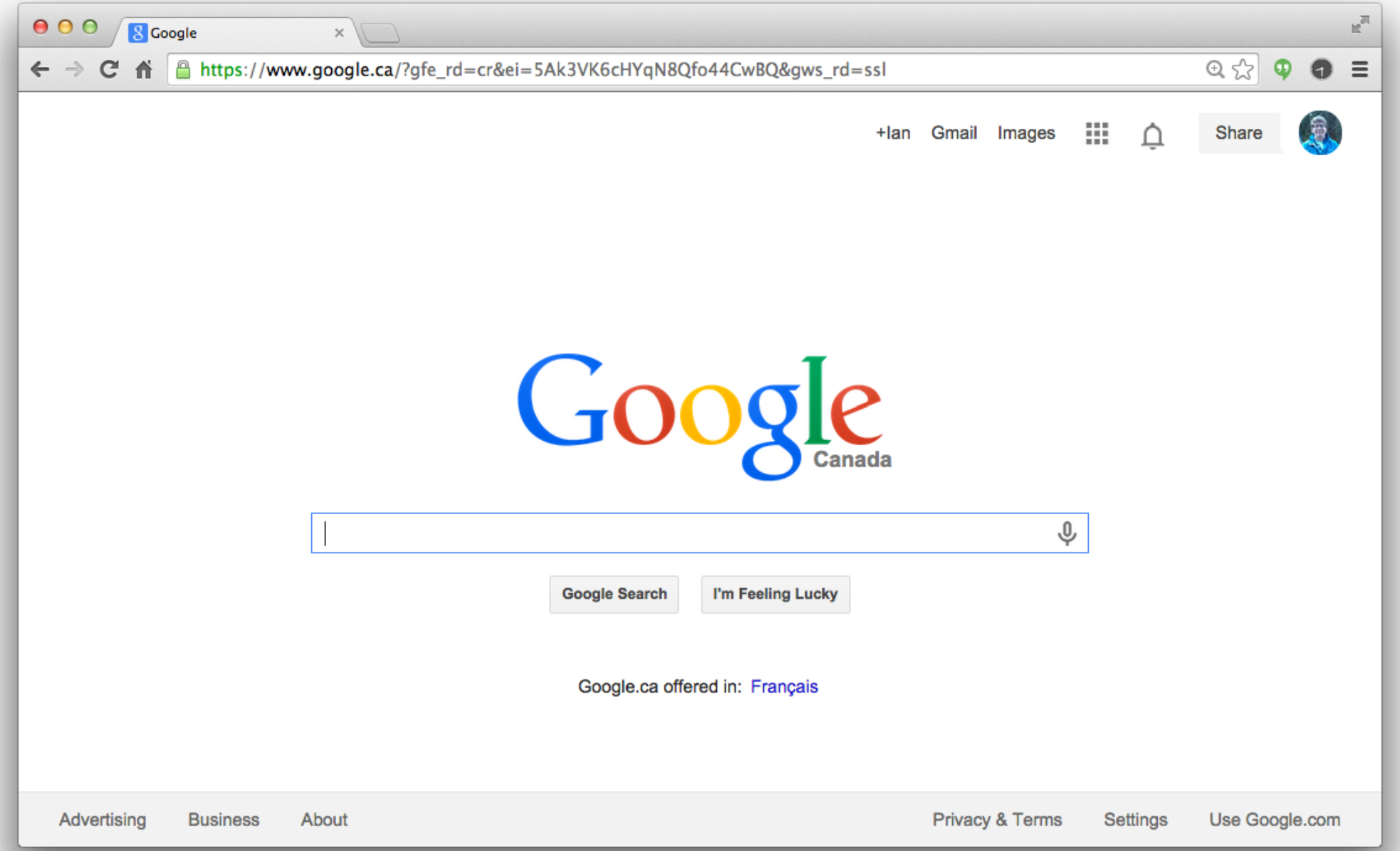
A Prologue: Big Data is Everywhere

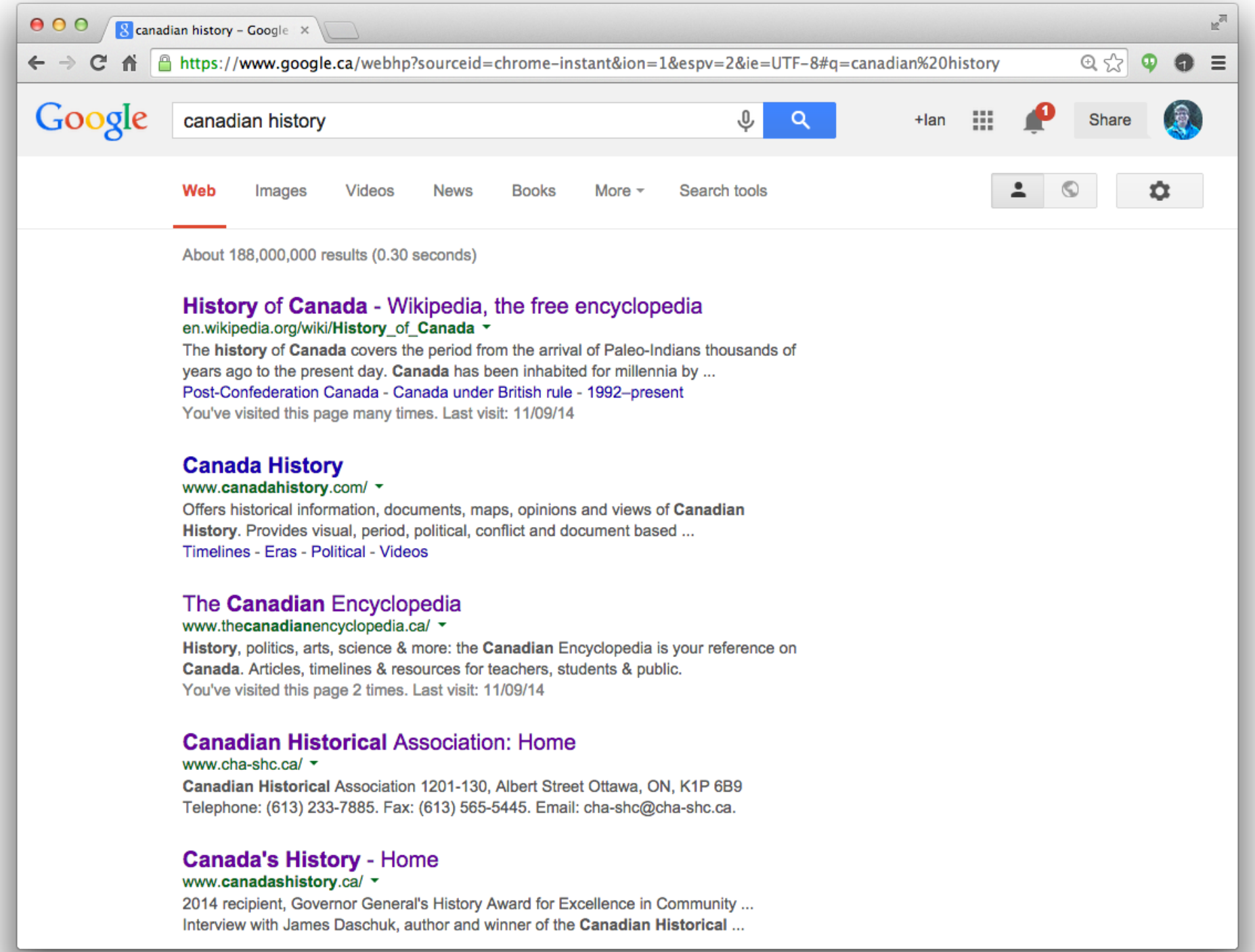
What do we mean by Big Data?

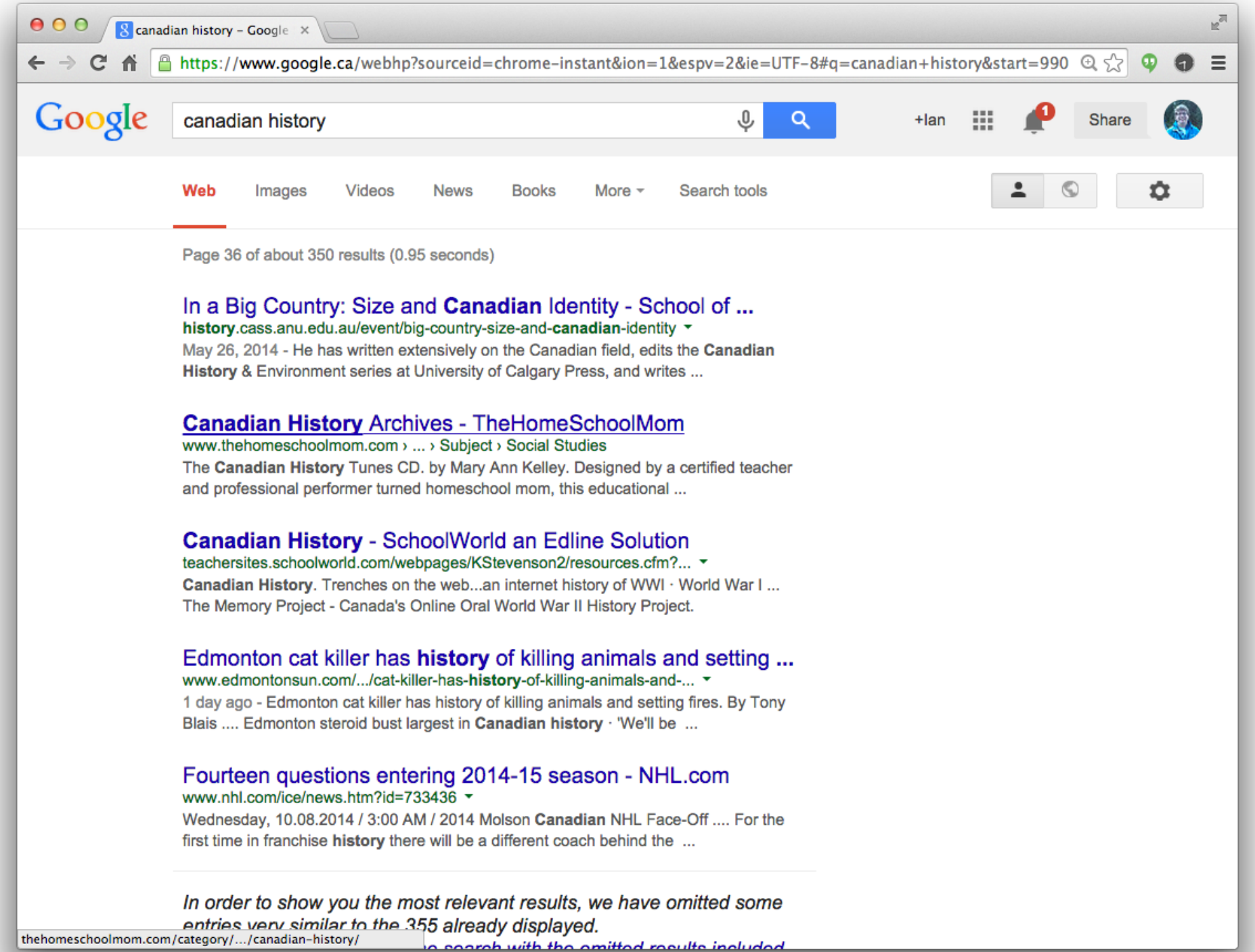
- Computational definition: the 5 Vs (Volume, Velocity, Variety, Veracity, and Value)

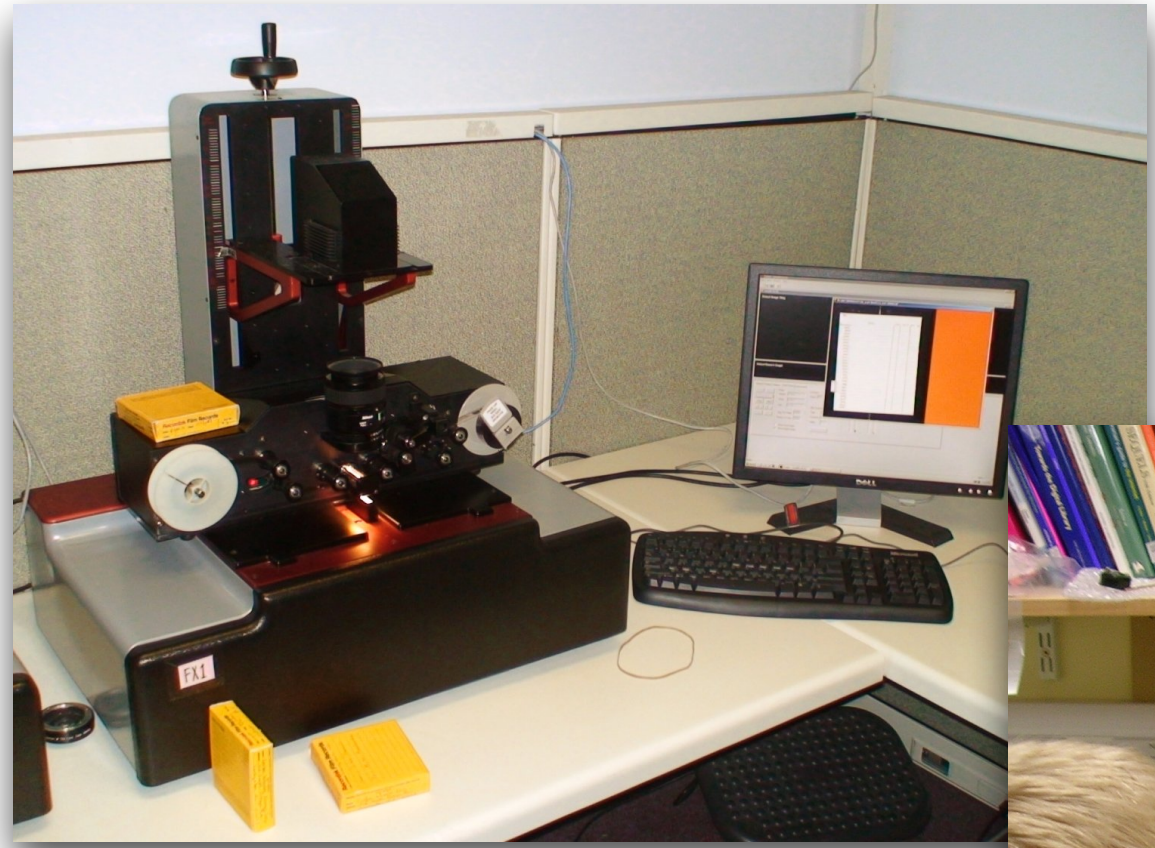
- "For us, as humanists, big is in the eye of the beholder. **If it's more data that you could conceivably read yourself in a reasonable amount of time, or that requires computational intervention to make new sense of it, it's big enough!**" (Shawn Graham, Ian Milligan, Scott Weingart, *Exploring Big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

Why is it everywhere?


