

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?**

g Google

https://www.google.ca/?gfe_rd=cr&ei=5Ak3VK6cHYqN8Qfo44CwBQ&gws_rd=ssl

+lan Gmail Images

Share

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 +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 Search +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%20Studies
The **Canadian History** Tunes CD. by Mary Ann Kelley. Designed by a certified teacher and professional performer turned homeschool mom, this educational ...

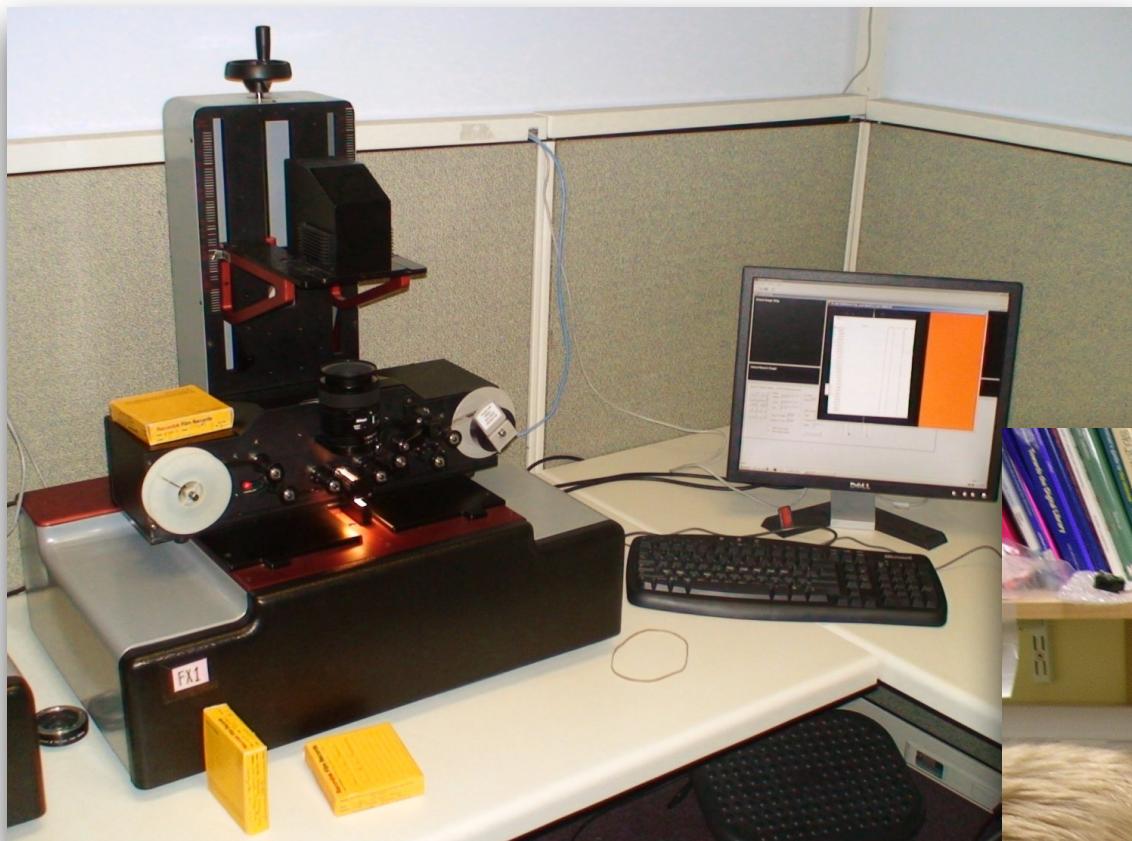
Canadian History - SchoolWorld an Edline Solution
teachersites.schoolworld.com/webpages/KStevenson2/resources.cfm?... ▾
Canadian History. Trenches on the web...an internet history of WWI · World War I ...
The Memory Project - Canada's Online Oral World War II History Project.

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/ [Search with the omitted results included](#)



Advanced Search - ProQuest

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

in Anywhere

in Anywhere

AND ([] OR []) AND ([] OR [])

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 *i*

Search | Clear form

Search subject areas

Use search forms customized for each subject.

The Arts

Business

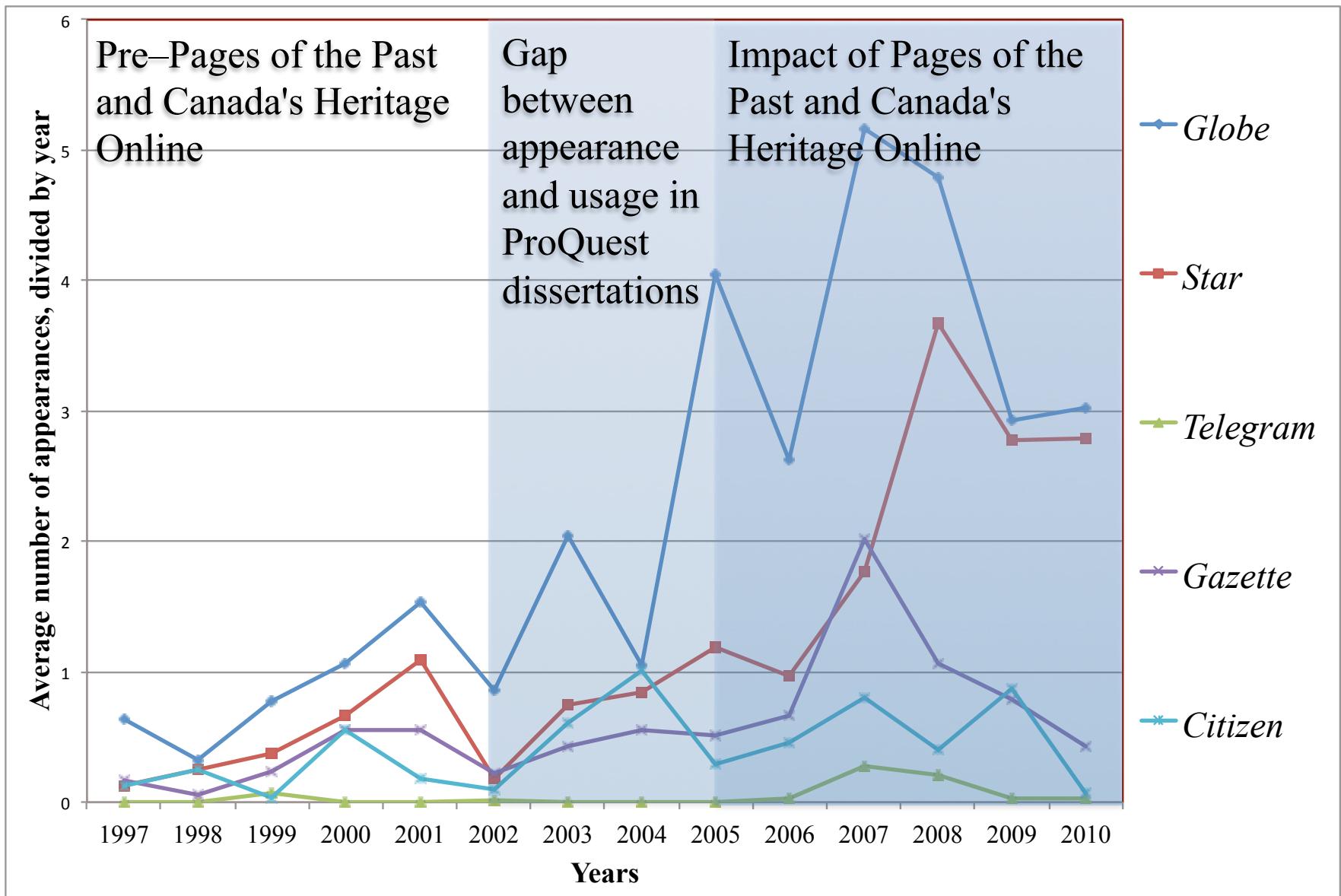
Dissertations & Theses

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
SERIES I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

370

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL
SERIES I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

371

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
SERIES I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

372

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
SERIES I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

373

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series II
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

374

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series II
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Subseries F Democratic Party
BOX 30

JOHN J. BURNS LIBRARY
BOSTON COLLEGE

375

Scarcity





WebARChive (WARC) File

HISTORIANS.ORG - The H

https://web.archive.org/web/19981212025605/http://www.historians.org/

INTERNET ARCHIVE
Wayback Machine

728 captures | 12 Dec 98 – 23 Nov 14

http://www.historians.org/ Go NOV DEC JAN 12 1998 2000 Close Help ?

