

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

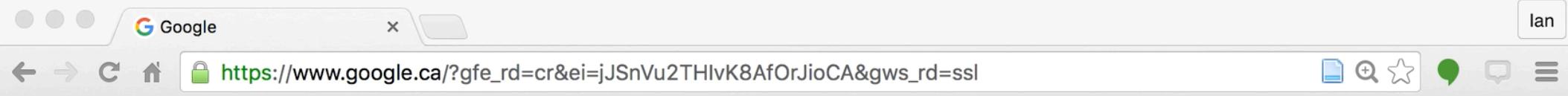
- **1. Prologue:** Big Data is everywhere
- **2. The Web Age:** Will accelerate this process
- **3. What can we do with big data?**

**A Prologue:
Big Data is
Everywhere**

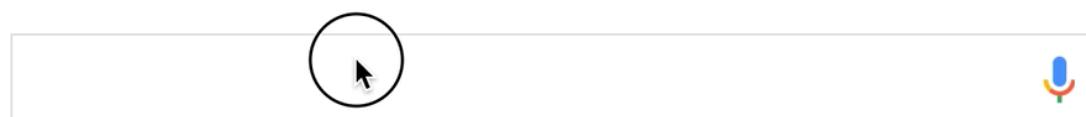
What do we mean by Big Data?

- 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?**



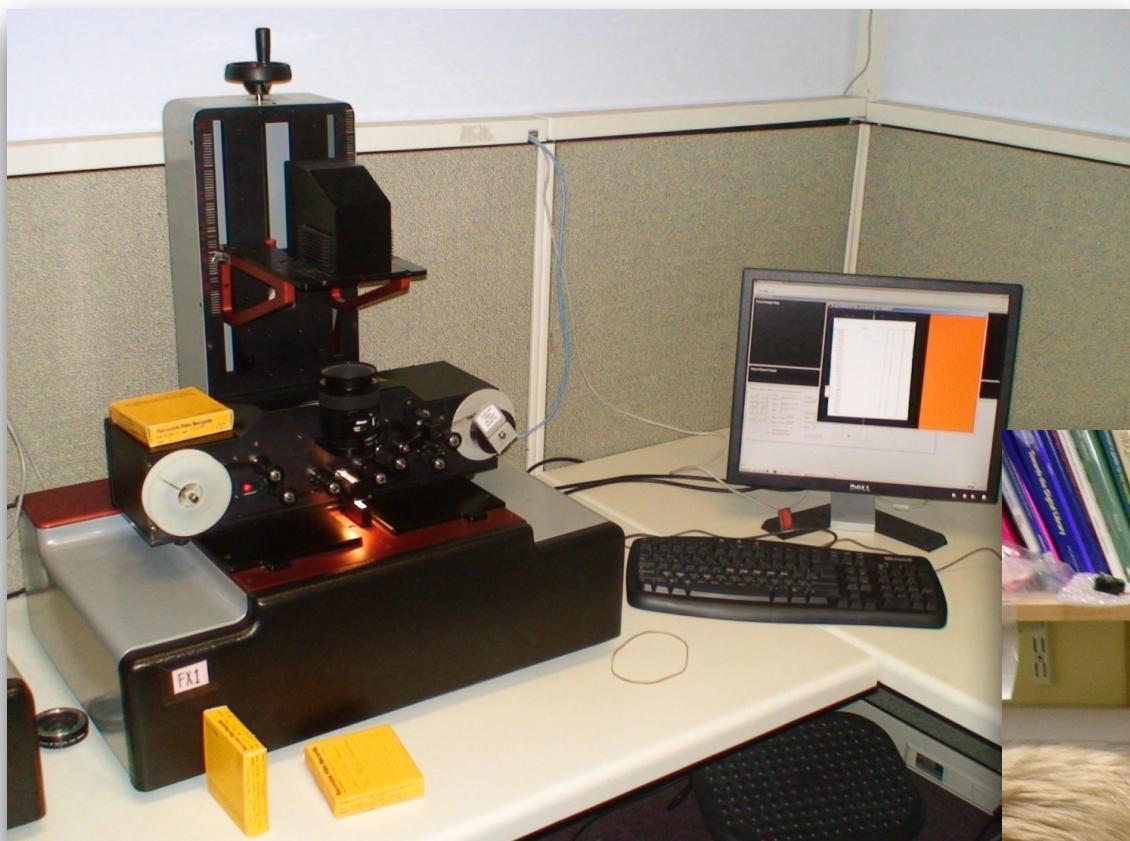
Ian Gmail Images



Google Search

I'm Feeling Lucky

Google.ca offered in: Français



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 | 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