Google
https://www.google.ca/?gfe_rd=cr&ei=5Ak3VK6cHYqN8Qfo44CwBQ&gws_rd=ssl
+Ian
Gmail
Images
Share
Google
Canada
Google Search
I'm Feeling Lucky
Google.ca offered in: Français
Advertising
Business
About
Privacy & Terms
Settings
Use Google.com


canadian history - Google
https://www.google.ca/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=canadian%20history
Google
canadian history
+Ian
Share
Web
Images
Videos
News
Books
More
Search tools
About 188,000,000 results (0.30 seconds)
History of Canada - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/History_of_Canada
The history of Canada covers the period from the arrival of Paleo-Indians thousands of years ago to the present day. Canada has been inhabited for millennia by ...
Post-Confederation Canada - Canada under British rule - 1992–present
You've visited this page many times. Last visit: 11/09/14
Canada History
www.canadahistory.com/
Offers historical information, documents, maps, opinions and views of Canadian History. Provides visual, period, political, conflict and document based ...
Timelines - Eras - Political - Videos
The Canadian Encyclopedia
www.thecanadianencyclopedia.ca/
History, politics, arts, science & more: the Canadian Encyclopedia is your reference on Canada. Articles, timelines & resources for teachers, students & public.
You've visited this page 2 times. Last visit: 11/09/14
Canadian Historical Association: Home
www.cha-shc.ca/
Canadian Historical Association 1201-130, Albert Street Ottawa, ON, K1P 6B9
Telephone: (613) 233-7885. Fax: (613) 565-5445. Email: cha-shc@cha-shc.ca.
Canada's History - Home
www.canadashistory.ca/
2014 recipient, Governor General's History Award for Excellence in Community ...
Interview with James Daschuk, author and winner of the Canadian Historical ...


canadian history - Google
https://www.google.ca/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=canadian+history&start=990
Google
canadian history
+Ian
Share
Web
Images
Videos
News
Books
More
Search tools
Page 36 of about 350 results (0.95 seconds)
In a Big Country: Size and Canadian Identity - School of ...
history.cass.anu.edu.au/event/big-country-size-and-canadian-identity
May 26, 2014 - He has written extensively on the Canadian field, edits the Canadian History & Environment series at University of Calgary Press, and writes ...
Canadian History Archives - TheHomeSchoolMom
www.thehomeschoolmom.com › ... › Subject › Social Studies
The Canadian History Tunes CD. by Mary Ann Kelley. Designed by a certified teacher and professional performer turned homeschool mom, this educational ...
Canadian History - SchoolWorld an Edline Solution
teachersites.schoolworld.com/webpages/KStevenson2/resources.cfm?...
Canadian History. Trenches on the web...an internet history of WWI · World War I ... The Memory Project - Canada's Online Oral World War II History Project.
Edmonton cat killer has history of killing animals and setting ...
www.edmontonsun.com/.../cat-killer-has-history-of-killing-animals-and-...
1 day ago - Edmonton cat killer has history of killing animals and setting fires. By Tony Blais .... Edmonton steroid bust largest in Canadian history · 'We'll be ...
Fourteen questions entering 2014-15 season - NHL.com
www.nhl.com/ice/news.htm?id=733436
Wednesday, 10.08.2014 / 3:00 AM / 2014 Molson Canadian NHL Face-Off .... For the first time in franchise history there will be a different coach behind the ...
In order to show you the most relevant results, we have omitted some entries very similar to the 355 already displayed.
thehomeschoolmom.com/category/.../canadian-history/


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

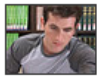
Search Clear form

Search subject areas

Use search forms customized for each subject.


The Arts


Business

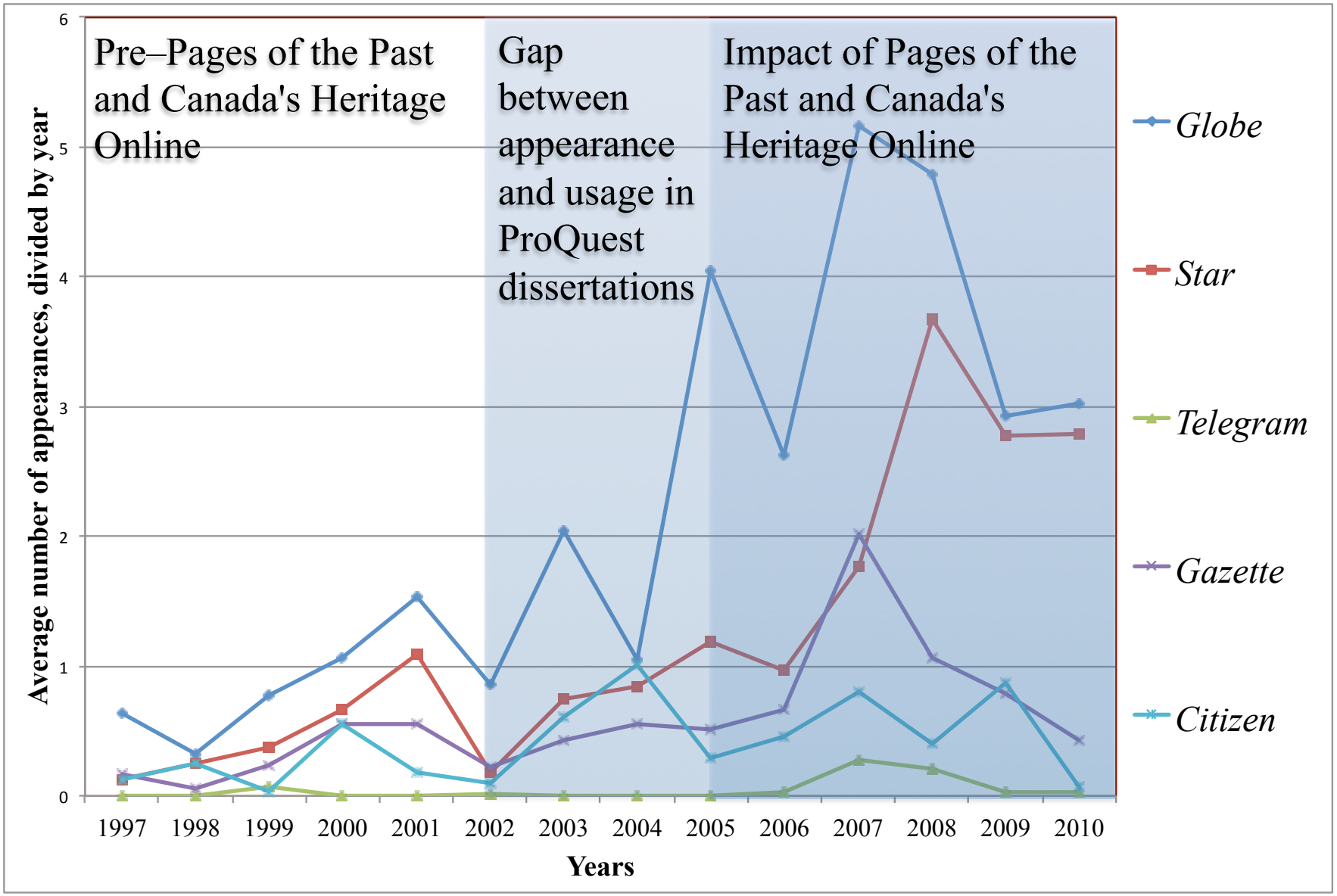

Dissertations & Theses


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 ...

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

A process that is only now beginning to accelerate.

Historians need to be ready to engage with web archives. They have the potential to fundamentally transform our trade.

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

[500 Historians Call for End to Impeachment Inquiry]

The Historians Committee for Open Debate

[About the HCFOD] [What's Hot!]

[How to Join] [Subscribe]

[Directory] [On-Line Forum]

[Press Releases] [Document Archive]

[New Books] [Links]

www.historians.org SEARCH :

(by keyword, name, or subject)

Search Now

AND Insensitive

For more information, contact info@historians.org

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

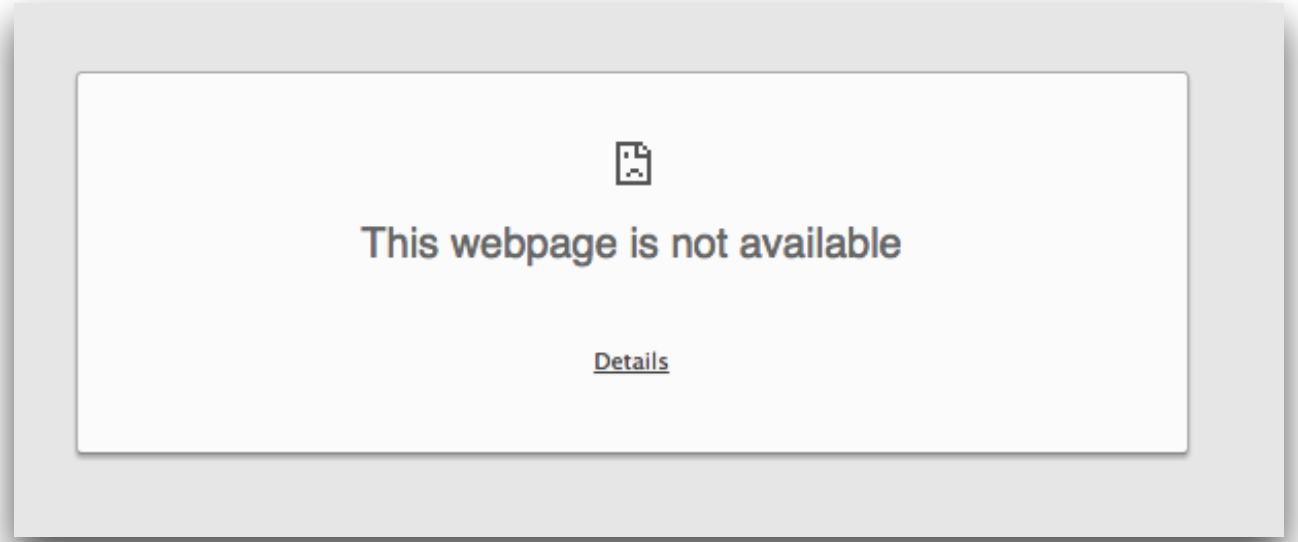
General Inquiries:
general@occupywallst.org
Press Inquiries:
press@occupywallst.org
Press Phone: +1 (347) 292-1444
Help & Directions: +1 (516) 708-4777
Watch: The world we're building
Read: This call to action
Liberty Square Eviction Defense: Text "@occupyalert" to 23559 to receive alerts in the event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

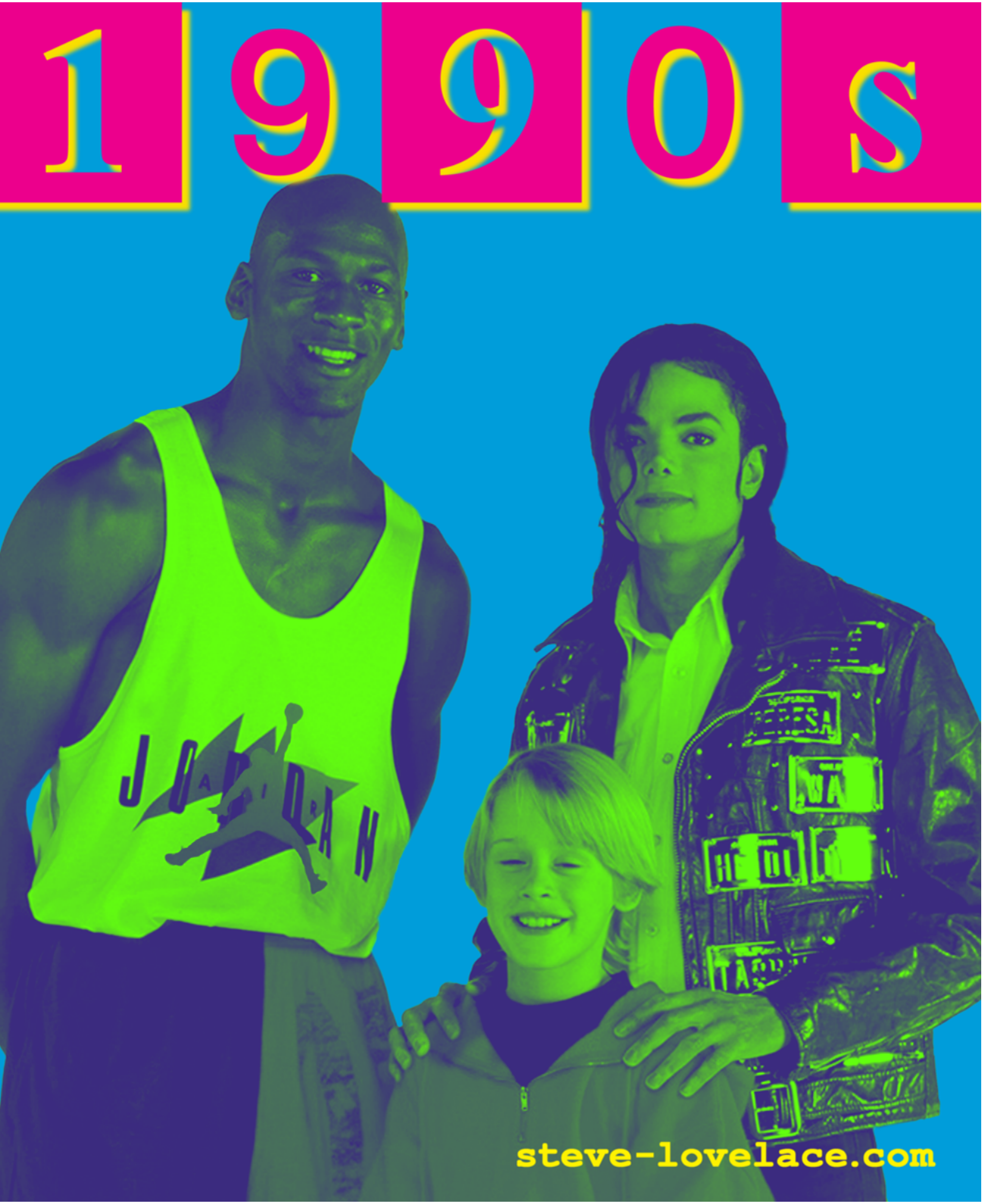
This #ows movement empowers real people to create real change from the bottom up. We want to see a general assembly in every backyard, on every street corner because we don't need Wall Street and we don't need politicians to build a better society.

the only solution is WorldRevolution

Click here for NYCGA committee meeting times.