[500 Historians Call for End to Impeachment Inquiry](#)

The Historians Committee for Open Debate

[About the HCFOD](#)

[How to Join](#)

[Directory](#)

[Press Releases](#)

[New Books](#)



[What's Hot!](#)

[Subscribe](#)

[On-Line Forum](#)

[Document Archive](#)

[Links](#)

www.
historians.org :
SEARCH

(by keyword, name, or subject)

Search Now
AND Insensitive

[For more information, contact info@historians.org](#)

© 1998 The Historians Committee for Open Debate. All rights reserved.

Waiting for web.archive.org...

You are viewing an archived web page, collected at the request of [Internet Archive Global Events](#) using [Archive-It](#). This page was captured on 5:07:27 Dec 03, 2011, and is part of the [Occupy Movement 2011/2012](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page. [Videos Metadata](#)

Welcome [login](#) | [signup](#)
Language [en](#) [es](#) [fr](#)

OccupyWallStreet

The revolution continues [worldwide!](#)

[News](#) [LiveStream](#) [#HowToOccupy](#) [Forum](#) [Chat](#) [User Map](#) [NYCGA](#) [About](#) [Donate](#)

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 [occupy movement: the liberation of vacant bank owned homes for those in need](#). The banks are

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](#)

You are viewing an archived web page, collected at the request of [Internet Archive Global Events](#) using [Archive-It](#). This page was captured on 5:07:27 Dec 03, 2011, and is part of the [Occupy Movement 2011/2012](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page. [Videos Metadata](#)

Welcome [login](#) | [signup](#)
Language [en](#) [es](#) [fr](#)

OccupyWallStreet

The revolution continues [worldwide!](#)

[News](#) [LiveStream](#) [#HowToOccupy](#) [Forum](#) [Chat](#) [User Map](#) [NYCGA](#) [About](#) [Donate](#)

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 [occupy movement: the liberation of vacant bank owned homes for those in need](#). The banks are

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.

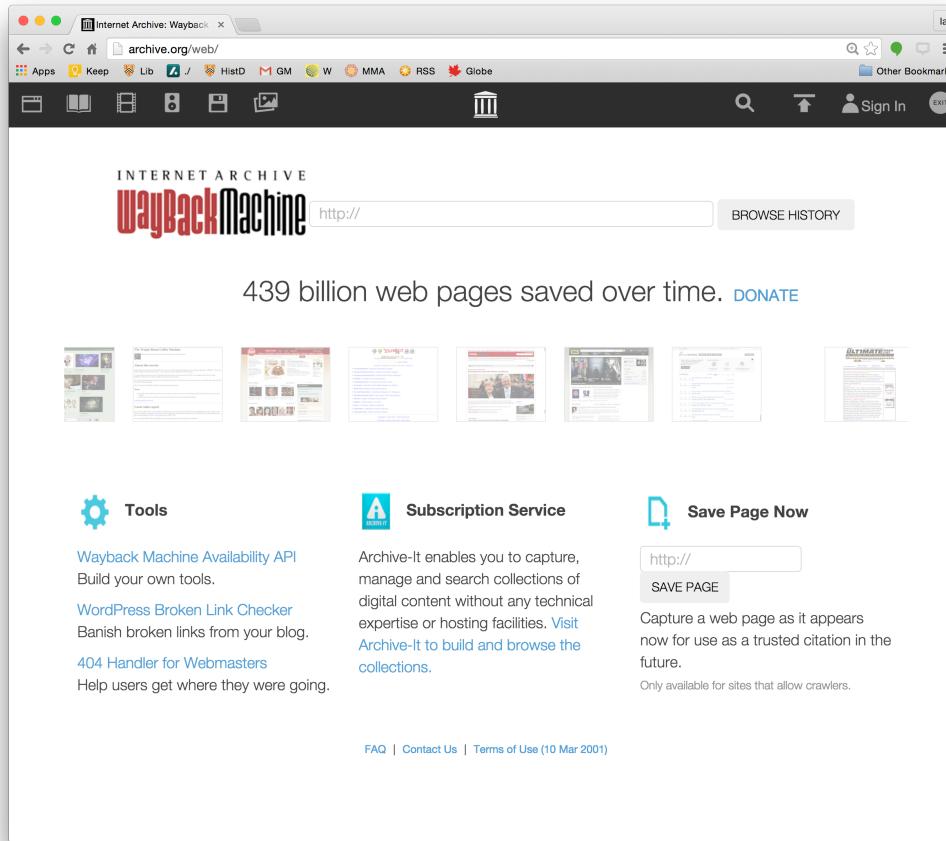
“.... [n]ow expectations have inverted. Everything may be recorded and preserved, at least potentially.”

- 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...

The screenshot shows a web browser window with the title "Archive-It - Canadian Political Parties and Groups". The URL in the address bar is <https://archive-it.org/collections/227>. The page features a header with the Archive-It logo, navigation links for HOME, EXPLORE, LEARN MORE, and CONTACT US, and a sub-navigation link for "Explore > University of Toronto > Canadian Political Parties and Political Interest Groups". Below this is a large green banner with the text "Canadian Political Parties and Groups" and "Collected by: University of Toronto". It also includes information about the collection being archived since Oct, 2005, and describing it as a collection of Canadian political parties and interest groups. A "Narrow Your Results" section lists categories like New Democratic Party of Canada (2), Assembly of First Nations (1), Bloc Québécois (1), Canada First (1), and Canada West Foundation (1). At the bottom, there are buttons for "Sites" and "Search Page Text", and a footer indicating "Page 1 of 1 (54 Total)".

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!

The screenshot shows a web browser window displaying the Archive-It.org collection page for "Canadian Political Parties and Political Interest Groups". The URL in the address bar is <https://archive-it.org/collections/227?q=Stephen+Harper&page=1&show=Sites>. The page features a header with the Archive-It logo, navigation links for HOME, EXPLORE, LEARN MORE, and CONTACT US, and a tagline about collecting and accessing cultural heritage on the web. Below the header, it says "Explore >> University of Toronto >> Canadian Political Parties and Political Interest Groups". The main content area displays the "Canadian Political Parties and Political Interest Groups" collection, which was collected by the University of Toronto and archived since Oct, 2005. It includes a brief description, subject terms (Politics & Elections), and a collector note. A search bar at the bottom left allows users to search within the collection. The results page shows a search term "Stephen Harper" and a count of 1,213,132 matches. The results are sorted by Best Match. One result is highlighted: "Stephen Harper | Facebook", with a link to <http://www.facebook.com/pages/Stephen-Harper/9106562109>.

The Internet Archive will la Ian

www.theverge.com/2015/10/22/9593656/internet-archive-wayback-machine-redesign-announced