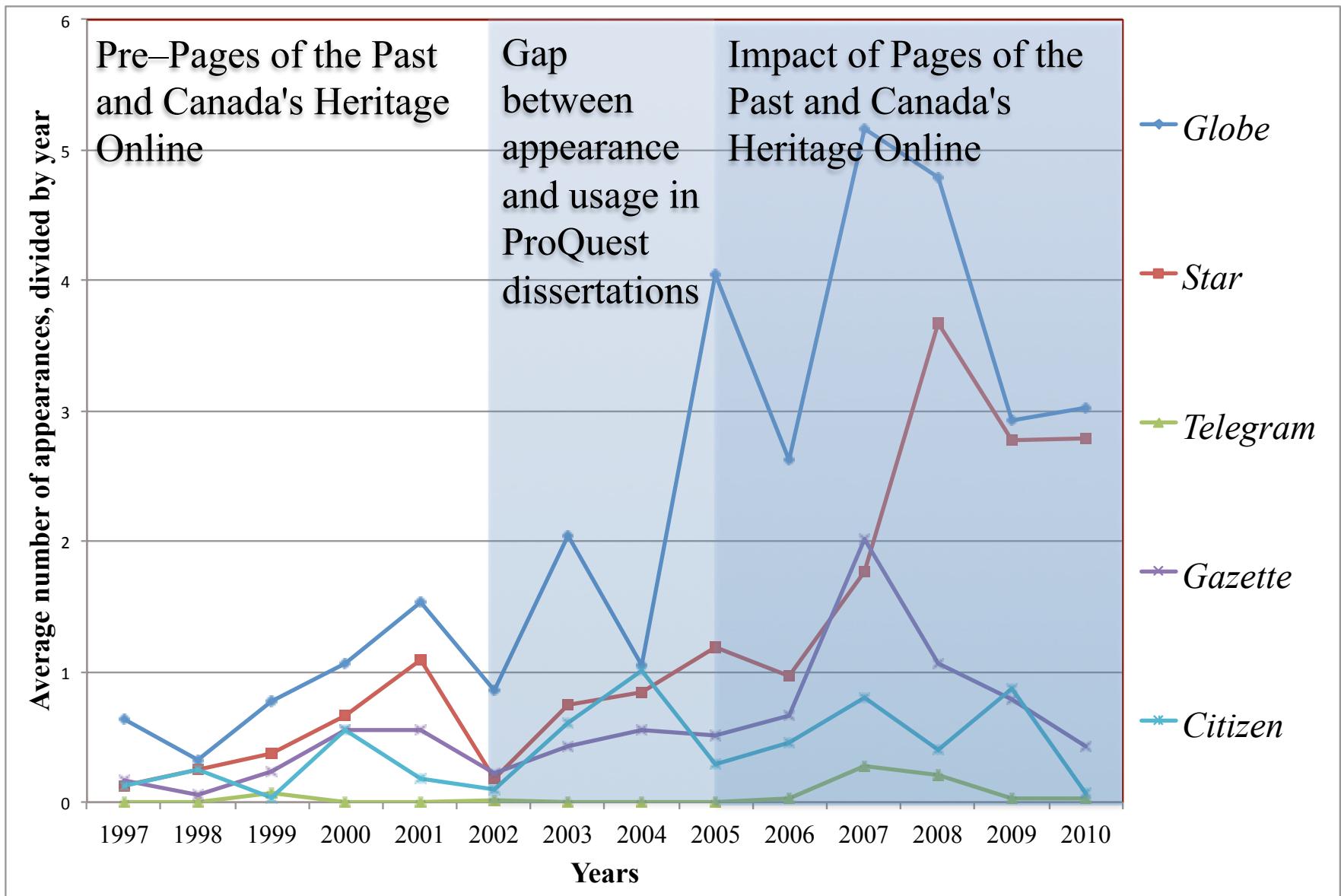
Duplicates: Include duplicate documents *i*

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 ...