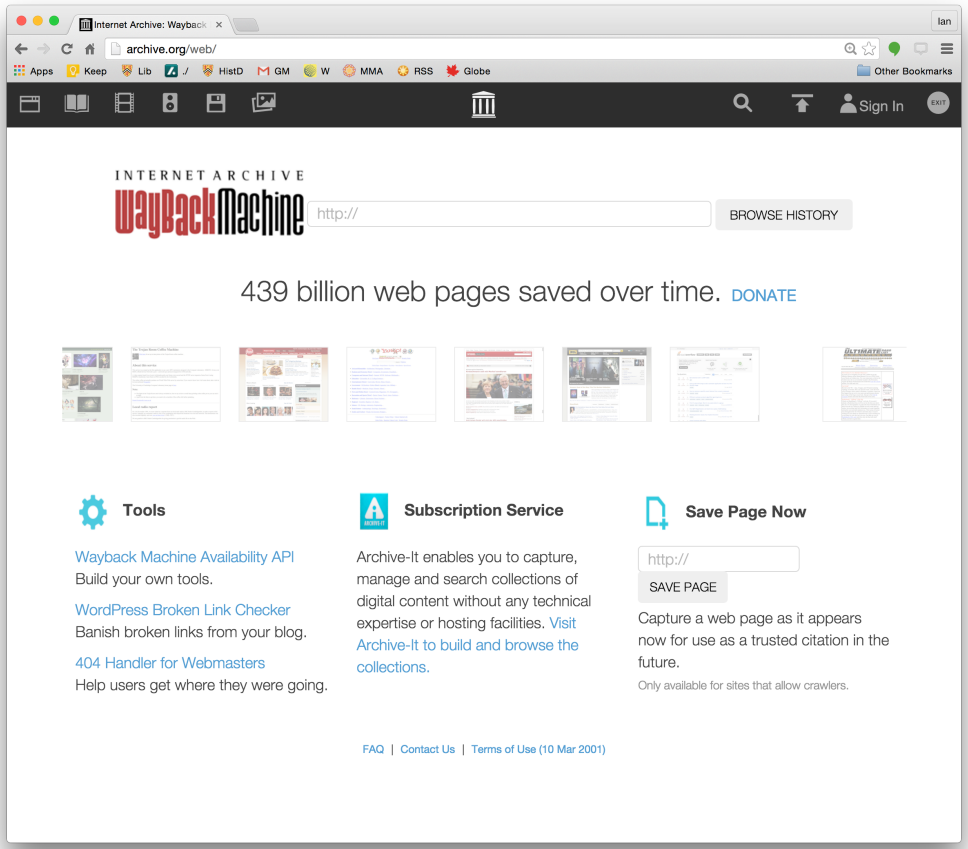
“…. [n]ow expectations have inverted. Everything may be recorded and preserved, at least potentially.”

- James Gleick



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

Nightmare Scenario



This won't be enough!

… but what will our search engines look like?

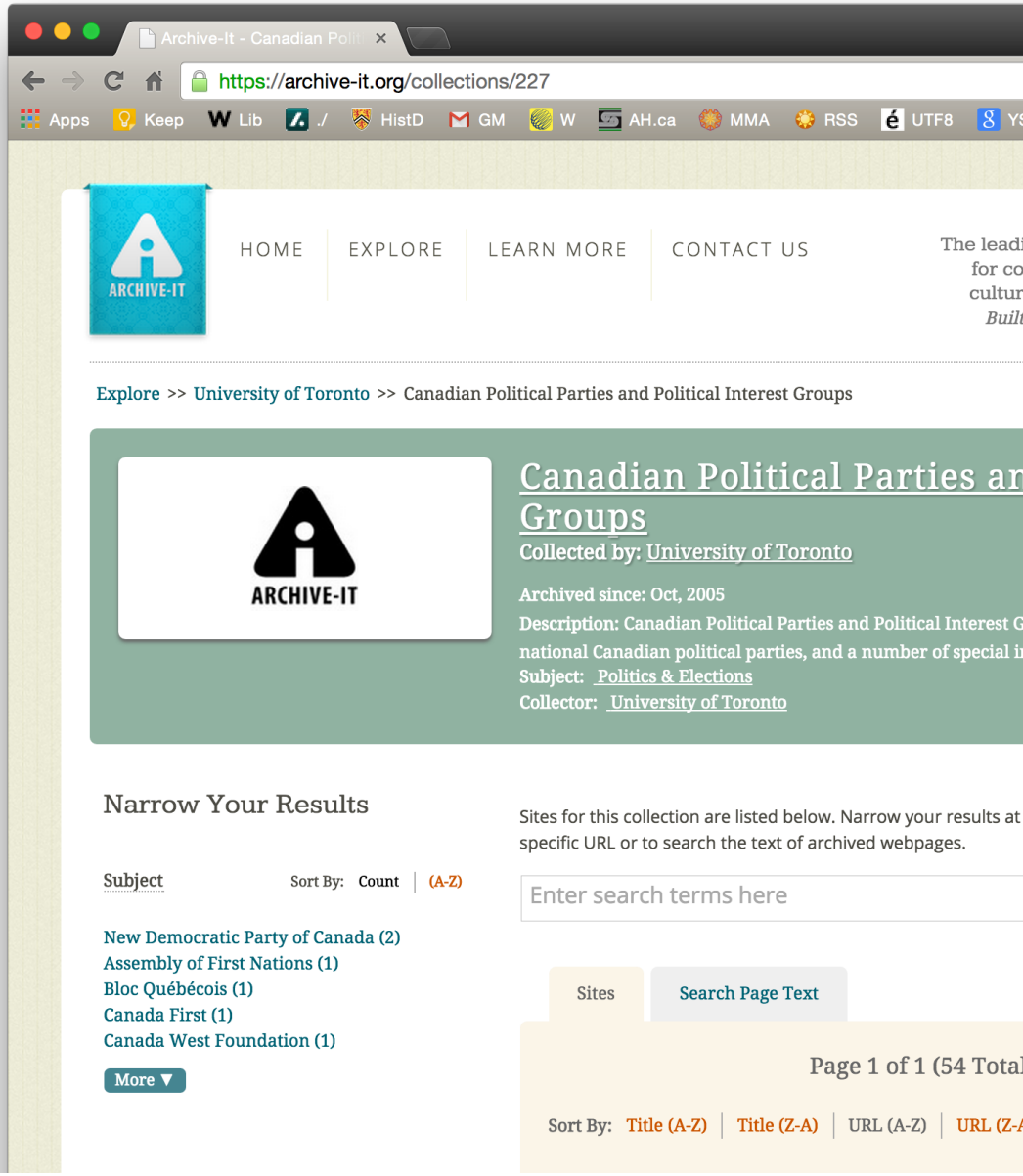
Nightmare Scenario

- 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…

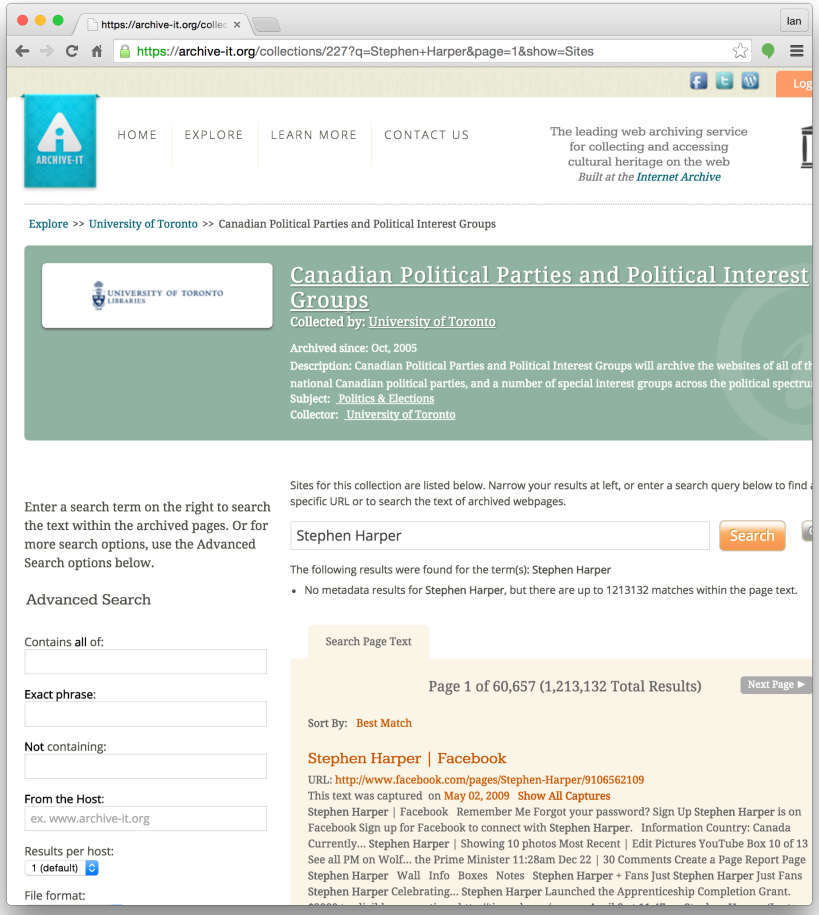


Pivotal Changes in Canadian Politics, 2005-2015

- 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

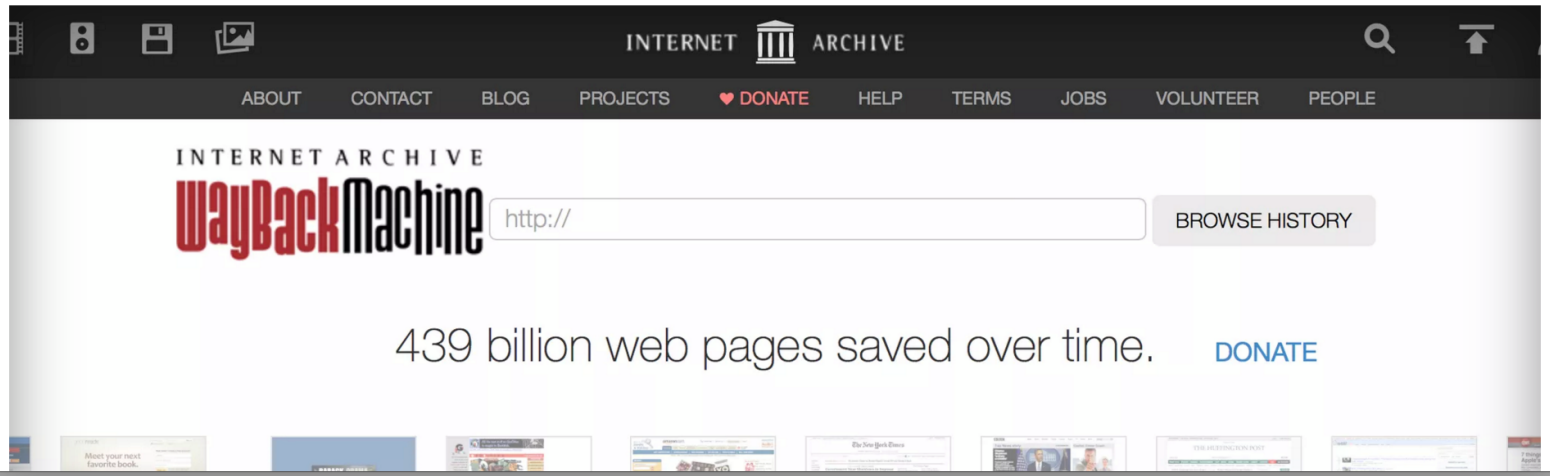
- Very limited - simple search engine, some advanced options; no facets
- Great collections.. but nobody uses them!