THE VERGE

TRENDING NOW
Amazon is opening its first physical bookstore today

34 NEW ARTICLES

PREVIOUS STORY
FCC passes rule cracking down on prison phone call charges

NEXT STORY
Researchers discover new attacks amid VoLTE rollout

TECH

The Internet Archive will launch a modernized Wayback Machine in 2017

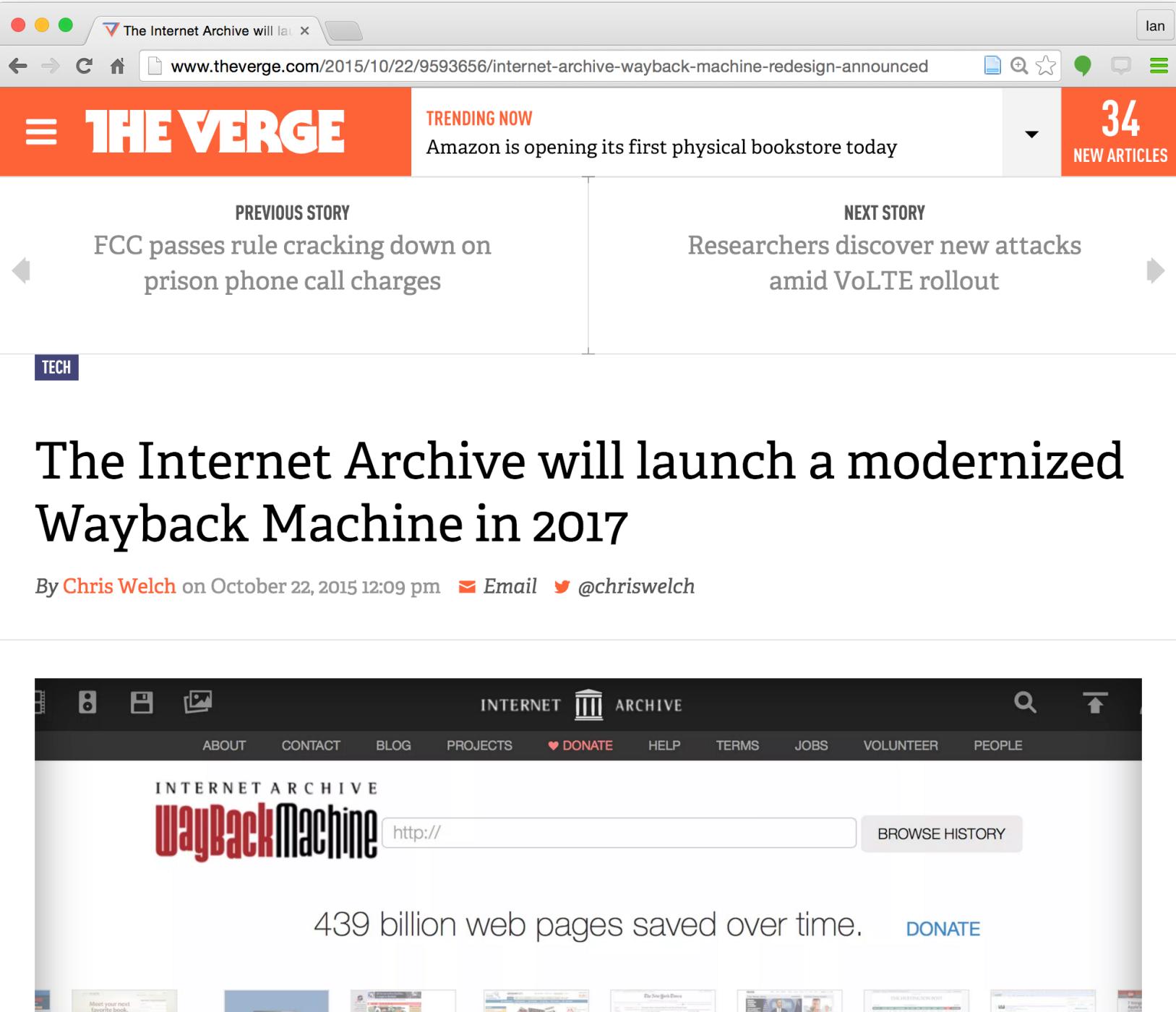
By Chris Welch on October 22, 2015 12:09 pm Email @chriswelch

INTERNET ARCHIVE

ABOUT CONTACT BLOG PROJECTS DONATE HELP TERMS JOBS VOLUNTEER PEOPLE

INTERNET ARCHIVE WayBack Machine http:// BROWSE HISTORY

439 billion web pages saved over time. DONATE



ukwa/shine lan

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

File	Description	Time
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>

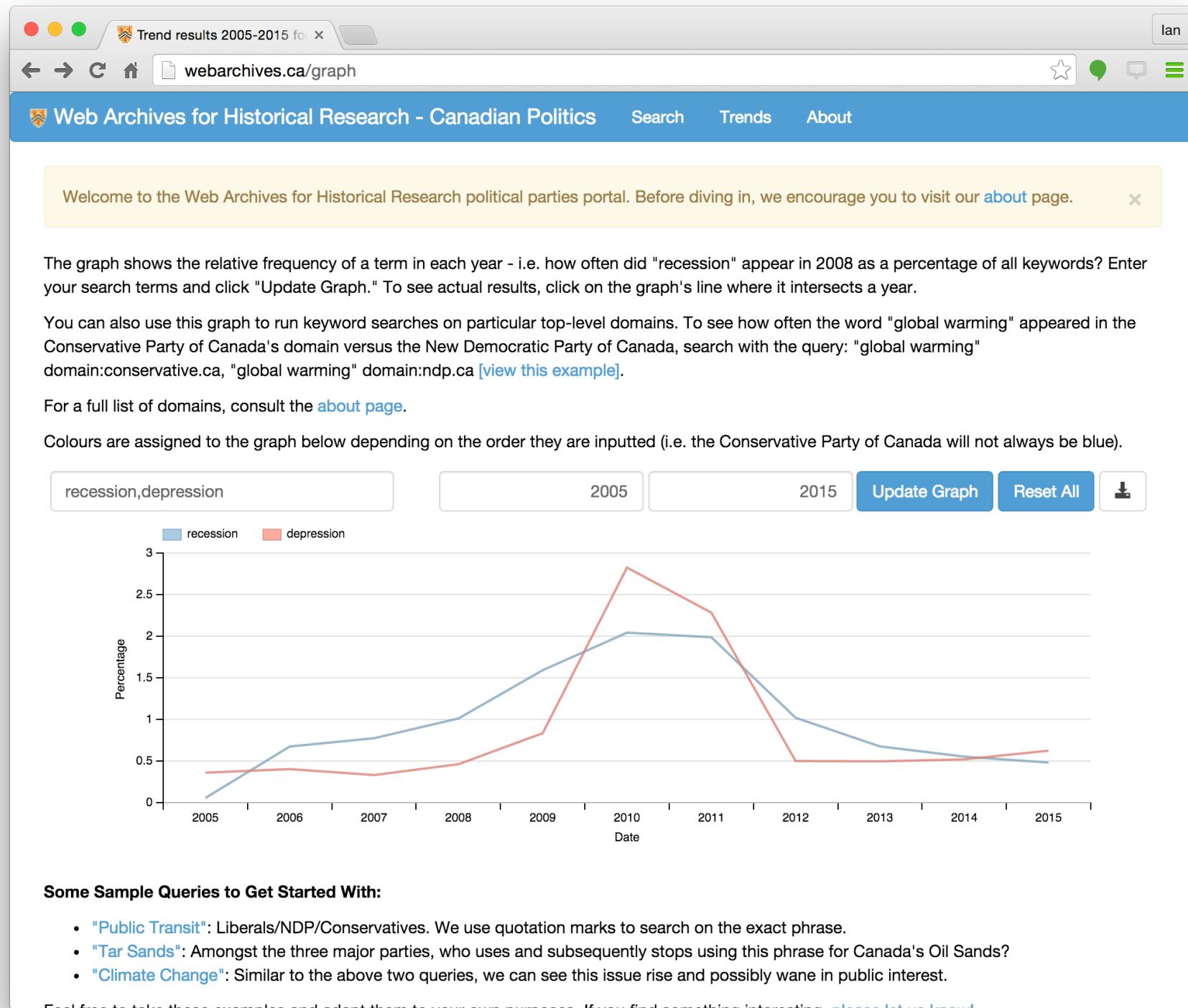
You can clone with [HTTPS](#), [SSH](#), or [Subversion](#).

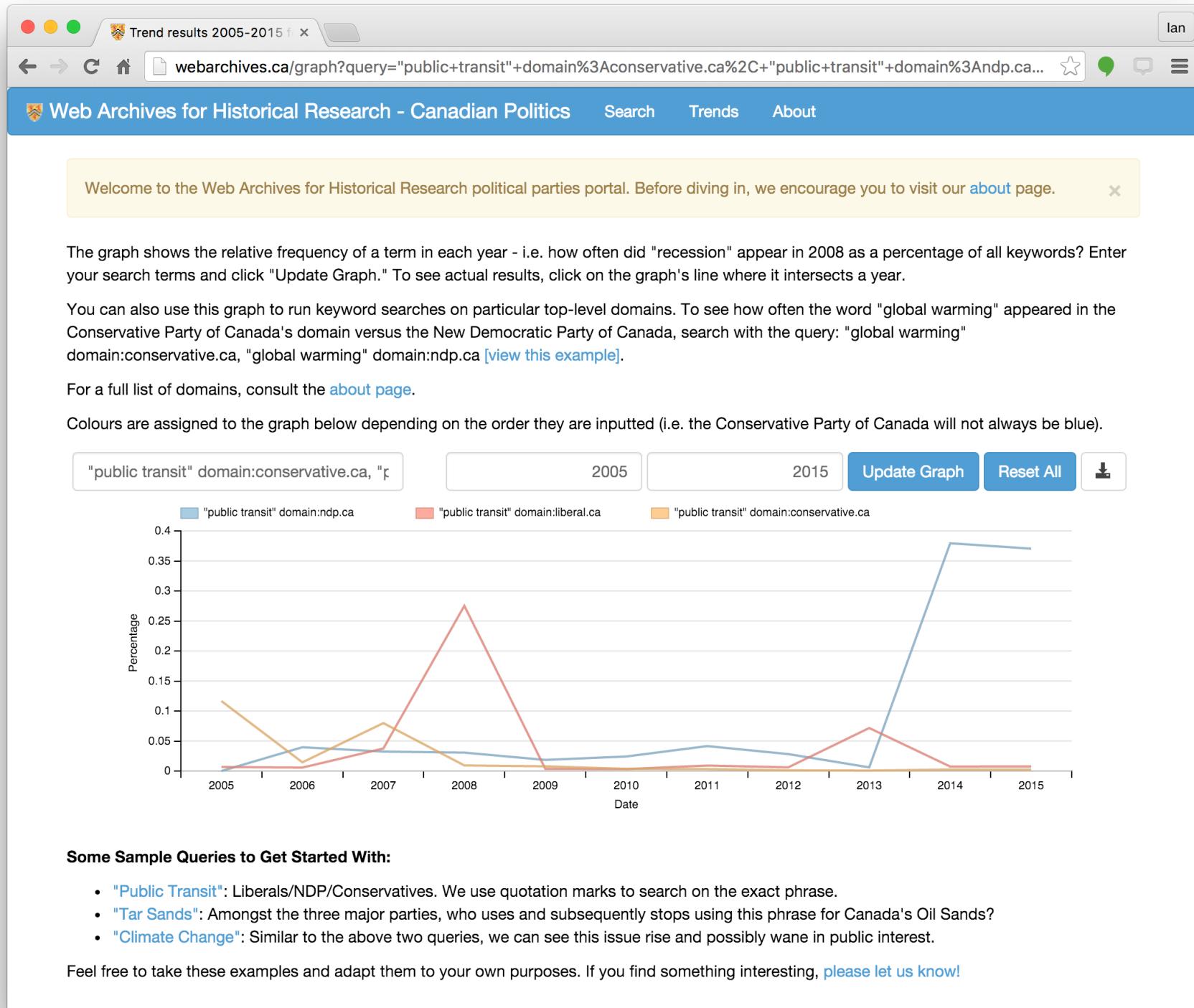
[Clone in Desktop](#)