**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
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

Danmarks Radio Online

https://web.archive.org/web/19961102165719/http://www.dr.dk/ Ian

INTERNET ARCHIVE Wayback Machine 2,813 captures 2 Nov 96 – 17 Jan 16 http://www.dr.dk/ Go OCT NOV DEC Close X 1995 1996 1997 Help ?

DR ONLINE

Om DR Online - English version - Tekst-version

TV PROGRAM RADIO PROGRAM NYHEDER AFDELINGER EMNER

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](#).

© 1996 DR Henvendelse til webmaster@dr.dk

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.

Scarcity Abundance



RIC'S GRILL
STEAK SEAFOOD
& CHOP HOUSE



October 17, 2015

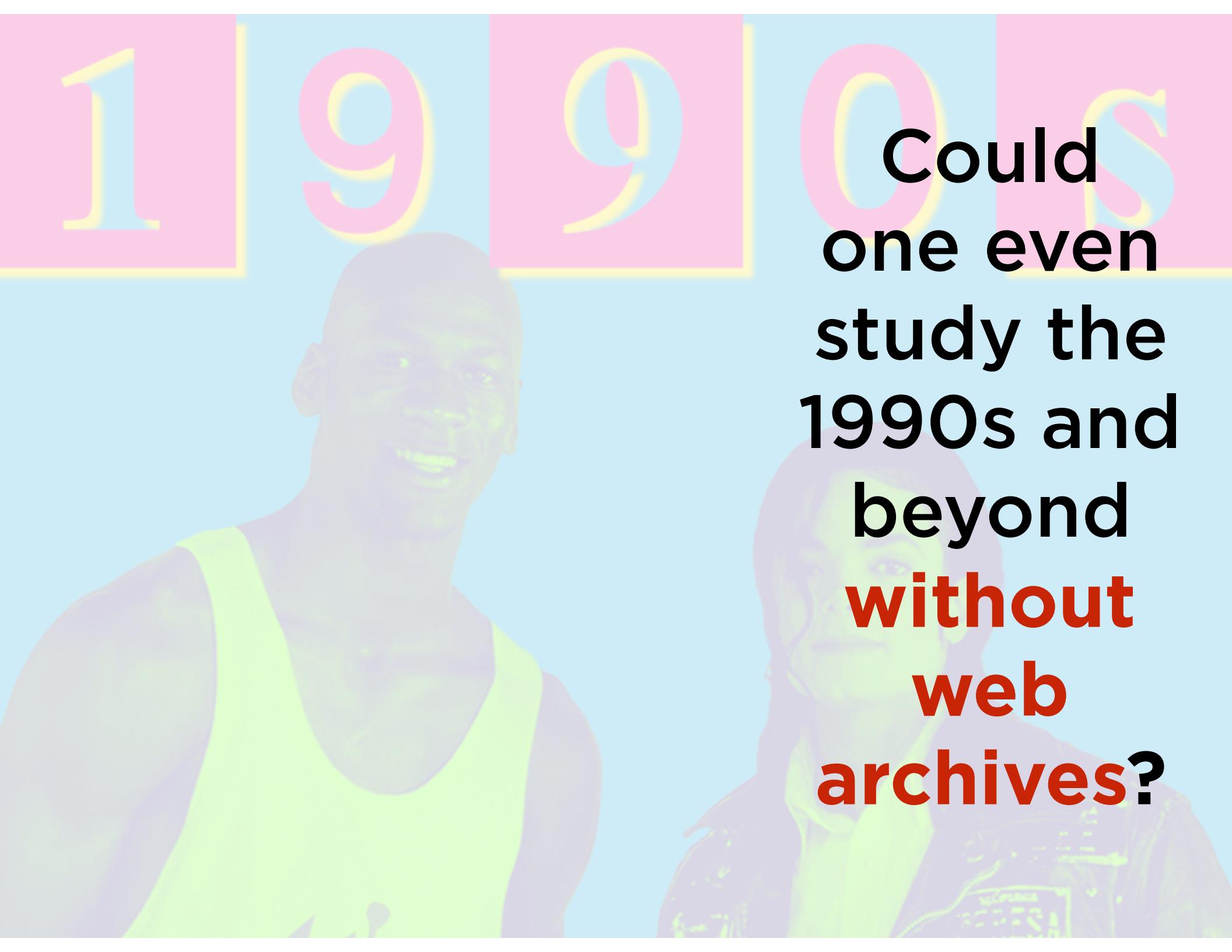
October 18, 2015

October 19, 2015

October 20, 2015

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

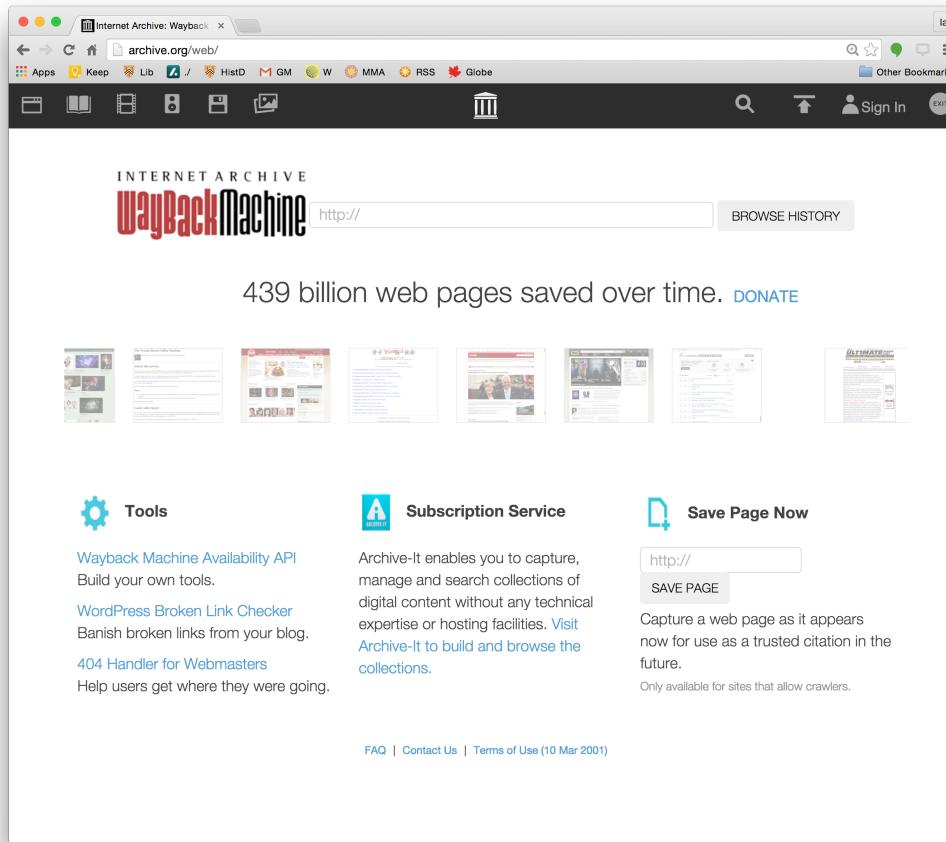
- James Gleick



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

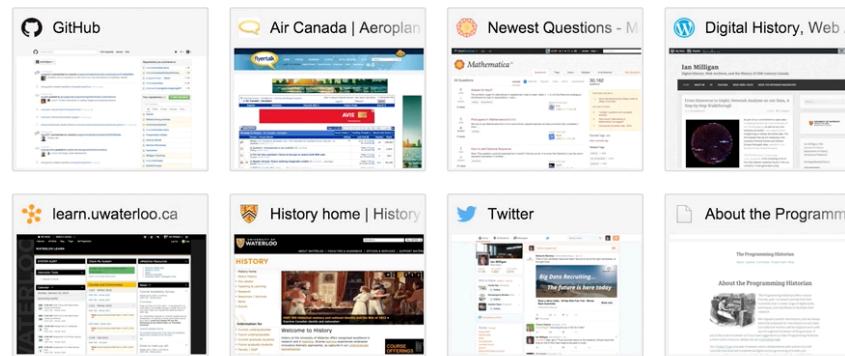
1990s

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...