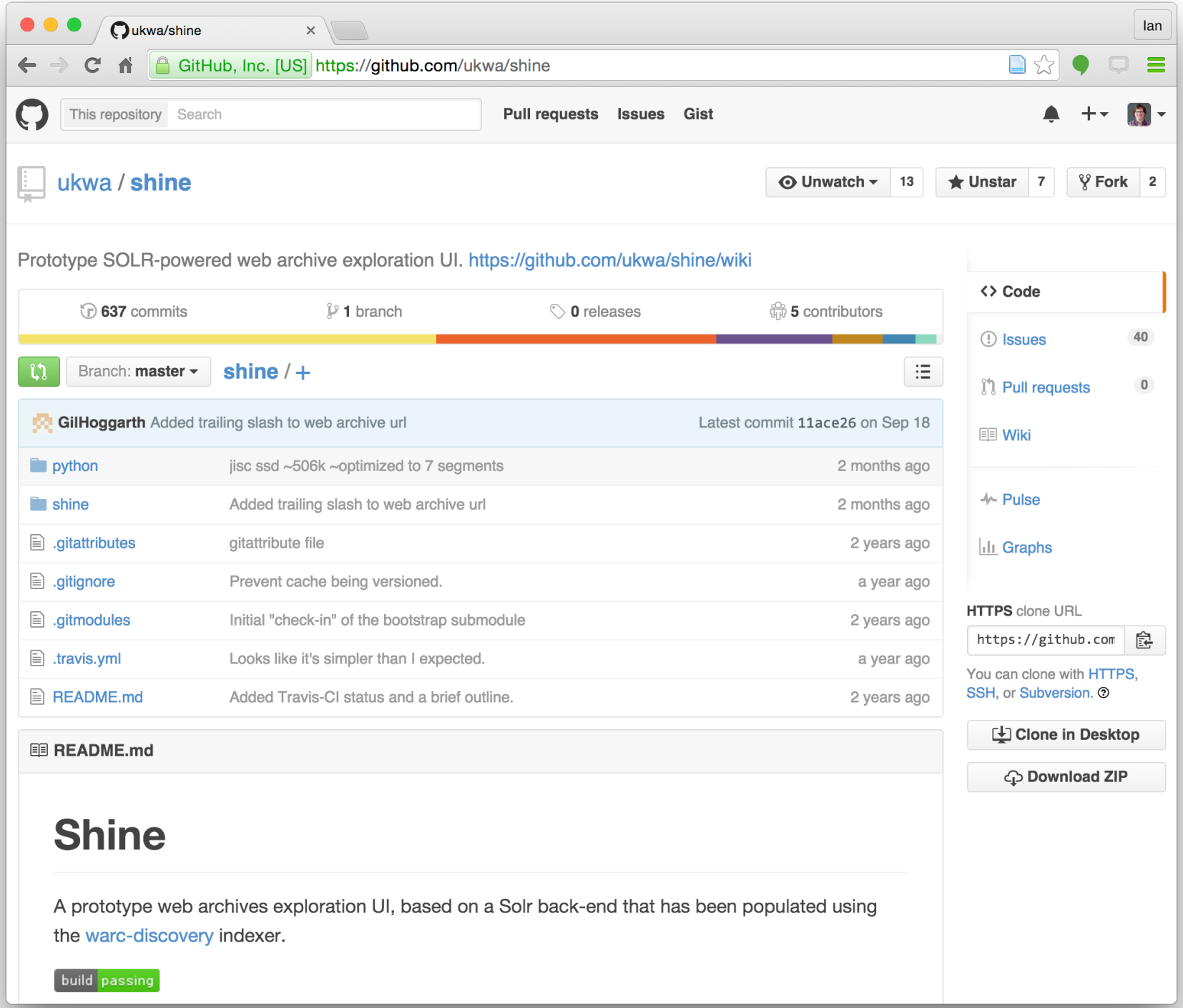


TECH

The Internet Archive will launch a modernized Wayback Machine in 2017

By Chris Welch on October 22, 2015 12:09 pm Email @chriswelch




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

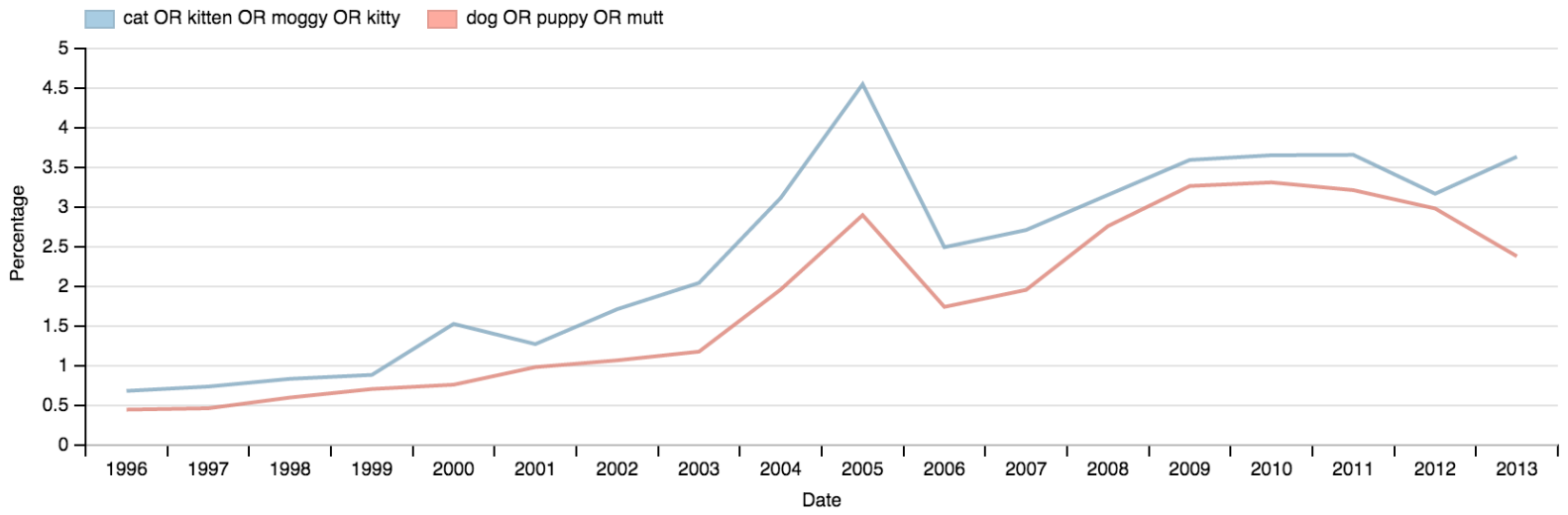
Trend results 1996-2013 fo

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

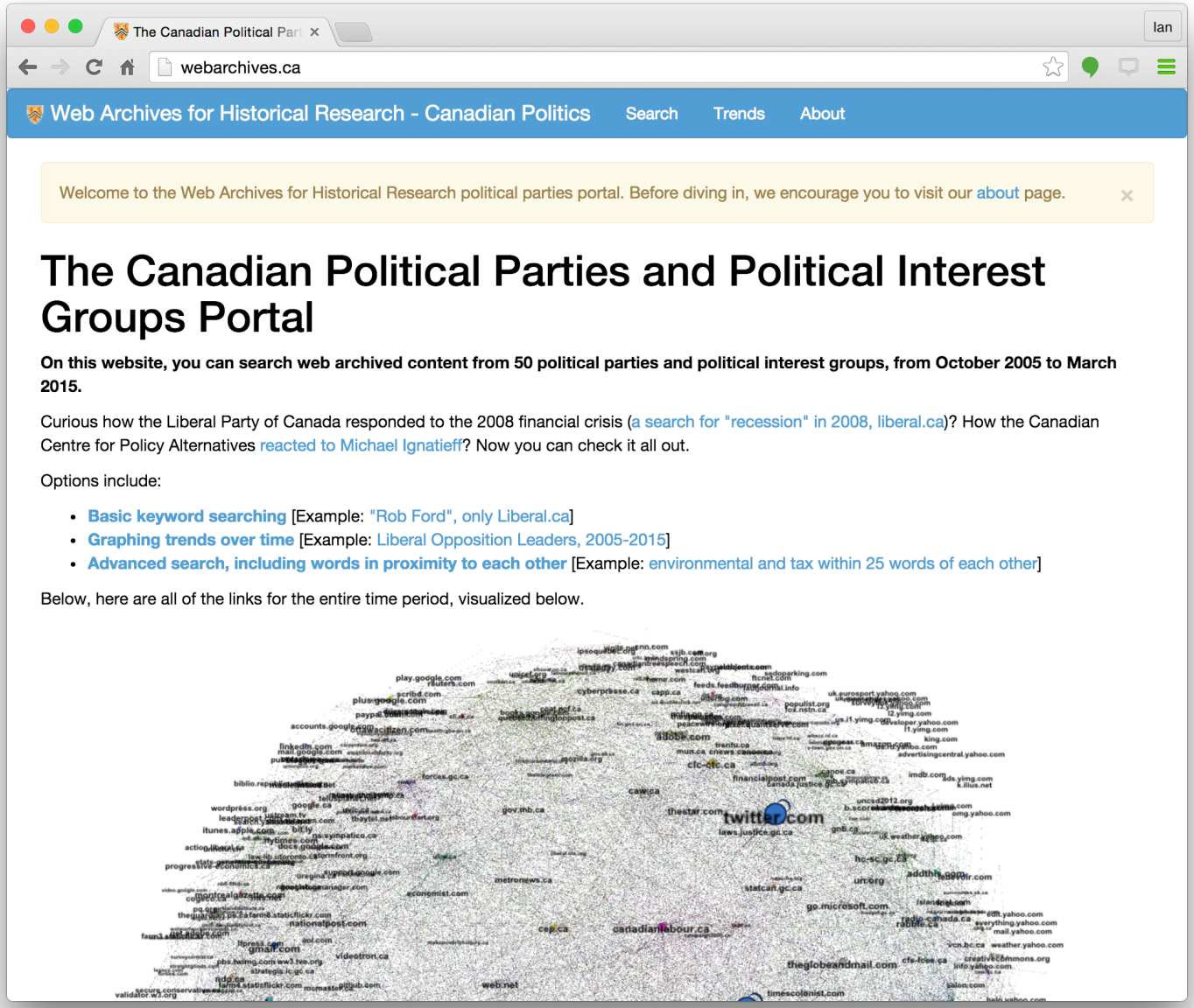
UK Web Archive | Search | Trends | Login

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)

Trend results 2005-2015 fo ×

webarchives.ca/graph

Web Archives for Historical Research - Canadian Politics

Search Trends About

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

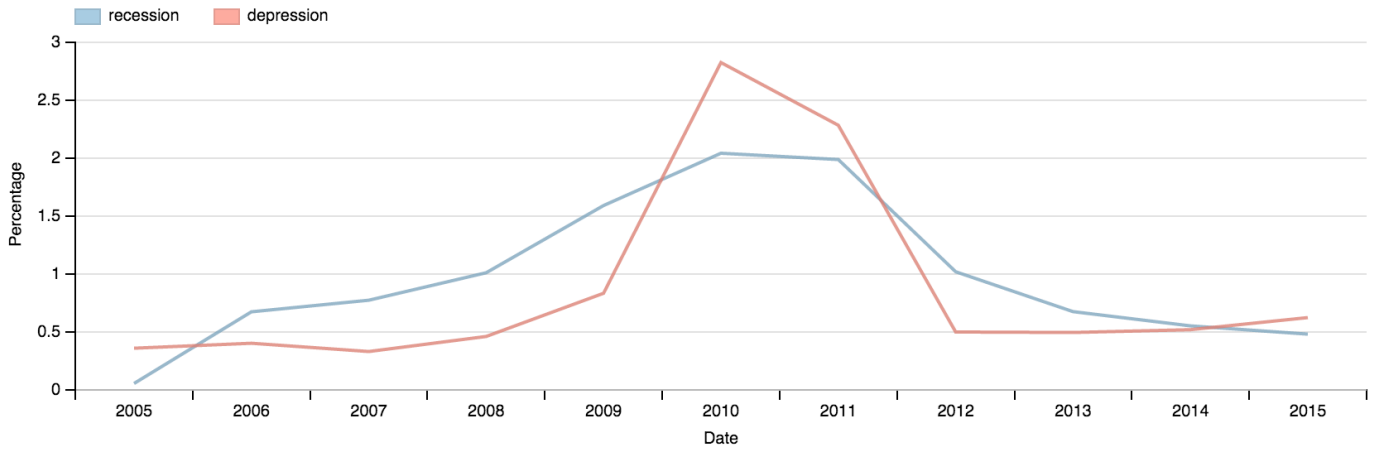
The graph shows the relative frequency of a term in each year - i.e. how often did "recession" appear in 2008 as a percentage of all keywords? Enter your search terms and click "Update Graph." To see actual results, click on the graph's line where it intersects a year.

You can also use this graph to run keyword searches on particular top-level domains. To see how often the word "global warming" appeared in the Conservative Party of Canada's domain versus the New Democratic Party of Canada, search with the query: "global warming" domain:conservative.ca, "global warming" domain:ndp.ca [view this example].

For a full list of domains, consult the about page.

Colours are assigned to the graph below depending on the order they are inputted (i.e. the Conservative Party of Canada will not always be blue).

recession,depression | 2005 | 2015 | Update Graph | Reset All



Some Sample Queries to Get Started With:

- "Public Transit": Liberals/NDP/Conservatives. We use quotation marks to search on the exact phrase.
- "Tar Sands": Amongst the three major parties, who uses and subsequently stops using this phrase for Canada's Oil Sands?
- "Climate Change": Similar to the above two queries, we can see this issue rise and possibly wane in public interest.

Feel free to take these examples and adapt them to your own purposes. If you find something interesting, please let us know!

Trend results 2005-2015 f

webarchives.ca/graph?query="public+transit"+domain%3Aconservative.ca%2C+"public+transit"+domain%3Andp.ca...

Web Archives for Historical Research - Canadian Politics

Search Trends About

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

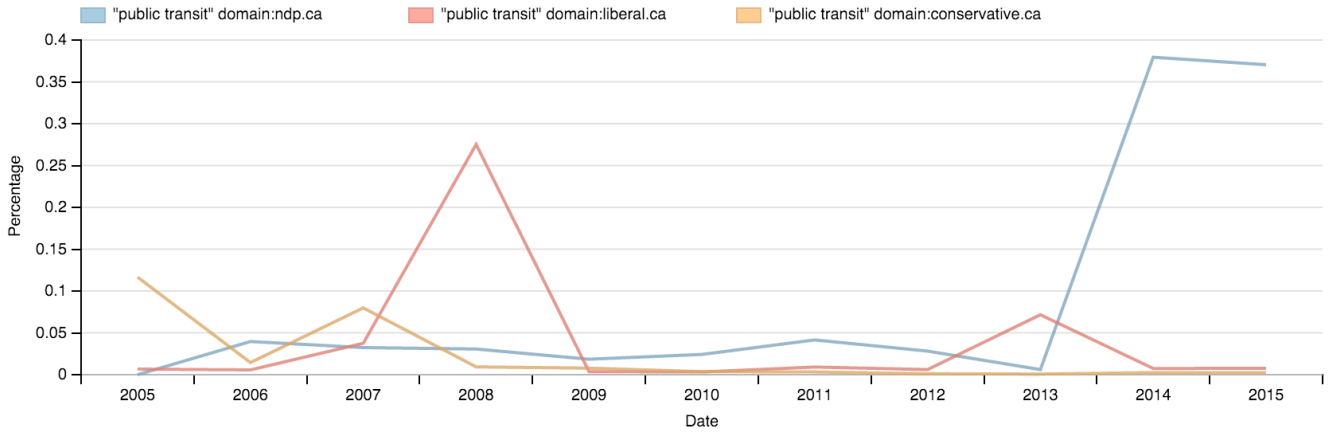
The graph shows the relative frequency of a term in each year - i.e. how often did "recession" appear in 2008 as a percentage of all keywords? Enter your search terms and click "Update Graph." To see actual results, click on the graph's line where it intersects a year.

You can also use this graph to run keyword searches on particular top-level domains. To see how often the word "global warming" appeared in the Conservative Party of Canada's domain versus the New Democratic Party of Canada, search with the query: "global warming" domain:conservative.ca, "global warming" domain:ndp.ca [view this example].

For a full list of domains, consult the about page.

Colours are assigned to the graph below depending on the order they are inputted (i.e. the Conservative Party of Canada will not always be blue).

"public transit" domain:conservative.ca, "p | 2005 | 2015 | Update Graph | Reset All



Some Sample Queries to Get Started With:

- "Public Transit": Liberals/NDP/Conservatives. We use quotation marks to search on the exact phrase.
- "Tar Sands": Amongst the three major parties, who uses and subsequently stops using this phrase for Canada's Oil Sands?
- "Climate Change": Similar to the above two queries, we can see this issue rise and possibly wane in public interest.

Feel free to take these examples and adapt them to your own purposes. If you find something interesting, please let us know!