[Download ZIP](#)



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. X

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

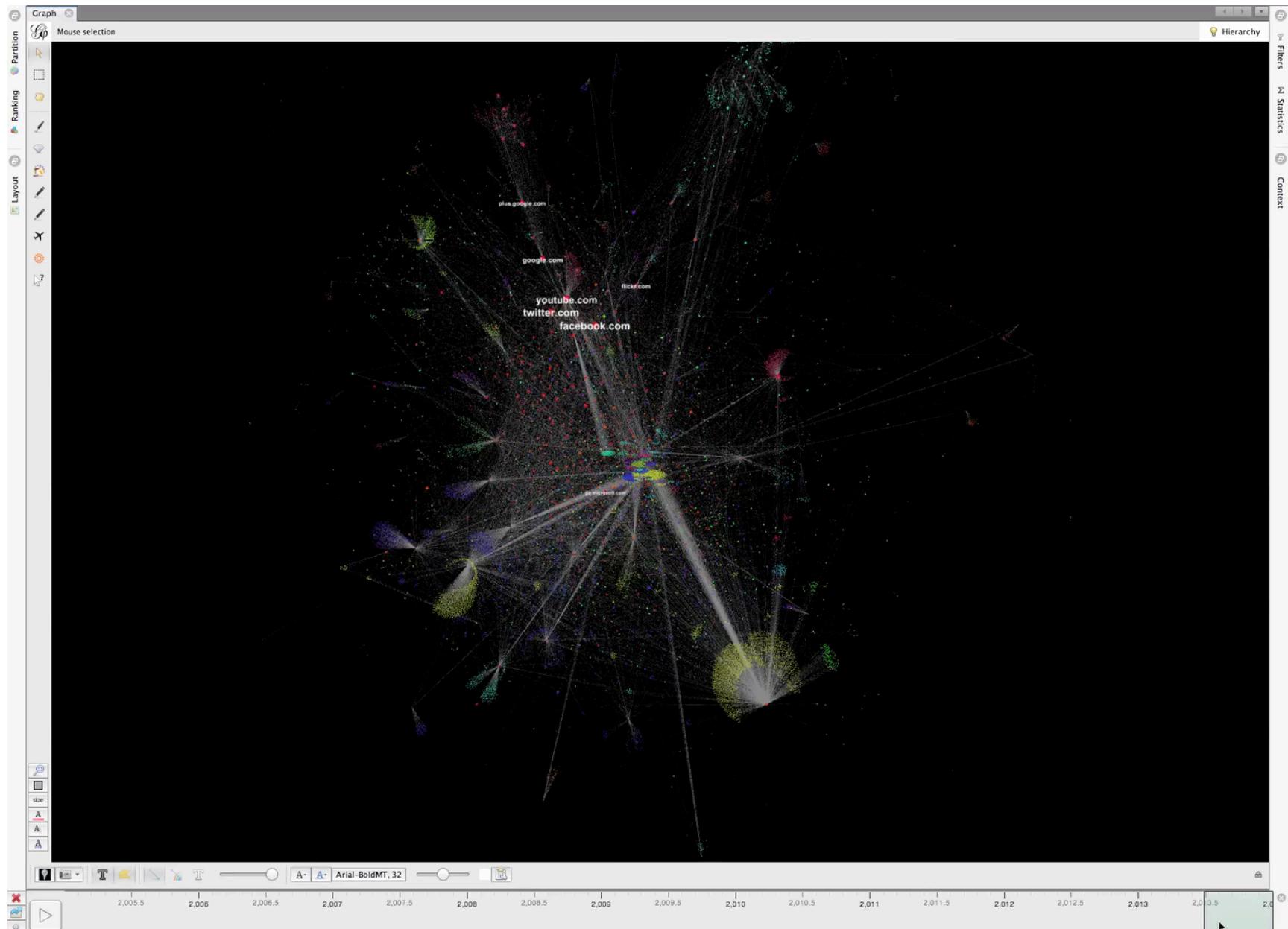
Good for public engagement - but limited for scholarship....

The screenshot shows a web browser window with the following details:

- Address Bar:** webarchives.ca/search?query=stephen+harper&tab=results&action=search
- Page Title:** Web Archives for Historical Research - Canadian Politics
- Header:** Search, Trends, About
- Welcome Message:** Welcome to the Web Archives for Historical Research political parties portal. Before diving in, we encourage you to visit our [about](#) page.
- Search Options:** Search, Advanced Search
- General Content Type:** html (1,085,201), other (71,691), pdf (3,947), audio (341), text (106), image (14)
- Sample Mode:** stephen harper (Search, Reset)
- Search Term(s):** stephen harper
- Crawl Years:** 2008 (443,448), 2010 (142,609), 2007 (109,236), 2006 (104,564), 2011 (83,910), 2014 (70,746)
- Navigation:** Results, Concordance
- Results Summary:** Results 1 to 10 of 1,161,300
- Download Options:** CSV, Asc

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

Metadata Extraction



December 2006

Stephane Dion Elected Leader of Party



December 2007
Rise of Social Media



April 2008

Fundraising with the Victory Fund/ Fonds de la Victoire



July 2008

The Green Shift Announced!