The screenshot shows a web browser window with the title "Archive-It - Canadian Political Parties and Political Interest Groups". The URL in the address bar is <https://archive-it.org/collections/227>. The page features the Archive-It logo and navigation links for HOME, EXPLORE, LEARN MORE, and CONTACT US. Below this, it displays the title "Canadian Political Parties Groups" collected by "University of Toronto". It notes the collection was archived since Oct, 2005, and describes it as containing national Canadian political parties and a number of specific political interest groups. The subject is listed as "Politics & Elections". A "Narrow Your Results" section allows users to filter by subject, with options like New Democratic Party of Canada (2), Assembly of First Nations (1), Bloc Québécois (1), Canada First (1), and Canada West Foundation (1). A search bar at the bottom right allows users to enter search terms here. The footer includes links for "Sites" and "Search Page Text", and a page number "Page 1 of 1 (54)".

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 collection 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 header includes the Archive-It logo, navigation links for HOME, EXPLORE, LEARN MORE, and CONTACT US, and a tagline: "The leading web archiving service for collecting and accessing cultural heritage on the web. Built at the Internet Archive". Below the header, the breadcrumb navigation shows "Explore > University of Toronto > Canadian Political Parties and Political Interest Groups". The main content area features the University of Toronto Libraries logo and the title "Canadian Political Parties and Political Interest Groups". It indicates the collection was "Collected by: University of Toronto" and "Archived since: Oct, 2005". A detailed description states: "Canadian Political Parties and Political Interest Groups will archive the websites of all the national Canadian political parties, and a number of special interest groups across the political spectrum. Subject: Politics & Elections. Collector: University of Toronto". Below this, a search bar allows users to enter a search term ("Stephen Harper") and a "Search" button. A message below the search bar says, "Enter a search term on the right to search the text within the archived pages. Or for more search options, use the Advanced Search options below." The search results page shows a summary of the search results: "The following results were found for the term(s): Stephen Harper • No metadata results for Stephen Harper, but there are up to 1213132 matches within the page text." The results table has a header row with columns for "Page Number", "Title", "URL", and "Captured Date". The first result listed is "Stephen Harper | Facebook". The interface also includes filters for "Contains all of:", "Exact phrase:", "Not containing:", "From the Host:", "Results per host:", and "File format:".