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

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

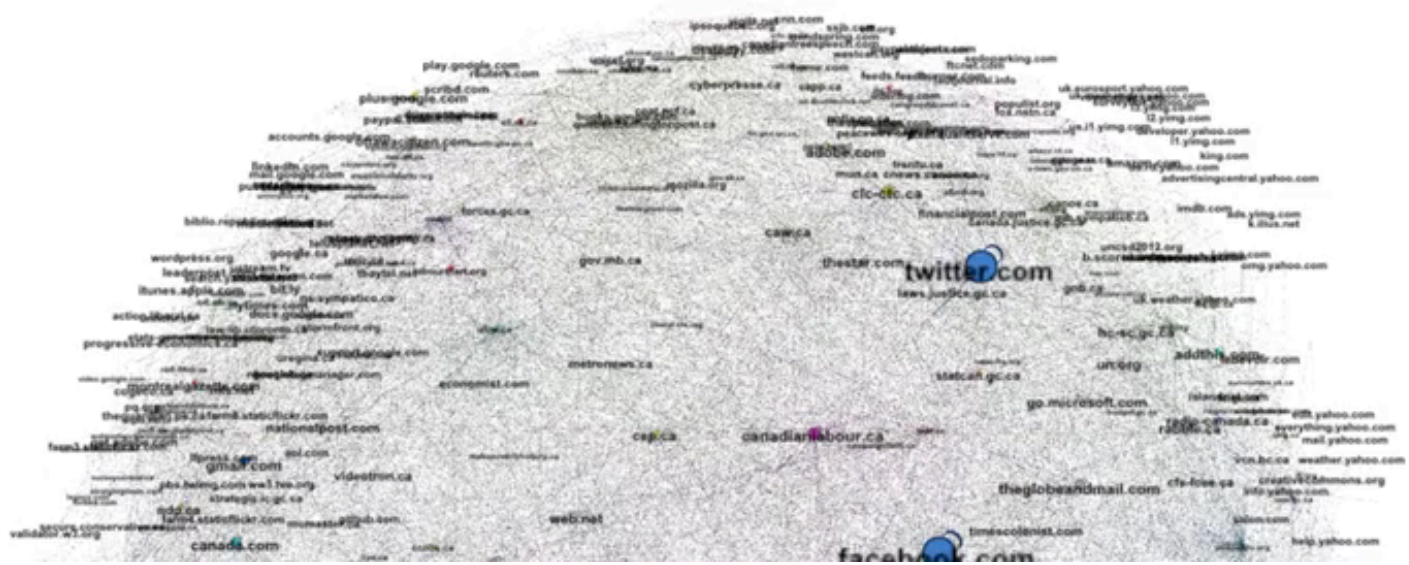
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

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]

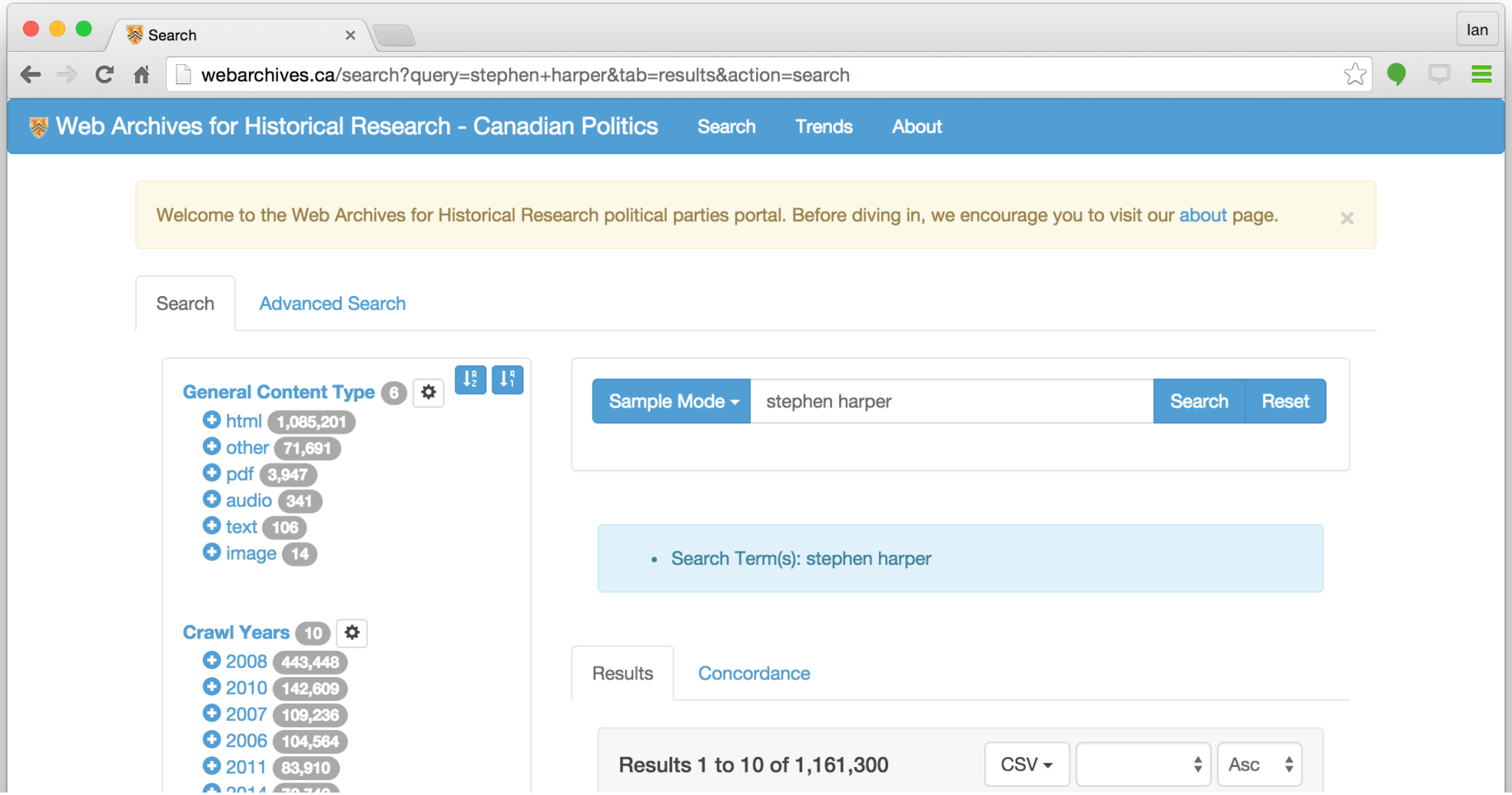
Below, here are all of the links for the entire time period, visualized below.



Five Things We Learned

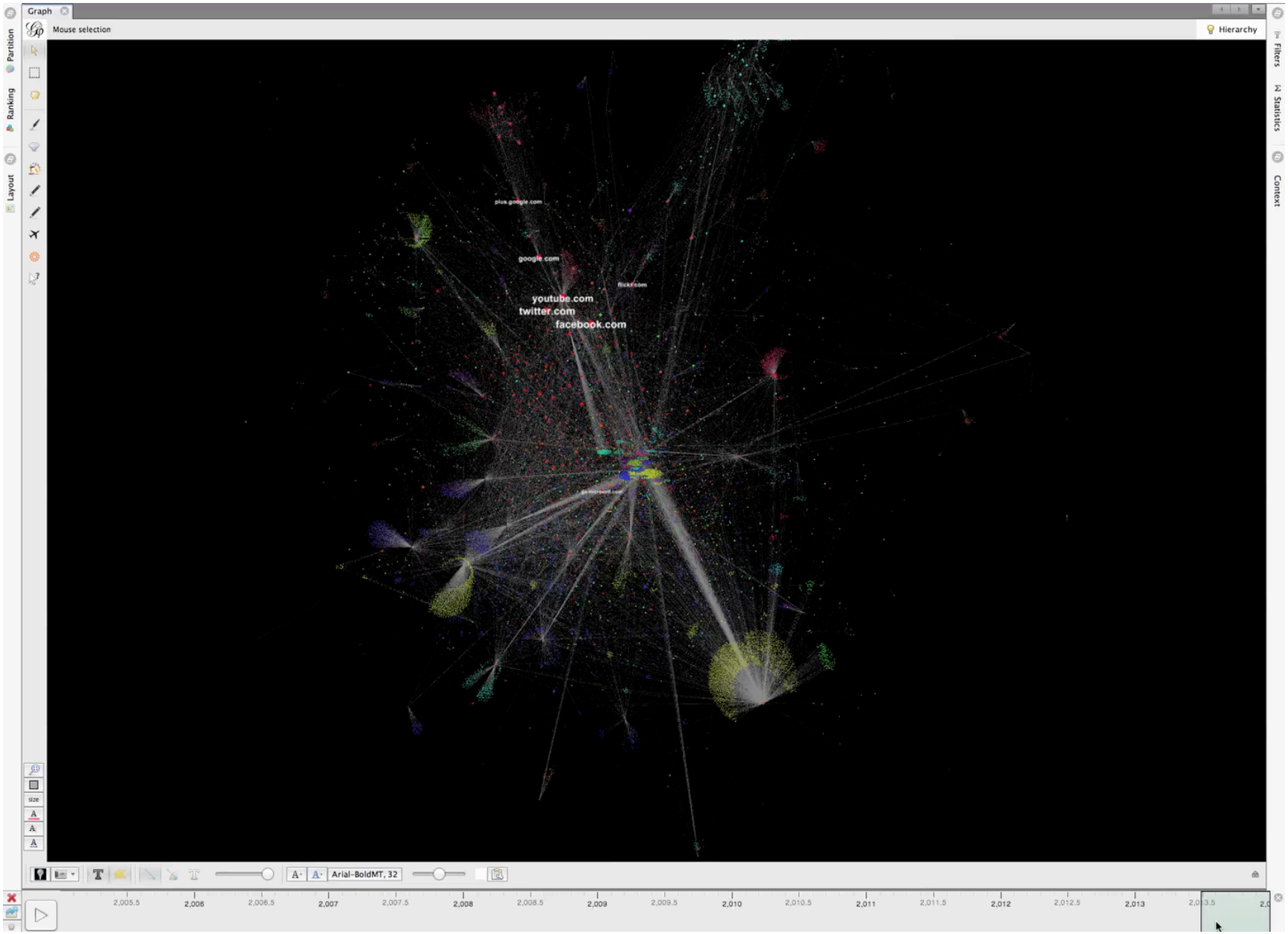
- Political parties delete content
- User-generated comments were more common in political parties
- Absences can be more informative than presences
- We can see the rise/fall of prominent people
- Enabling user access is truly transformative

Good for public engagement - but limited for scholarship….