October 2008

Election Campaign - Advertisement Sites

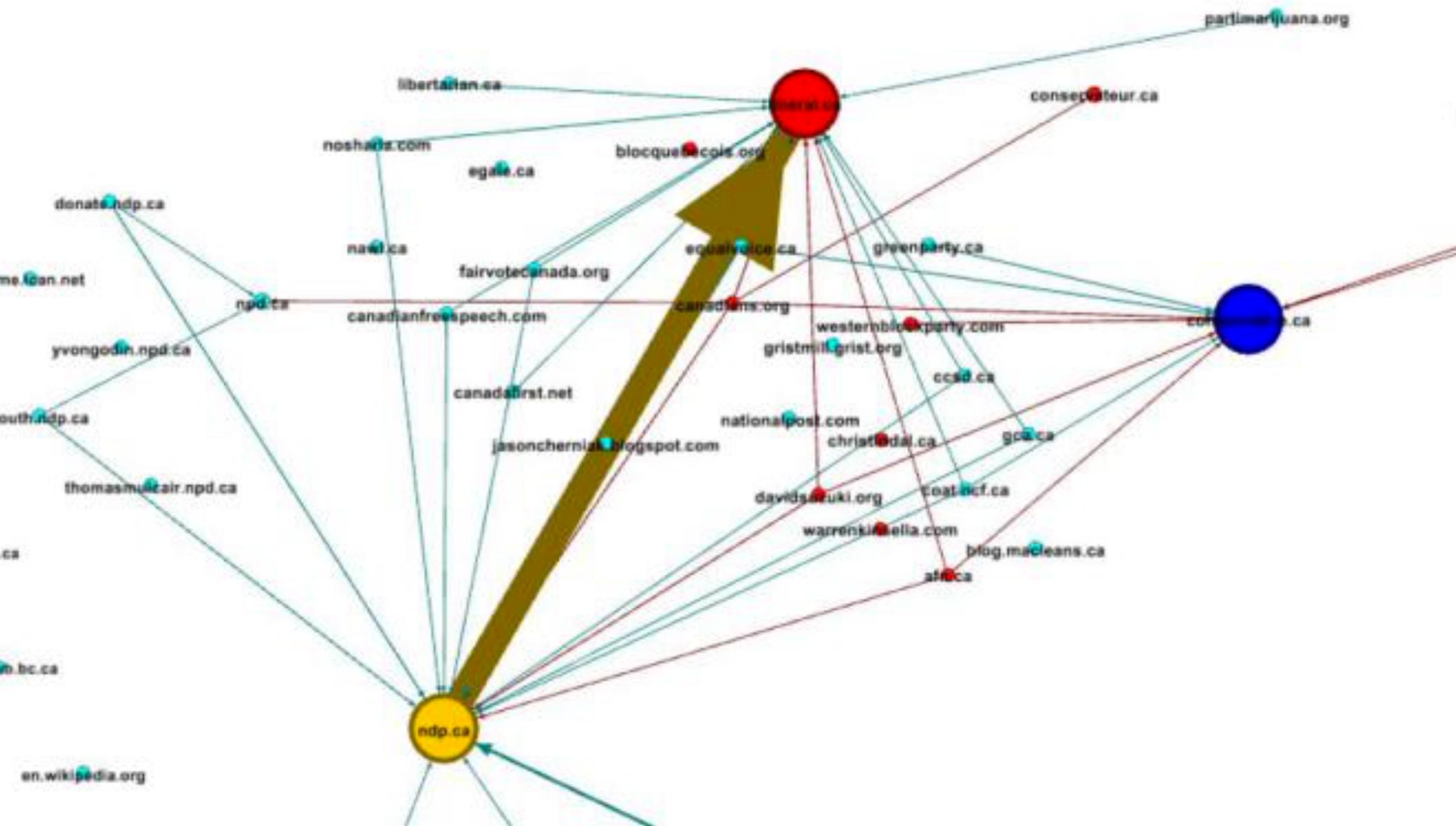


December 2008

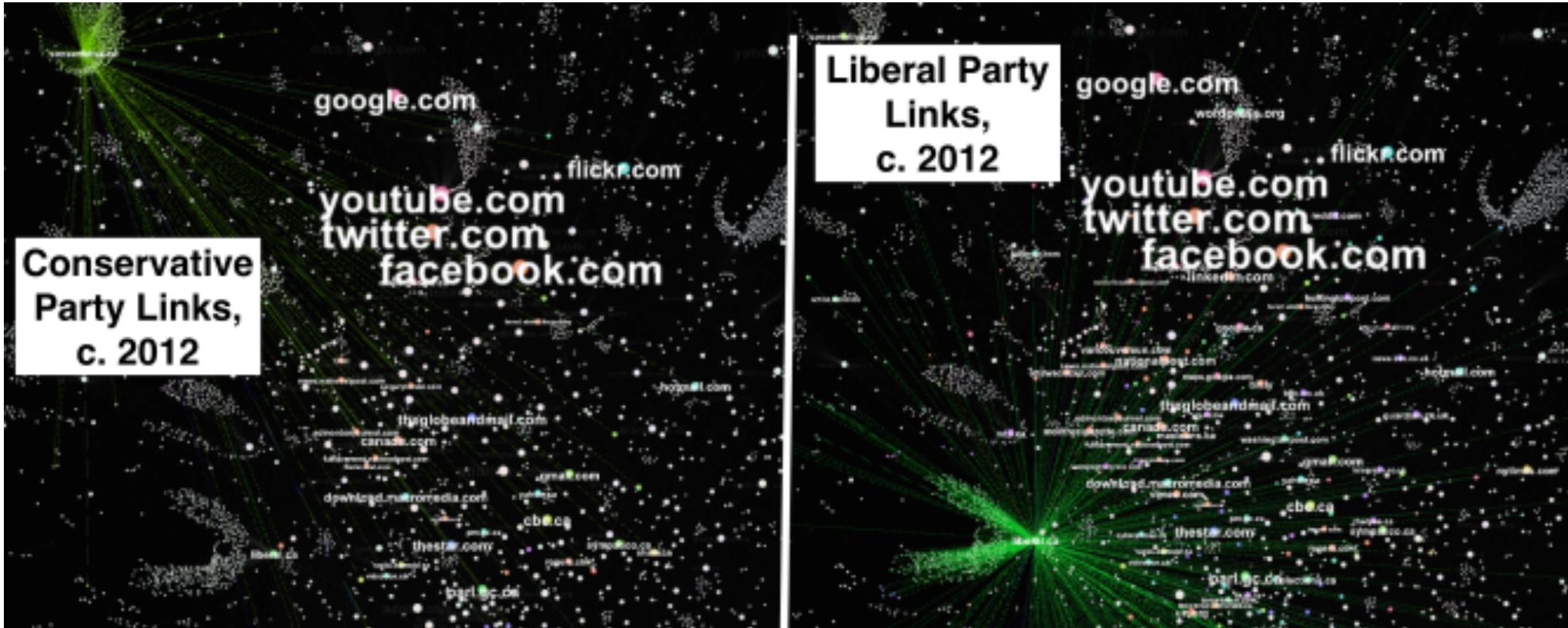
Election campaign Ends; Attacking Harper on Anti-American Grounds (bushharper)



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.

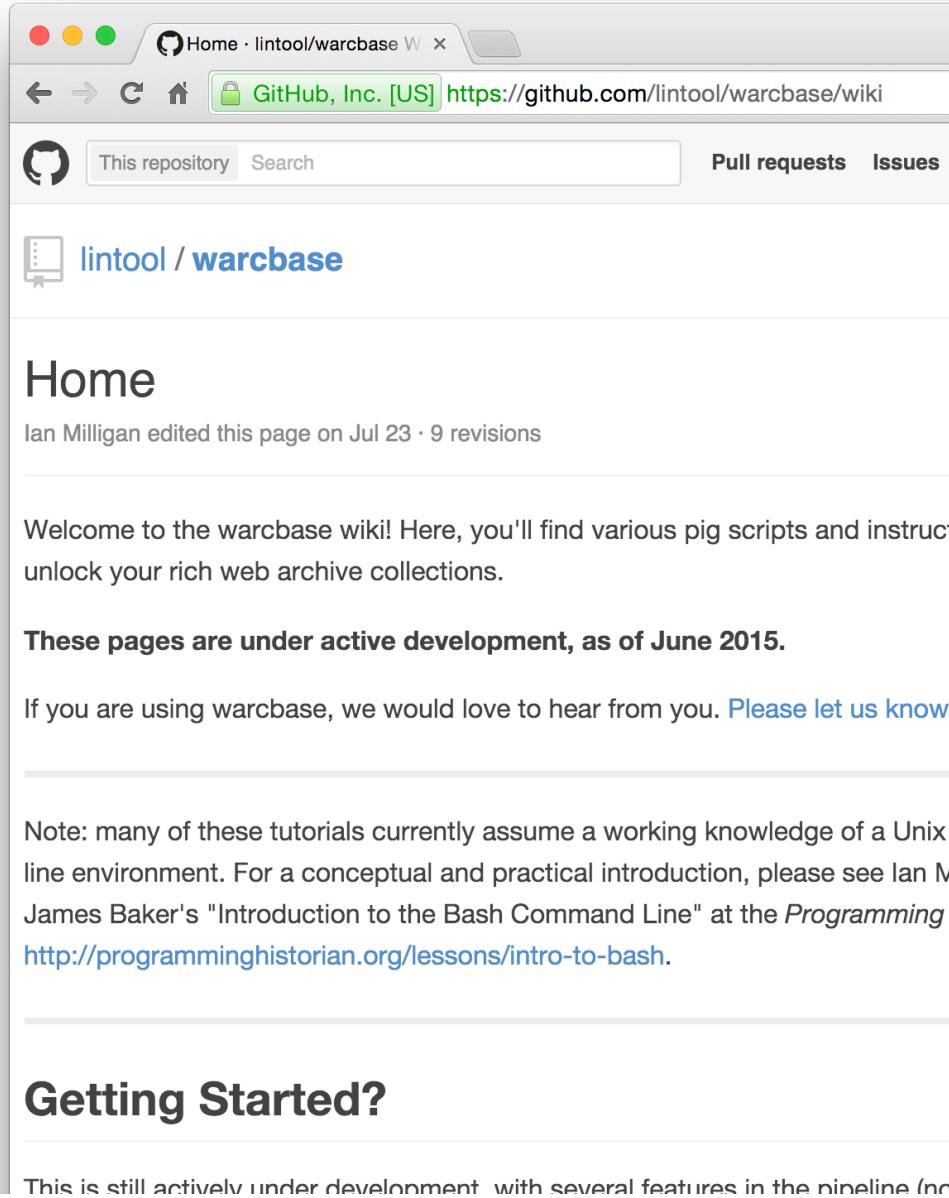
Metadata Extraction

- 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...