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

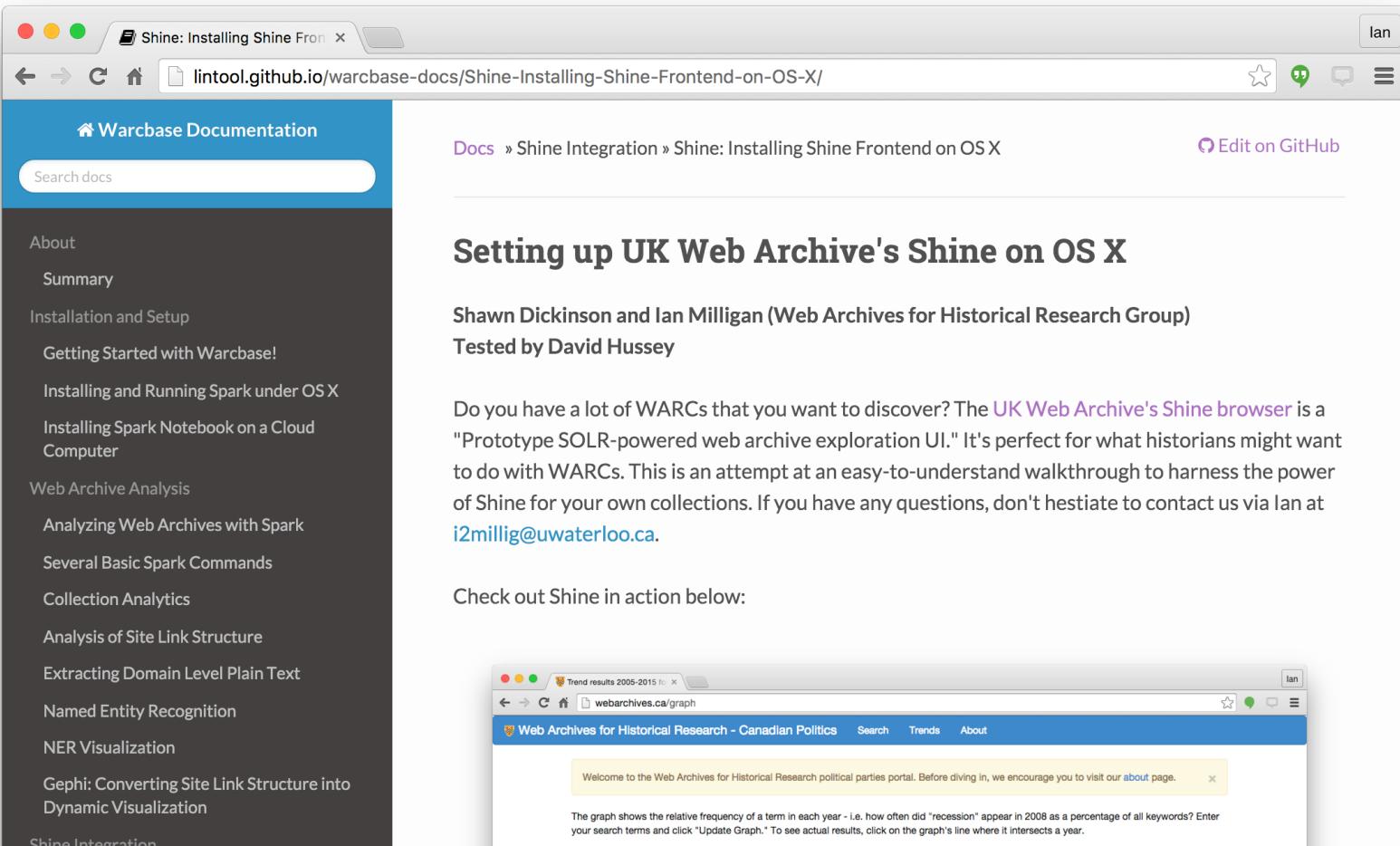
Code Issues Pull requests Wiki Pulse Graphs

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



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

- Title Bar:** Shine: Installing Shine Front
- Address Bar:** lintool.github.io/warcbase-docs/Shine-Installing-Shine-Frontend-on-OS-X/
- Page Content:**
 - Header:** Warcbase Documentation
 - Breadcrumbs:** Docs » Shine Integration » Shine: Installing Shine Frontend on OS X
 - Buttons:** Edit on GitHub
 - Section:** **Setting up UK Web Archive's Shine on OS X**
 - Text:** Shawn Dickinson and Ian Milligan (Web Archives for Historical Research Group)
Tested by David Hussey
 - Text:** Do you have a lot of WARCs that you want to discover? The [UK Web Archive's Shine browser](#) is a "Prototype SOLR-powered web archive exploration UI." It's perfect for what historians might want to do with WARCs. This is an attempt at an easy-to-understand walkthrough to harness the power of Shine for your own collections. If you have any questions, don't hesitate to contact us via Ian at i2millig@uwaterloo.ca.
 - Text:** Check out Shine in action below:
 - Image:** A smaller screenshot of a web browser showing the Shine interface with a graph visualization.