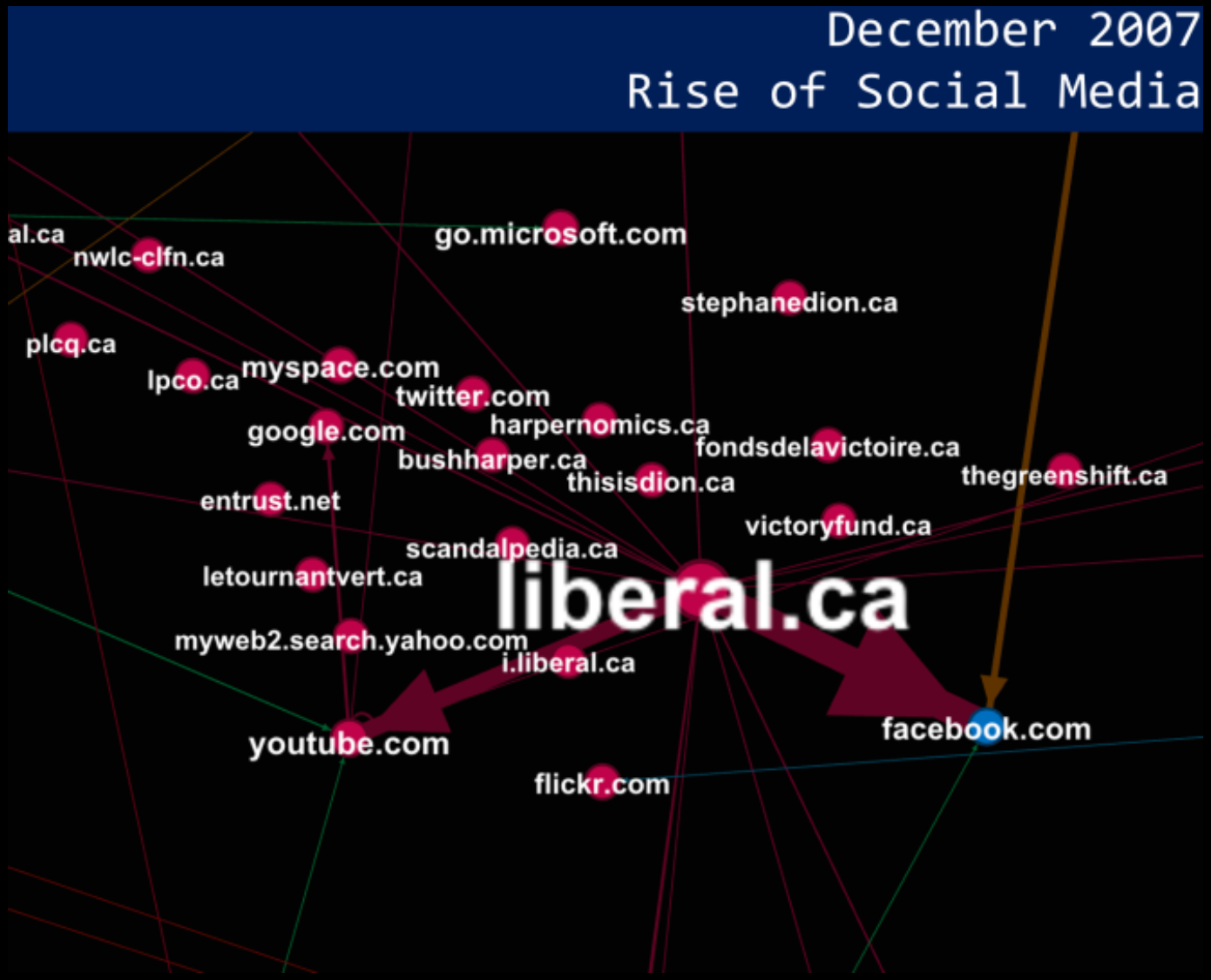


Getting over my bias towards content **and embracing metadata**

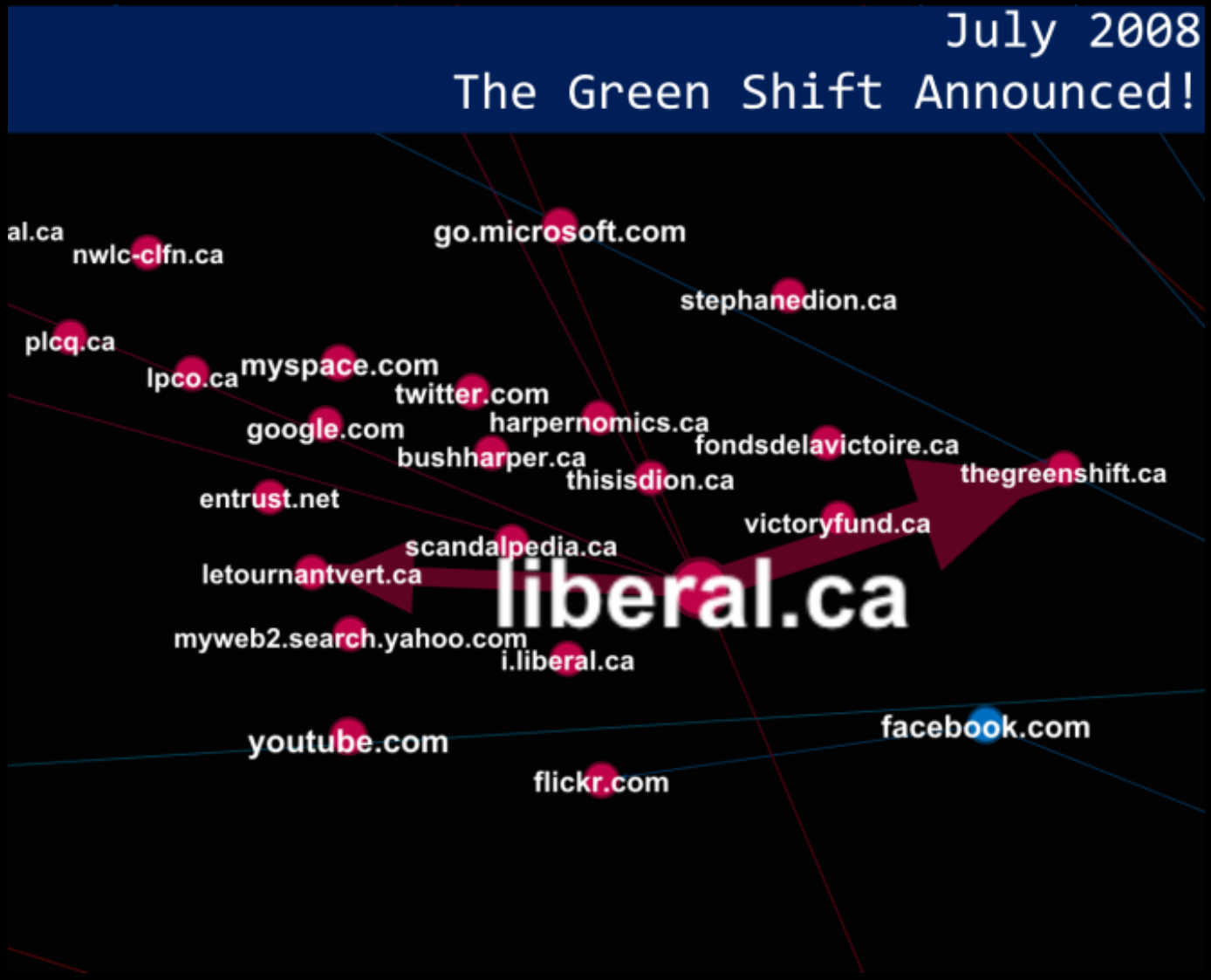
Metadata Extraction

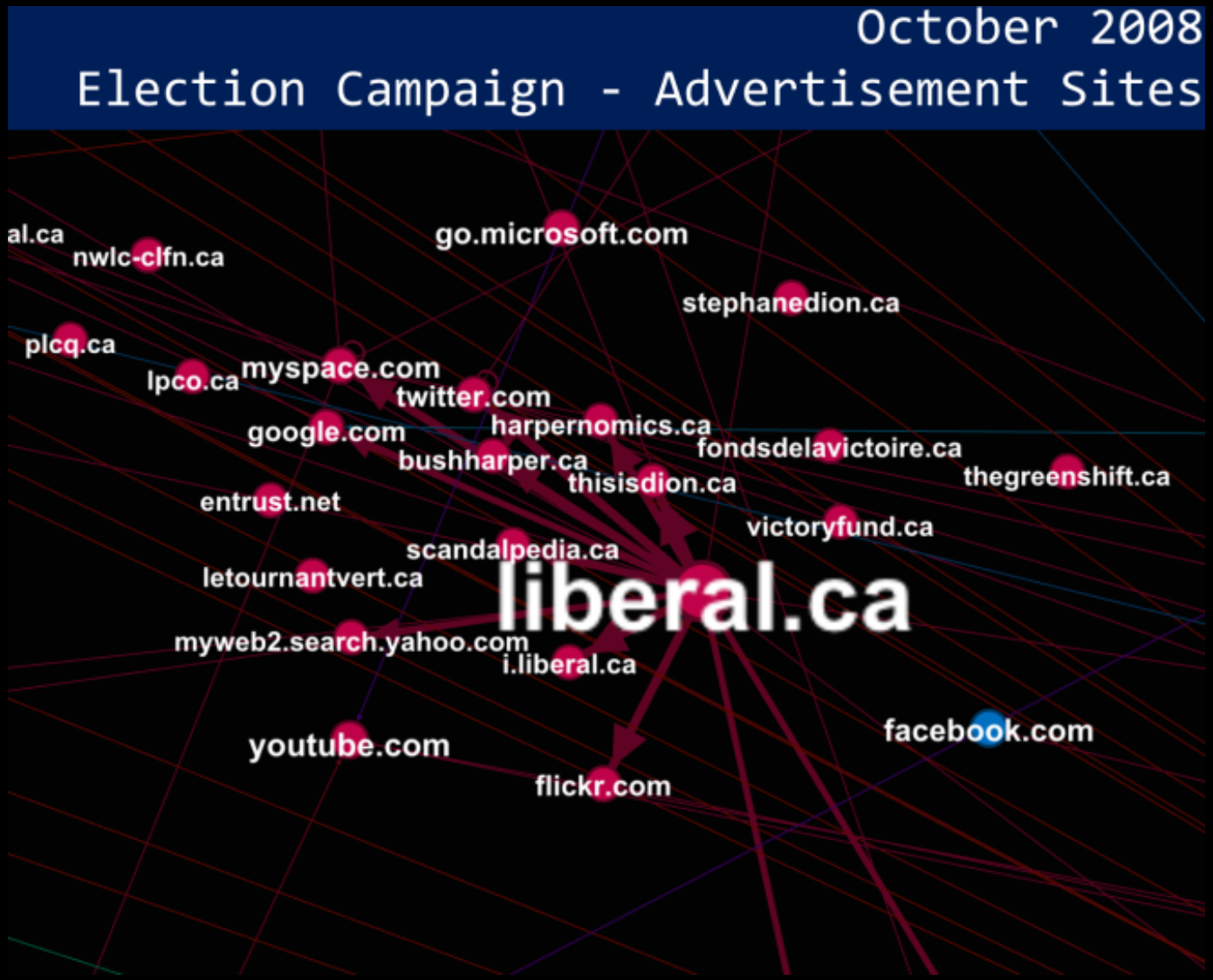



December 2006
Stephane Dion Elected Leader of Party
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca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

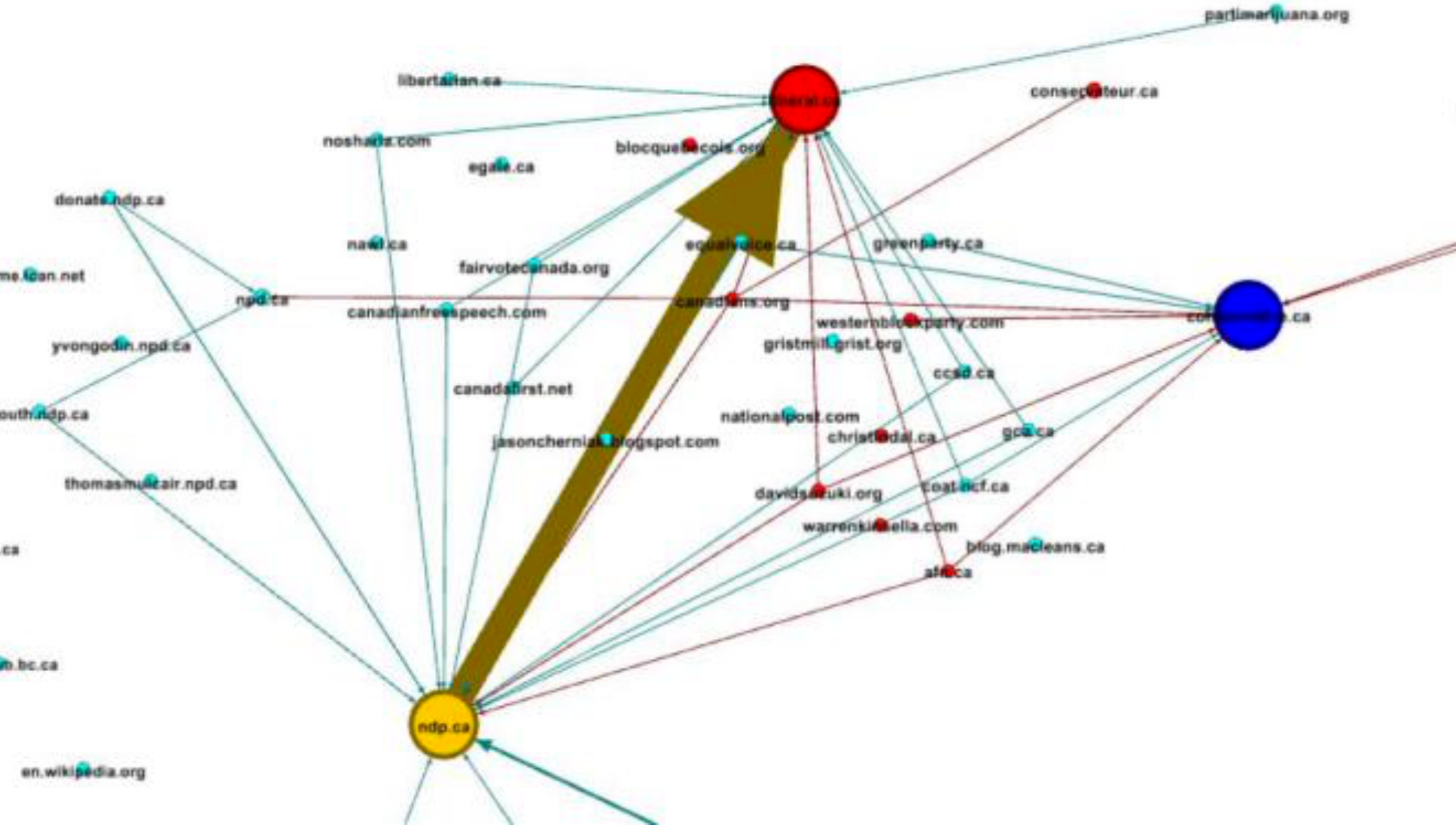

April 2008
Fundraising with the Victory Fund/
Fonds de la Victoire
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca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

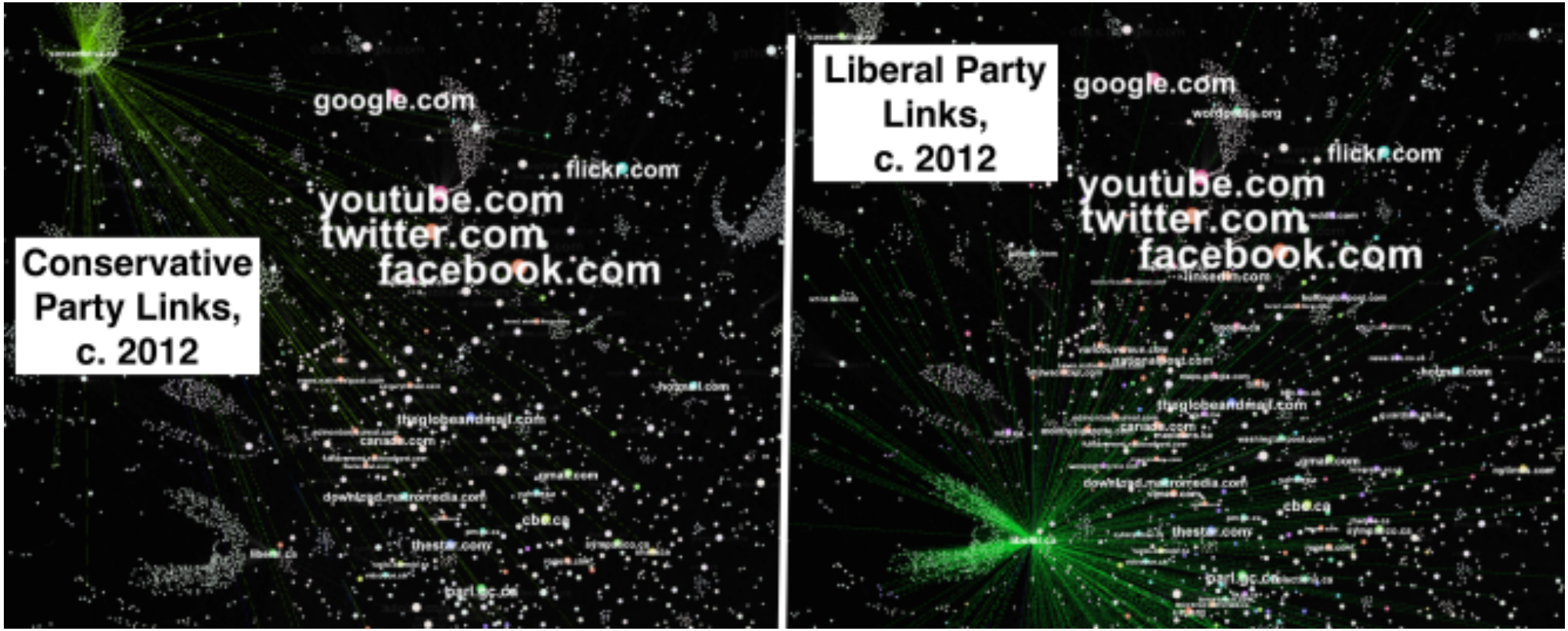

October 2008
Election Campaign - Advertisement Sites
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca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



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

- Results @ http://ianmilligan.ca/2015/02/05/topic-modeling-web-archive-modularity-classes/

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.