The screenshot shows a Mac OS X window displaying a GitHub wiki page. The title bar reads "Home · lintool/warcbase". The address bar shows "GitHub, Inc. [US] https://github.com/lintool/warcbase/wiki". The main content area is titled "warcbase" and contains the following text:

Home
Ian Milligan edited this page on Jul 23 · 9 revisions

Welcome to the warcbase wiki! Here, you'll find various pig scripts and instructions to unlock your rich web archive collections.

These pages are under active development, as of June 2015.

If you are using warcbase, we would love to hear from you. [Please let us know](#)

Note: many of these tutorials currently assume a working knowledge of a Unix line environment. For a conceptual and practical introduction, please see Ian MacLennan and James Baker's "Introduction to the Bash Command Line" at the [Programming Historian](http://programminghistorian.org/lessons/intro-to-bash).

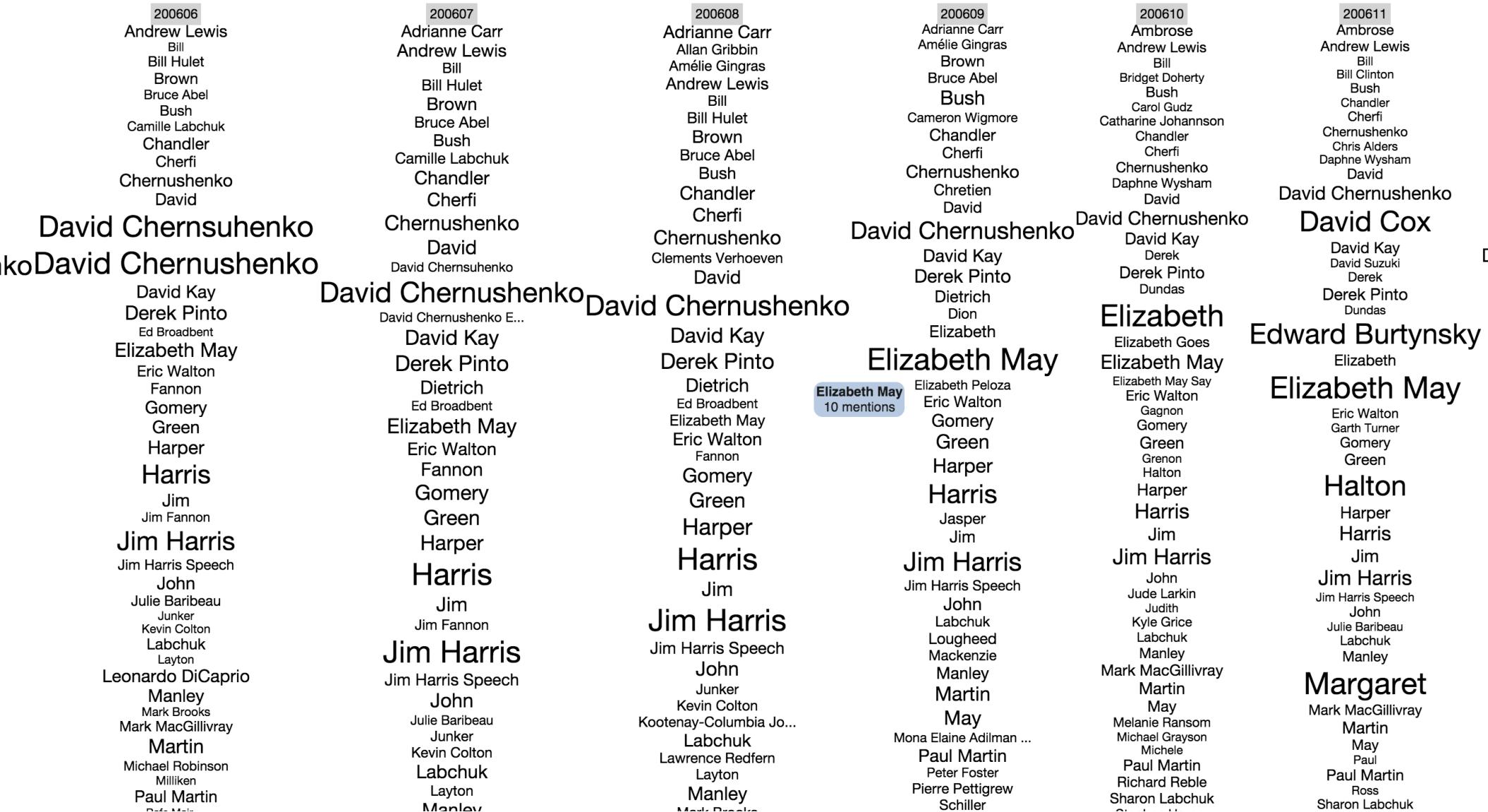
Getting Started?

This is still actively under development with several features in the pipeline (no

Extract all Plain Text

```
1. i2millig@rho: ~/derivatives/cpp.all.plaintext (ssh)
Python      bash      bash      bash      i2millig@rho:... i2millig@rho:... bash
(20060222,liberal.ca,http://liberal.ca/bio_e.aspx?id=35049,Liberal Party of Canada HOME THE TEAM THE P
ARTY MEDIA CENTRE COMMISSIONS YOUR RIDING      Omar Alghabra www.omaralghabra.com Home > Mississauga--E
rindale Riding Map (PDF) Omar Alghabra came to Canada at a very young age, and immediately knew Canada
was his home. He was first elected 2006 as the Member of Parliament for Mississauga-Erindale. Mr. Al
ghabra is an experienced entrepreneur. For the past six years, he has worked for a large multinational
corporation, carrying out different responsibilities including quality assurance, project management, s
ales, contract management and management of a complete department handling a global mandate. Mr. Alghab
ra is an active member of his community. He is the former National President of the Canadian Arab Feder
ation (2004-2005) and a former member of the Community Editorial Board for the Toronto Star. (2003-2004
). Mr. Alghabra is currently a member of the Diversity Council for General Electric Canada and is activ
e in Junior Achievement for the Toronto Region. He was a member of the Multicultural Inter-Agency of Pe
ople from 2001 to 2002. Mr. Alghabra has a degree in Mechanical Engineering from Ryerson University and a
Masters in Business Administration (MBA) from York University.    Omar Alghabra 790 Burnamthorpe West,
Unit 10 905-276-2806   info@omaralghabra.ca Riding President Elias Hazineh Send an email
Home | News | Your Riding | Contact Us | français This website is the property of the
Liberal Party of Canada and may not be reproduced in whole or in part without express written permission.
© Liberal Party of Canada 2006. All rights reserved. Authorized by the registered agent for the Libe
ral Party of Canada. Privacy Policy)
(20060222,liberal.ca,https://liberal.ca/news_e.aspx?id=11470,Liberal.ca HOME THE TEAM THE PARTY MEDIA C
ENTRE COMMISSIONS YOUR RIDING  Celebrating our National Flag  February 15, 2006 February 15 is Nationa
l Flag Day in Canada, which marks the 41st anniversary of the first raising of the maple leaf flag on P
arliament Hill. Today is a celebration of our shared values, common citizenship and sense of pride in t
his great country we call home. The Canadian flag is one of the most recognizable symbols in the world
and flies proudly to remind us all of who we are and where we come from. The maple leaf's symbolic orig
ins date back to the beginning of our nation's history, while the red and white bars on the flag repres
ent strength and unity. Canada's flag was adopted in 1964 under the courageous leadership of Liberal Pr
ime Minister Lester B. Pearson. The idea of changing the Red Ensign which featured the Britain's Union
Jack, was very controversial at the time, with the Conservative Party strongly opposed to changing the
status quo. Facing strong Conservative resistance in the House of Commons, Pearson's minority governme
nt fought hard in the name of national unity and Canada's multicultural future to make the new flag a r
eality. In an impassioned speech to the House of Commons, Pearson said: "Mr. Speaker, it is for this g
eneration, for this Parliament, to give them and to give us all a common flag; a Canadian flag which, w
hile bringing together but rising above the landmarks and milestones of the past, will say proudly to t
he world and to the future: I stand for Canada." Thanks to Pearson's courageous leadership, Canadians a
cross this great nation celebrate our flag and what it stands for – a country and a citizenship that ar
e the envy of the world.
Home | News | Your Riding | Contact Us | fran
cais This website is the property of the Liberal Party of Canada and may not be reproduced in whole or
```