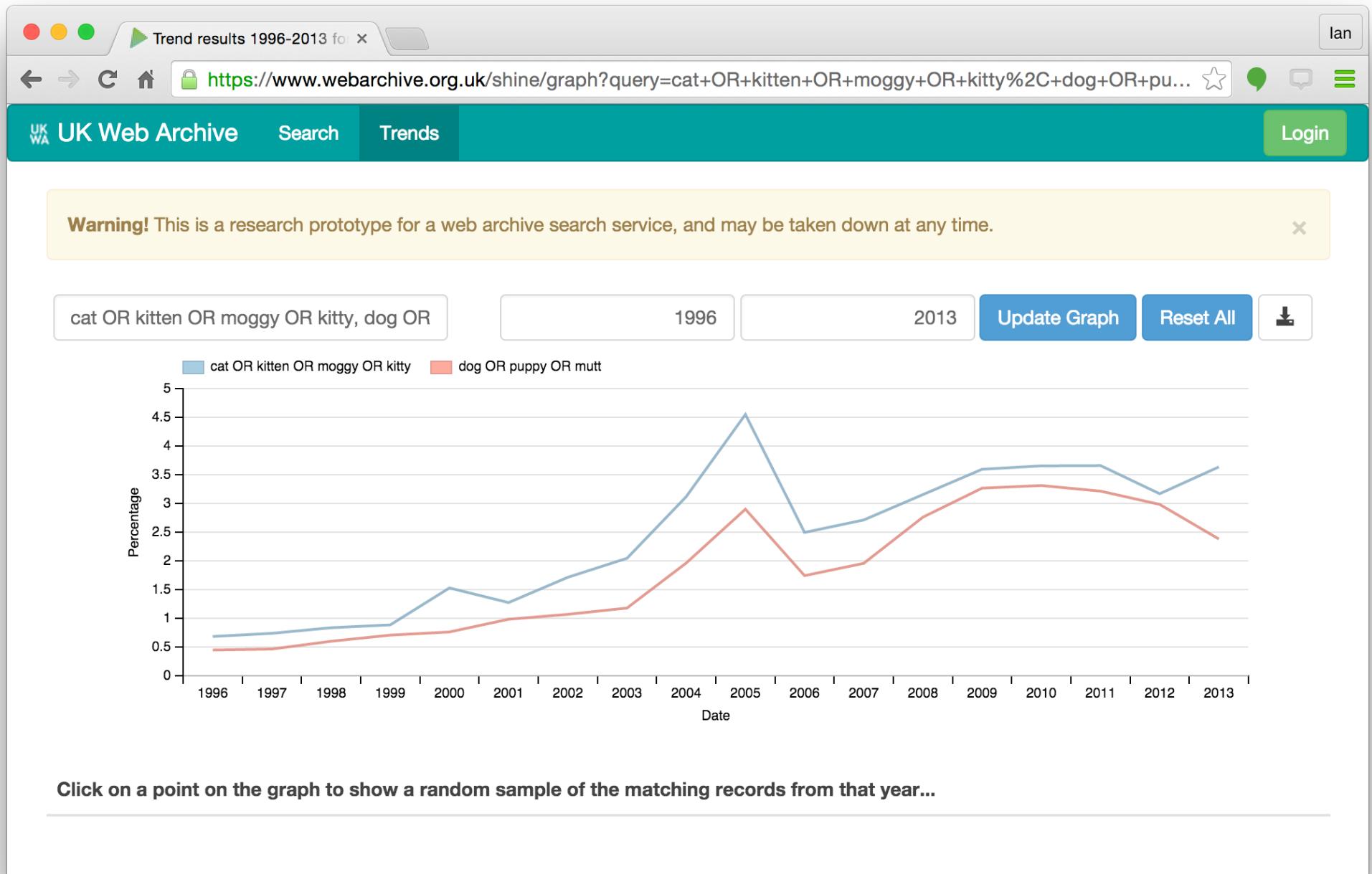
**Great research question that our
contemporary historians were
studying (Canada changing)**