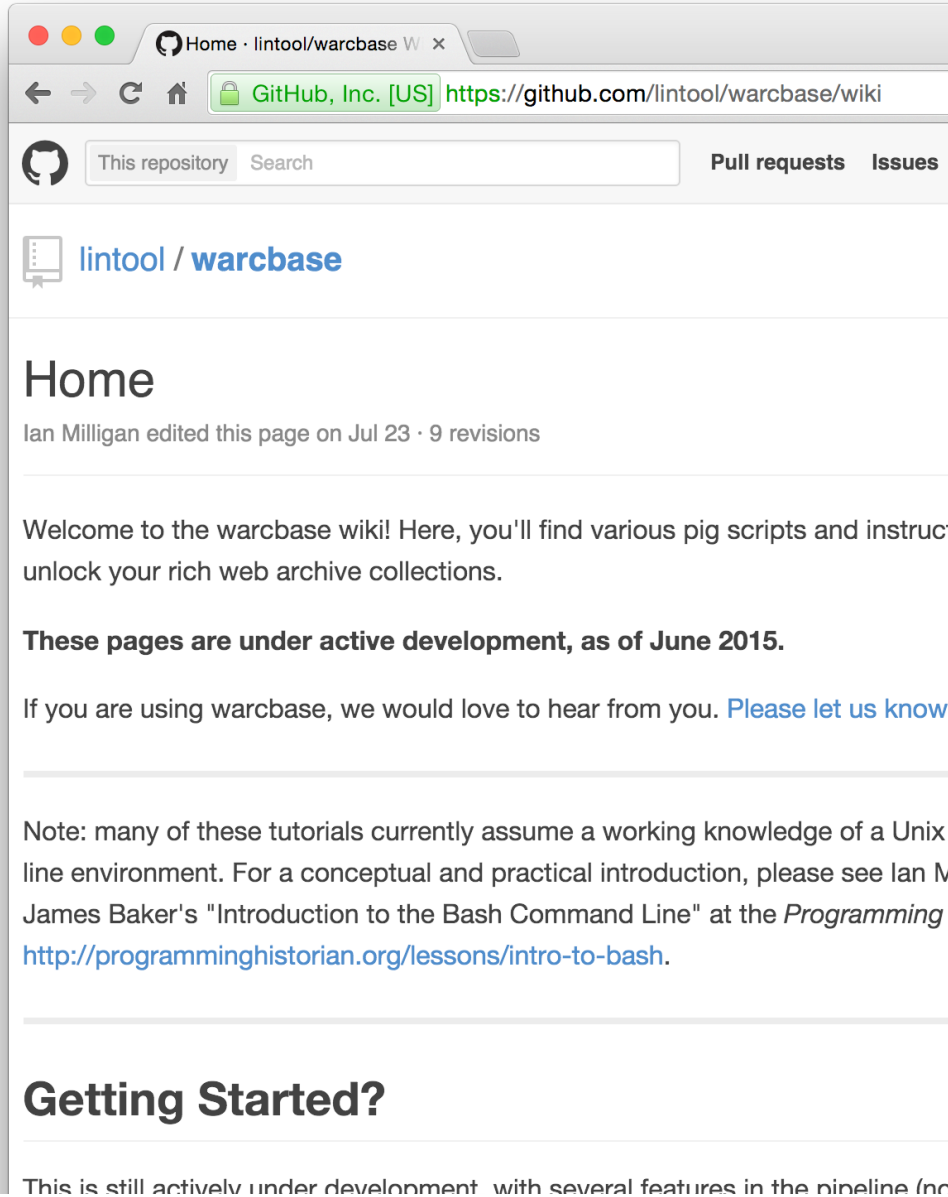
Metadata Extraction

- 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.

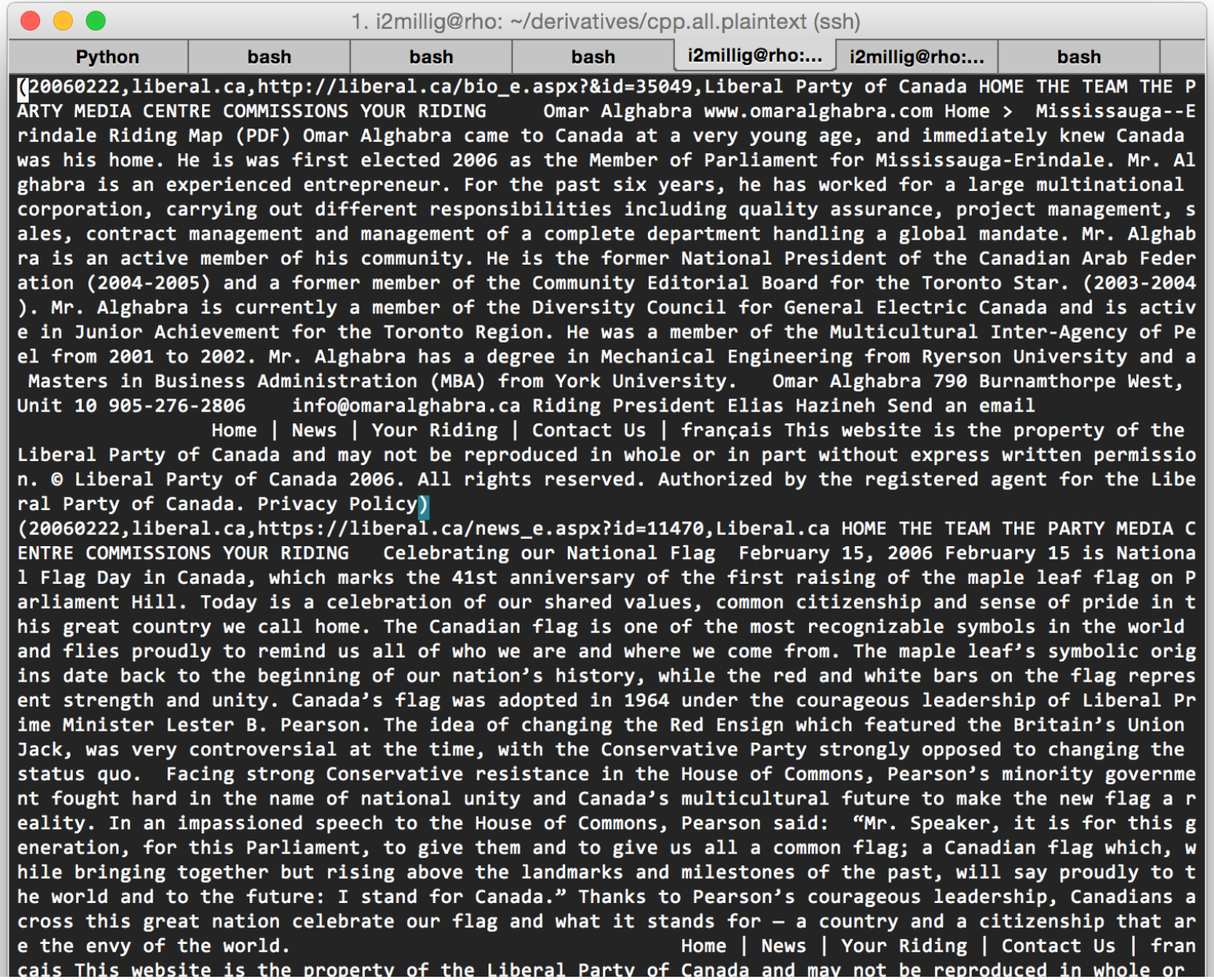
Interdisciplinarity

Warcbase

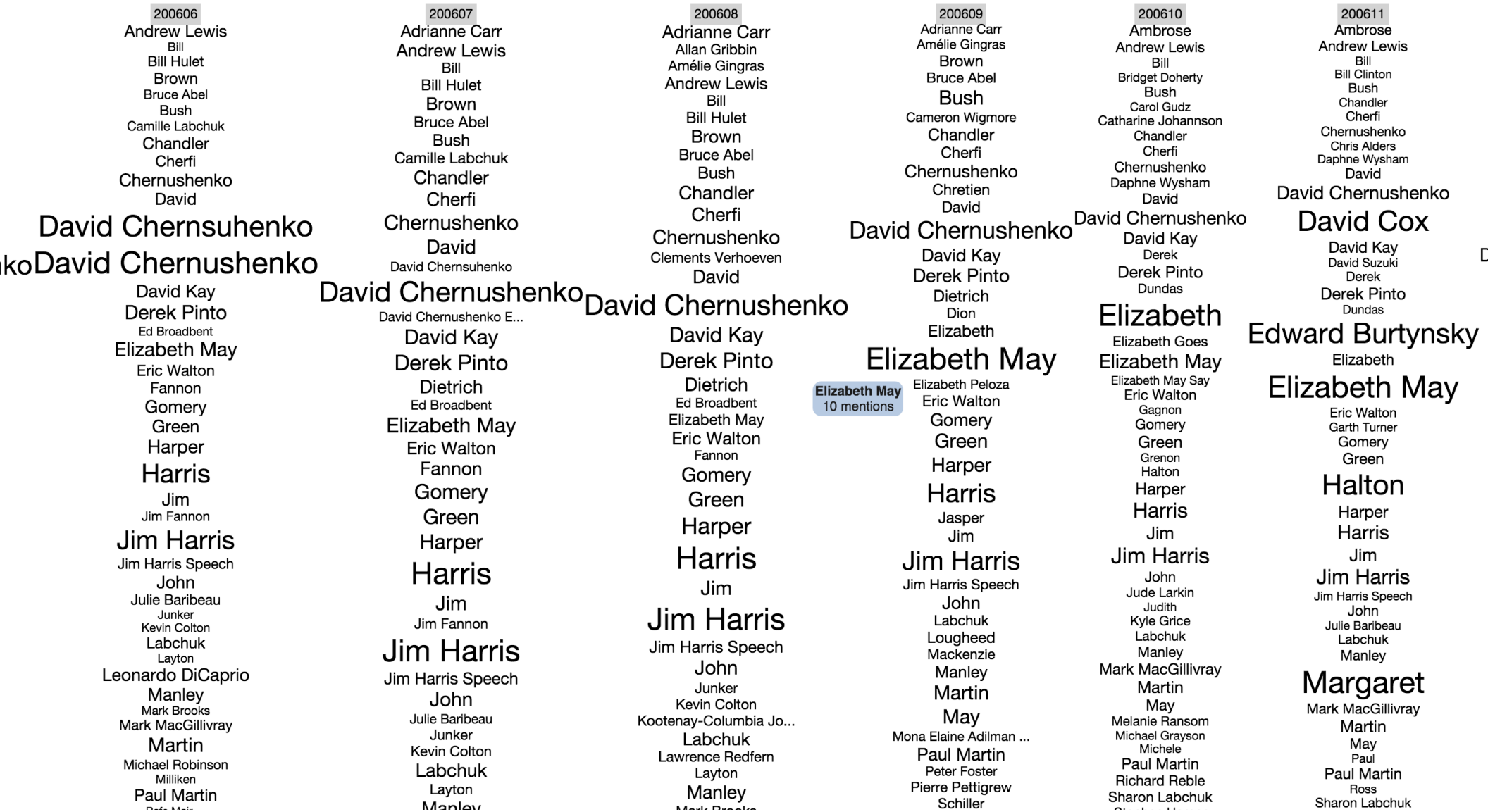
- Jimmy Lin (main developer), Ian Milligan, Jeremy Wiebe, Alice Zhou, all University of Waterloo
- Developing code, walkthroughs, and instructions for historians to be able to take a directory of WARCs and…



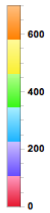
Extract all Plain Text



Extract Entities



Extract Entities

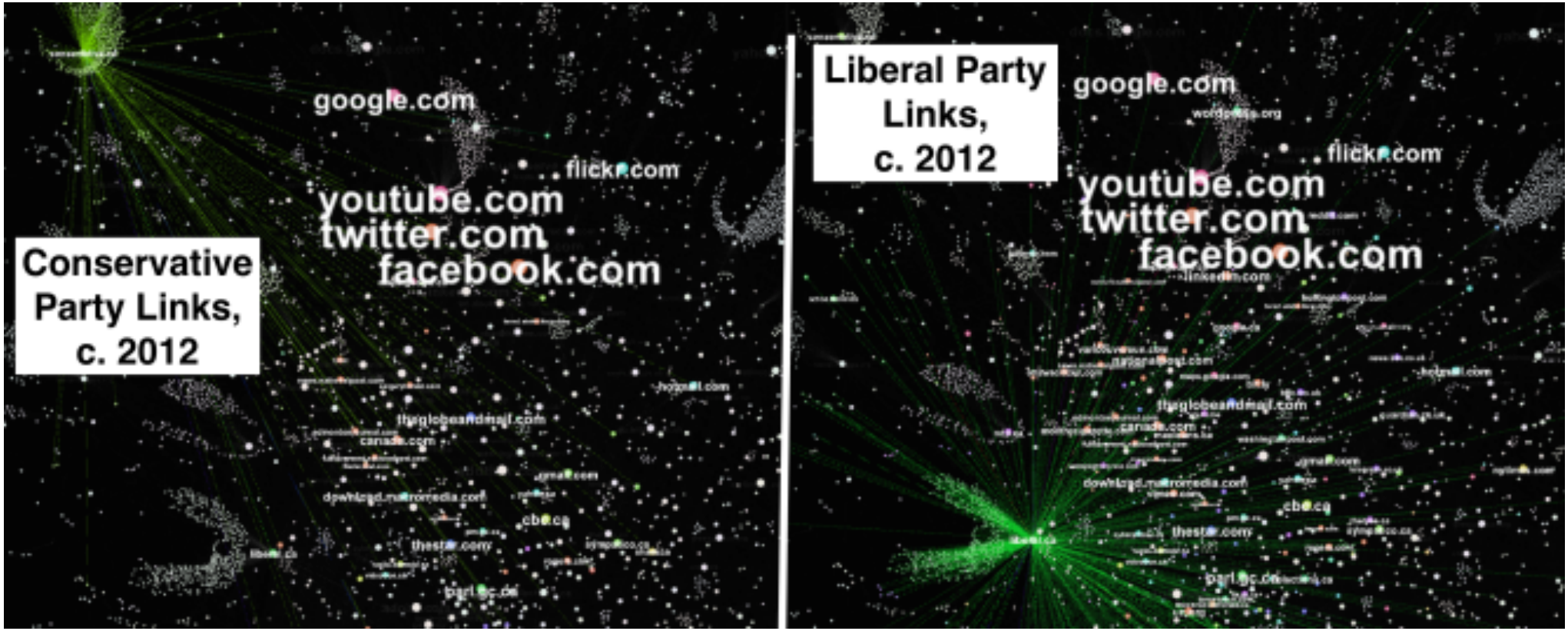



```
In[95]:= int = SemanticInterpretation[#] & /@ processedfreq[[All, 1]]
```