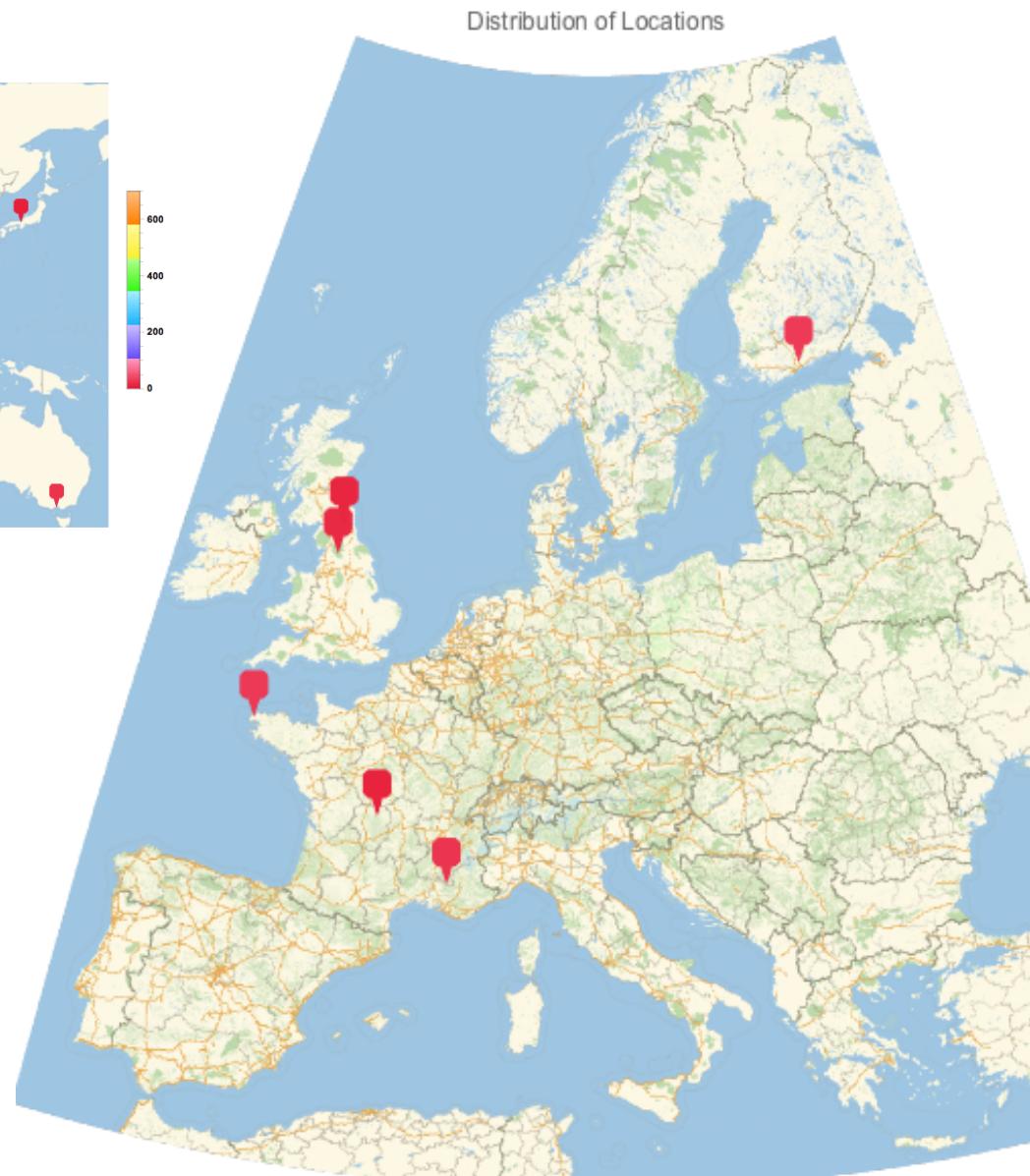
Extract Entities



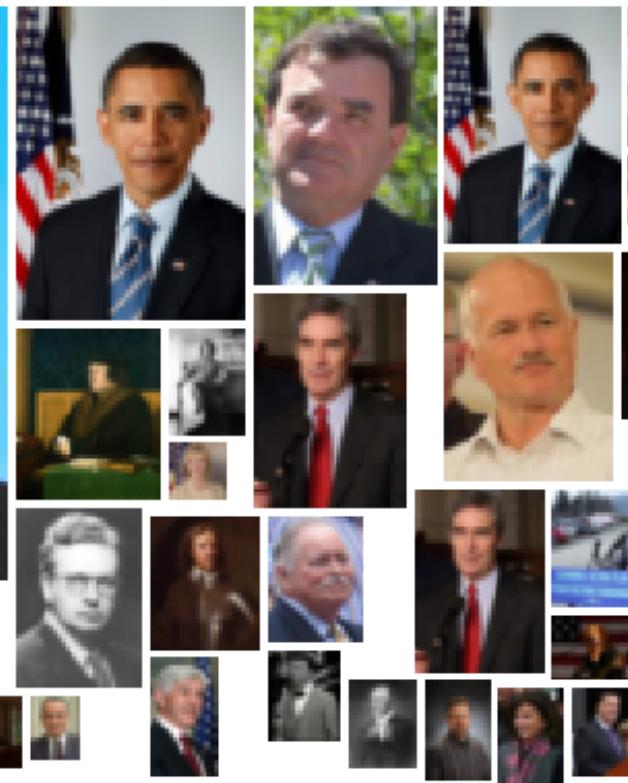
Extract Entities



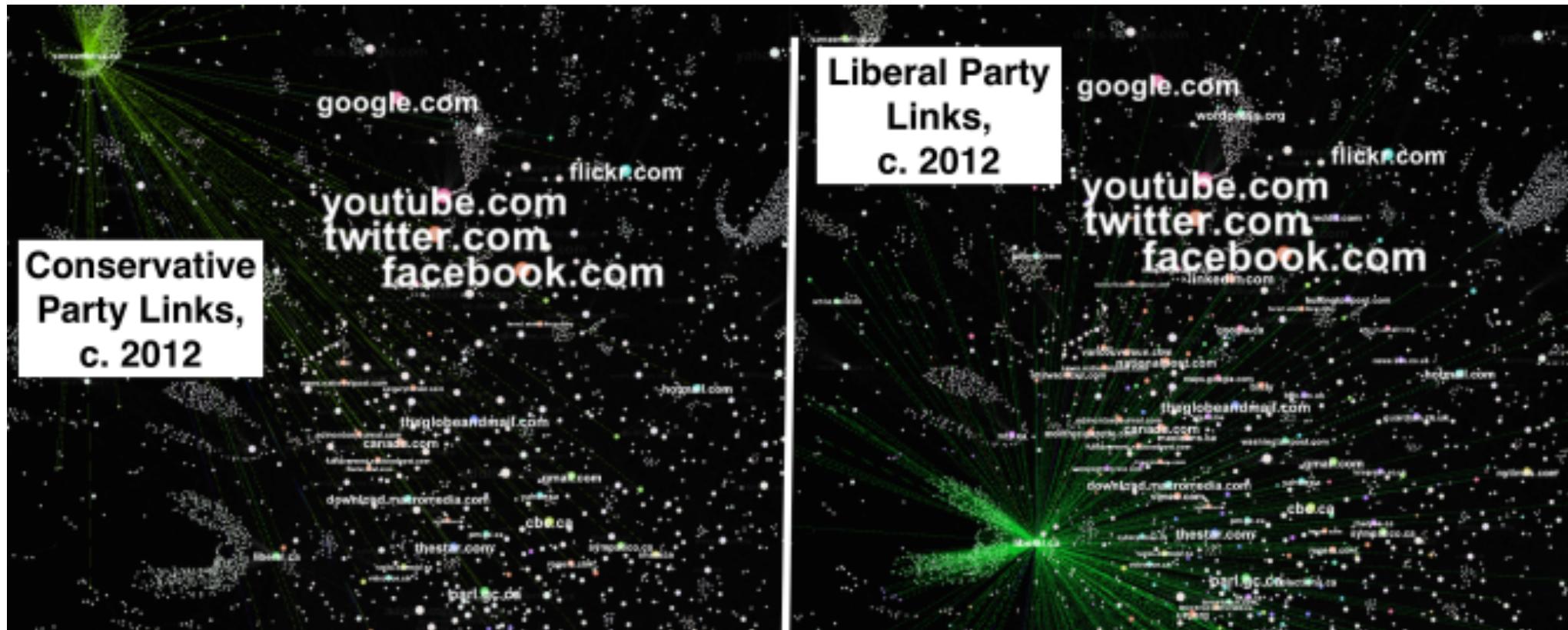
```
In[95]:= int = SemanticInterpretation[#] & /@ processedfreq[[All, 1]]  
Out[95]= {Canada, Calgary, Colombia, Montreal, $Failed, Ontario, Canada, Afghanistan,  
Ottawa, Manitoba, Canada, British Columbia, Canada, Toronto, Nova Scotia, Canada,  
Saskatchewan, Canada, Quebec, Canada, Alberta, Canada,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



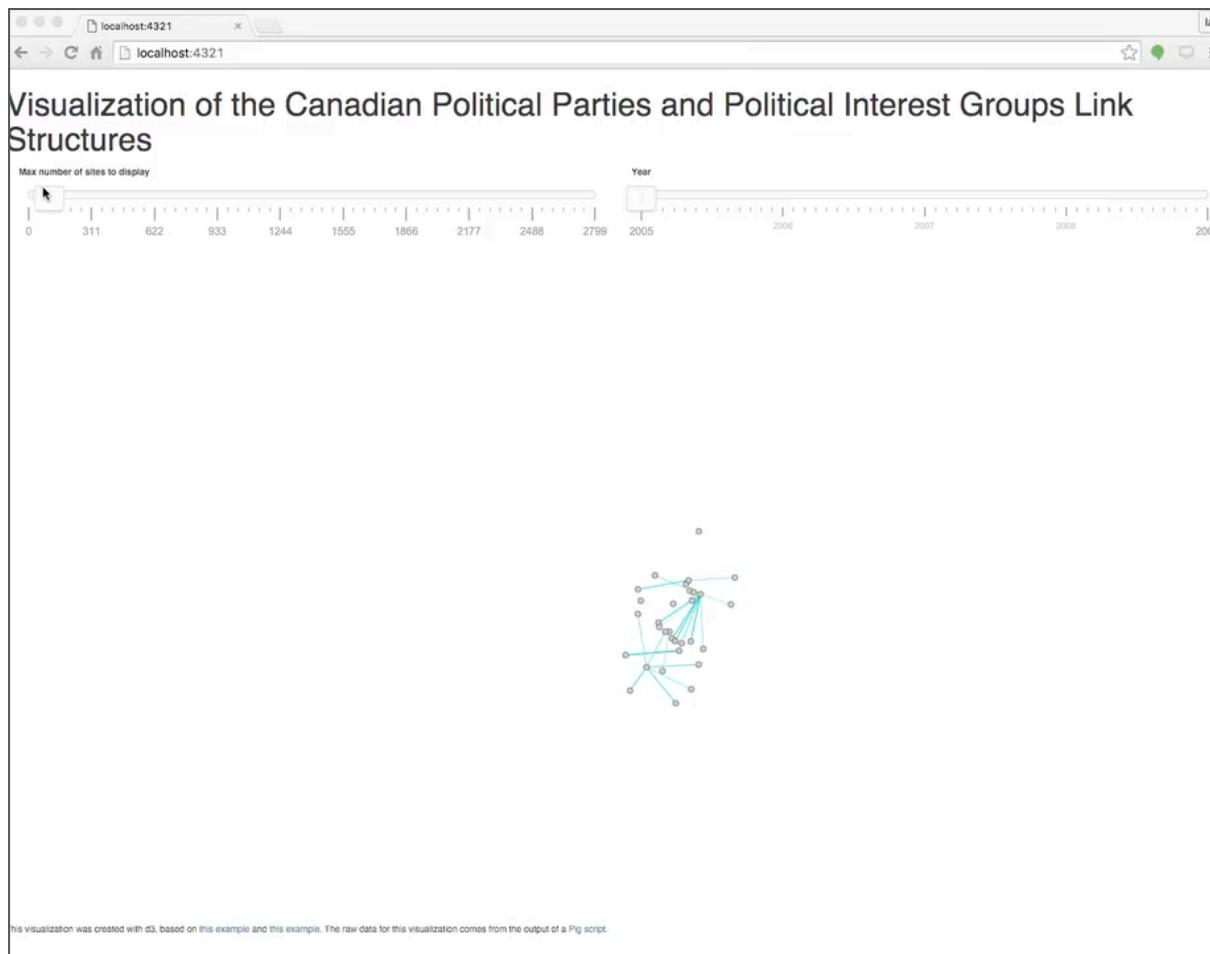
Extract Entities



Extract Links/Gephi Connector



Or D3.js link networks in browser



Bringing it all together in
a notebook environment



Spark Notebook Government Information Day - Demo (unsaved changes) 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]

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.org.arc.gz"
var warc="/Users/ianmilligan1/dropbox/wahr/sample-data/arc-warc/ARCHIVEIT-227-QUARTERLY-XUGECV-20091218231727-00039-crawling06.us.archive.org-8091.warc.gz"
var armdir="/Users/ianmilligan1/dropbox/warcs-workshop";

arc: String = /Users/ianmilligan1/Dropbox/warcs-workshop/227-20051004191331-00000-crawling015.archive.org.arc.gz
warc: String = /Users/ianmilligan1/dropbox/wahr/sample-data/arc-warc/ARCHIVEIT-227-QUARTERLY-XUGECV-20091218231727-00039-crawling06.us.archive.org-8091.warc.gz