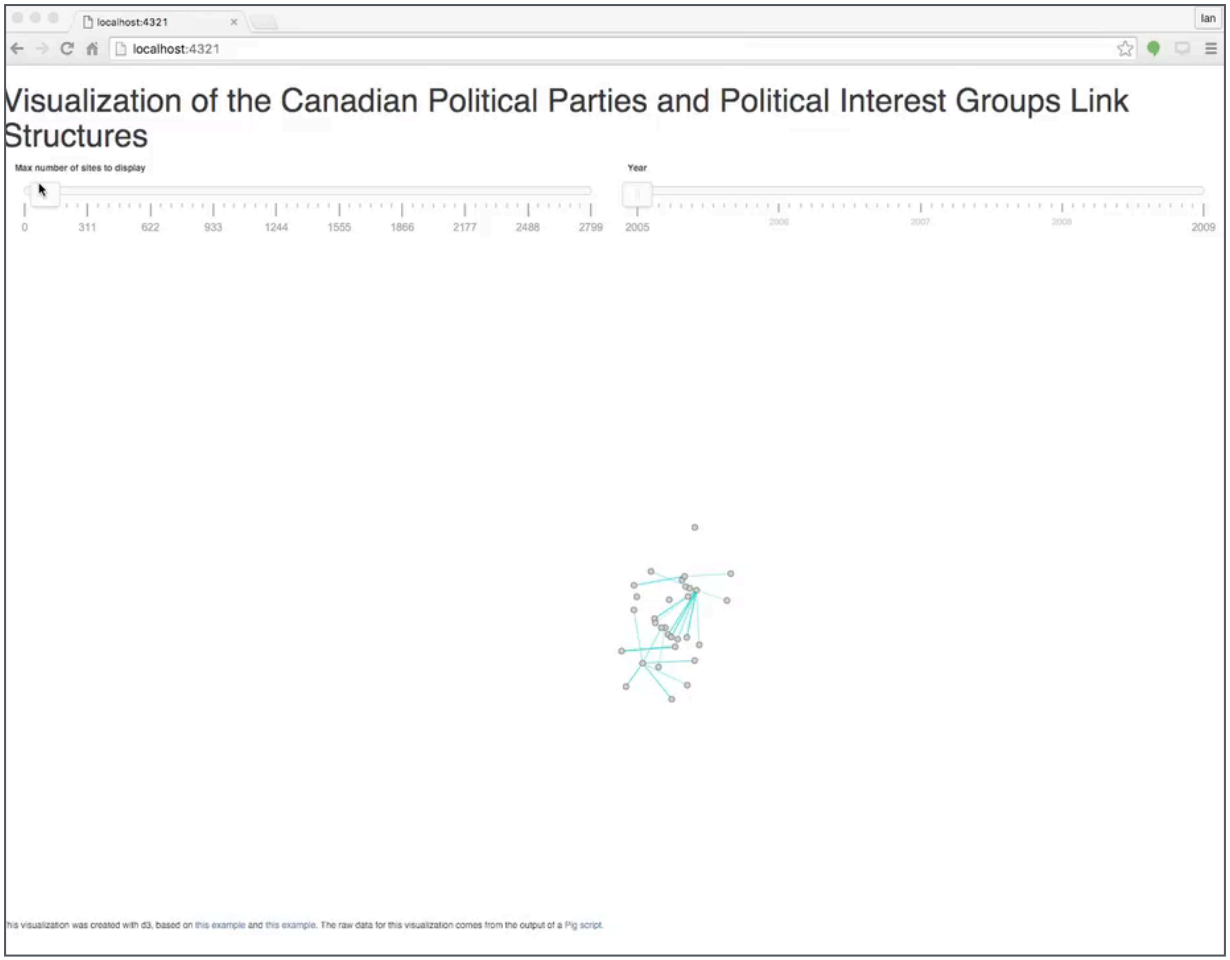
Out[95]= { Canada , Calgary , Colombia , Montreal , $Failed , Ontario, Canada , Afghanistan ,
Ottawa , Manitoba, Canada , British Columbia, Canada , Toronto , Nova Scotia, Canada ,
Saskatchewan, Canada , Quebec, Canada , Alberta, Canada , Winnipeg , Washington ,
Canada , Nunavut, Canada , White Horse , 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

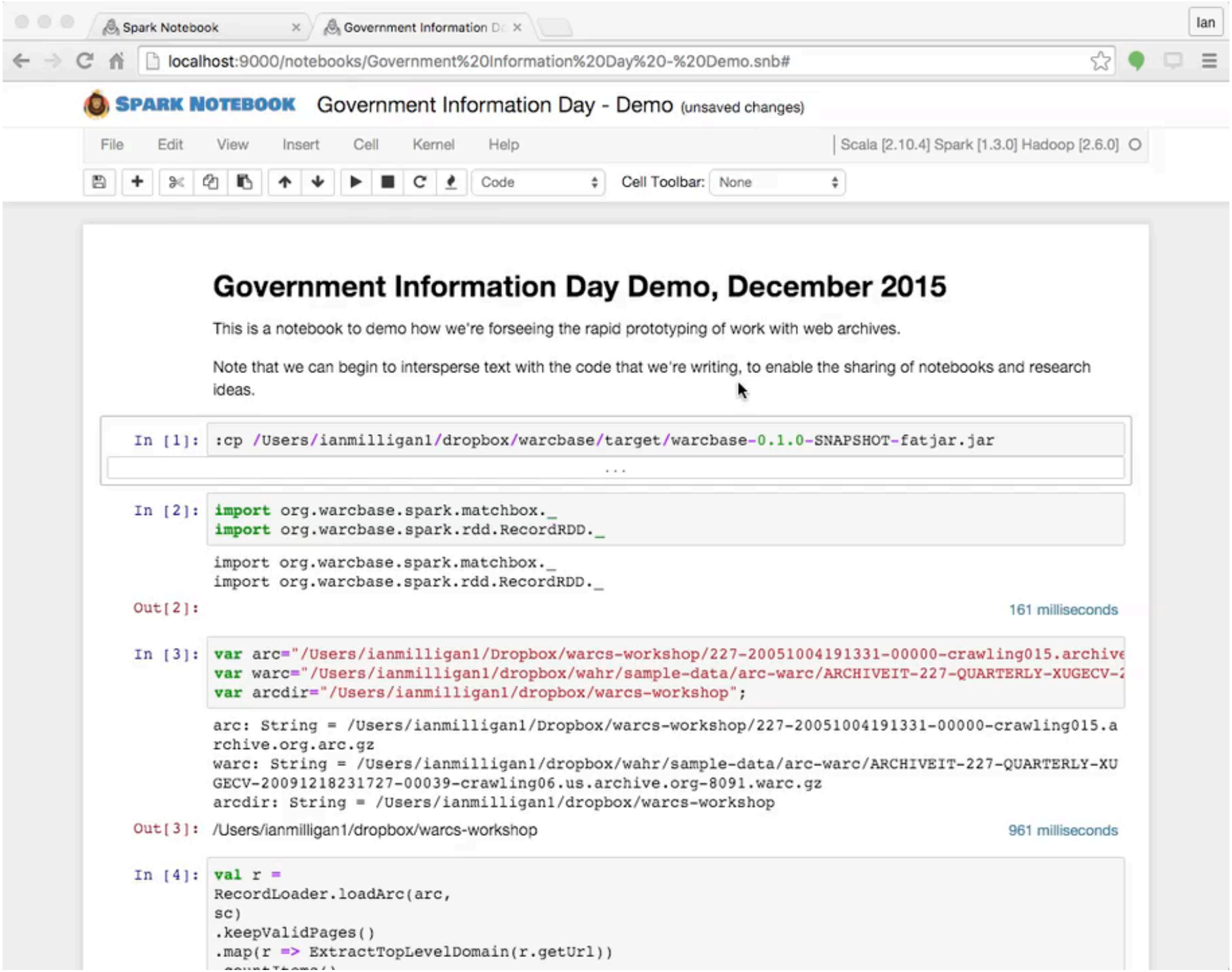
Extract Links/Gephi Connector



Or D3.js link networks in browser

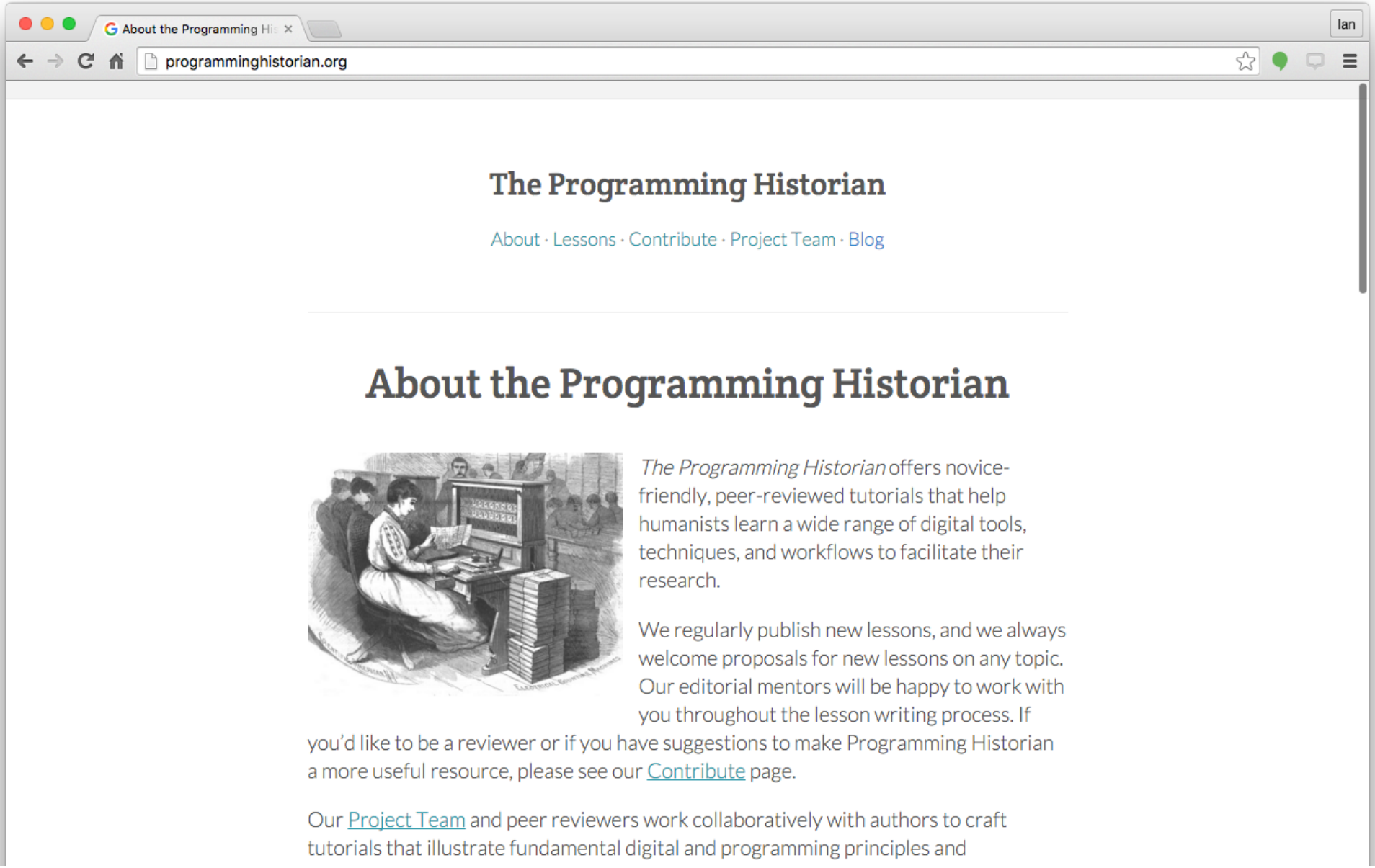


Bringing it all together in a notebook environment


Spark Notebook
Government Information D
Ian
localhost:9000/notebooks/Government%20Information%20Day%20-%20Demo.snb#
SPARK NOTEBOOK Government Information Day - Demo (unsaved changes)
File Edit View Insert Cell Kernel Help
Scala [2.10.4] Spark [1.3.0] Hadoop [2.6.0]
Code
Cell Toolbar: None
Government Information Day Demo, December 2015
This is a notebook to demo how we're forseeing the rapid prototyping of work with web archives.
Note that we can begin to intersperse text with the code that we're writing, to enable the sharing of notebooks and research ideas.
In [1]: :cp /Users/ianmilligan1/dropbox/warcbase/target/warcbase-0.1.0-SNAPSHOT-fatjar.jar
...
In [2]: import org.warcbase.spark.matchbox._
import org.warcbase.spark.rdd.RecordRDD._
import org.warcbase.spark.matchbox._
import org.warcbase.spark.rdd.RecordRDD._
Out[2]:
161 milliseconds
In [3]: var arc="/Users/ianmilligan1/Dropbox/warcs-workshop/227-20051004191331-00000-crawling015.archive
var warc="/Users/ianmilligan1/dropbox/wahr/sample-data/arc-warc/ARCHIVEIT-227-QUARTERLY-XUGECV-2
var arcdir="/Users/ianmilligan1/dropbox/warcs-workshop";
arc: String = /Users/ianmilligan1/Dropbox/warcs-workshop/227-20051004191331-00000-crawling015.a
rchive.org.arc.gz
warc: String = /Users/ianmilligan1/dropbox/wahr/sample-data/arc-warc/ARCHIVEIT-227-QUARTERLY-XU
GECV-20091218231727-00039-crawling06.us.archive.org-8091.warc.gz
arcdir: String = /Users/ianmilligan1/dropbox/warcs-workshop
Out[3]: /Users/ianmilligan1/dropbox/warcs-workshop
961 milliseconds
In [4]: val r =
RecordLoader.loadArc(arc,
sc)
.keepValidPages()
.map(r => ExtractTopLevelDomain(r.getUrl))

Walkthroughs at https://github.com/lintool/warcbase/wiki

Where to learn?

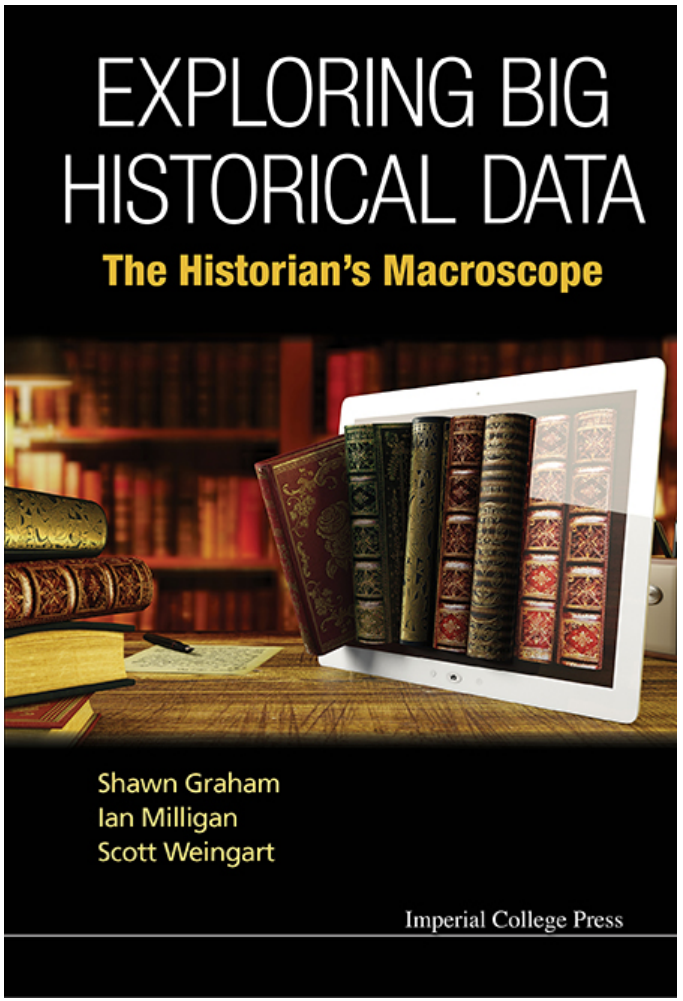


Programming Historian

- 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 can show 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