armdir: 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](https://github.com/lintool/warcbase/wiki)

Where to learn?



The screenshot shows a web browser window with the title bar "About the Programming His x" and the URL "programminghistorian.org". The page content includes the site's name, navigation links, and a section titled "About the Programming Historian" with a historical illustration and descriptive text.

The Programming Historian

About · Lessons · Contribute · Project Team · Blog

About the Programming Historian



The Programming Historian offers novice-friendly, peer-reviewed tutorials that help humanists learn a wide range of digital tools, techniques, and workflows to facilitate their research.

We regularly publish new lessons, and we always welcome proposals for new lessons on any topic. Our editorial mentors will be happy to work with you throughout the lesson writing process. If you'd like to be a reviewer or if you have suggestions to make *Programming Historian* a more useful resource, please see our [Contribute](#) page.

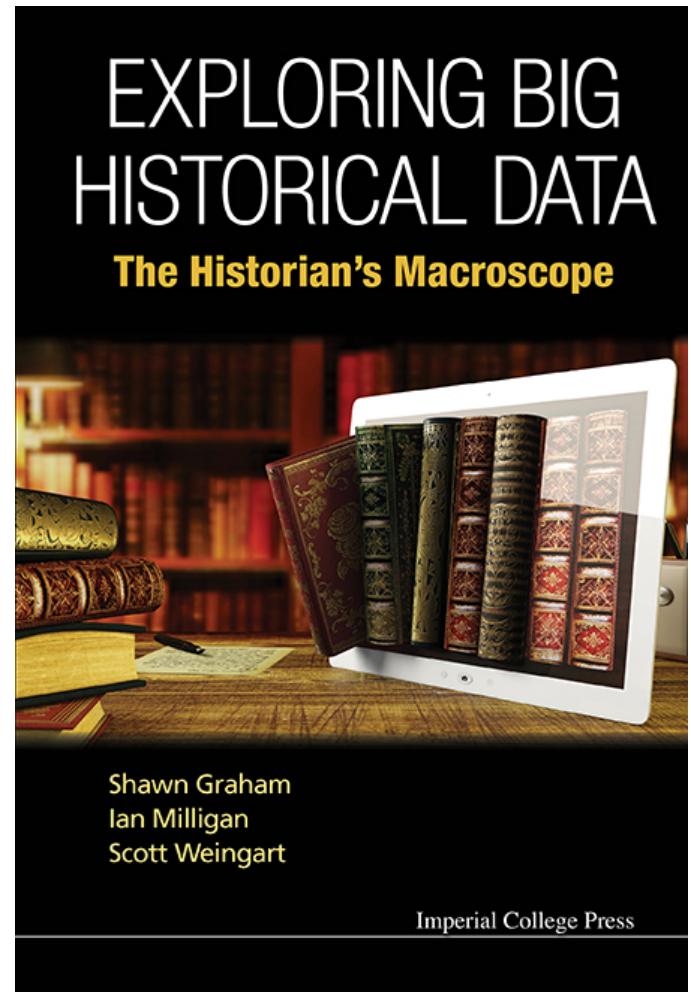
Our [Project Team](#) and peer reviewers work collaboratively with authors to craft tutorials that illustrate fundamental digital and programming principles and

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