+

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

+

Ope Source Software





**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

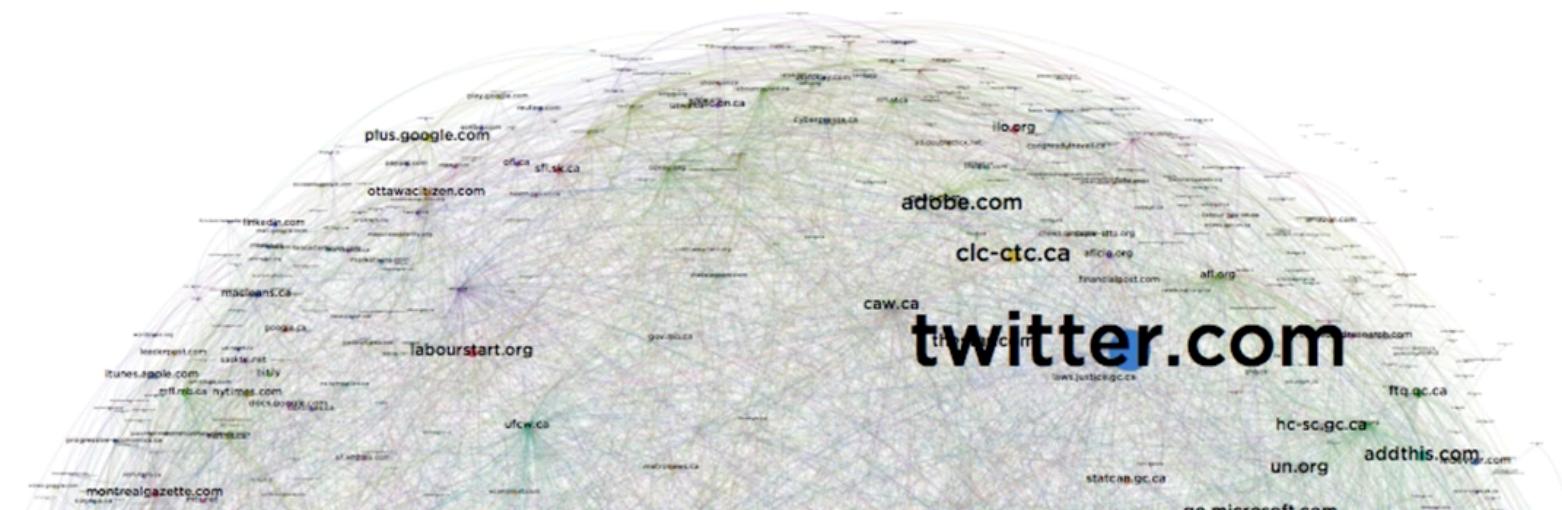
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

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....

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)
- Results:** Results (selected), Concordance
- Page Footer:** Results 1 to 10 of 1,161,300, CSV ▾, Asc ▾

**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!

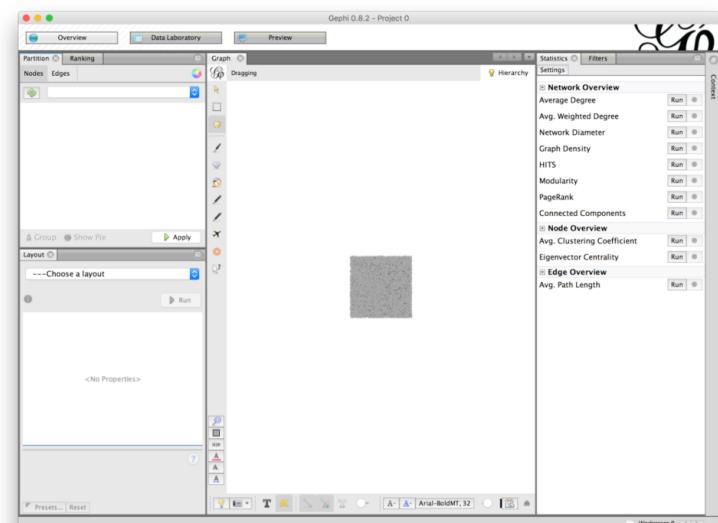
below.

Step-by-Step Walkthrough

Once you've downloaded the file, open up Gephi.

On the opening screen, you want to select “Open a Graph File...” and select the `all-links-cpp-link.graphml` file that you downloaded from our Dataverse page.

You then want to click ‘ok’ on the next page. Create a ‘new graph.’



Do you want to make this link graph yourself from our data? Read on.

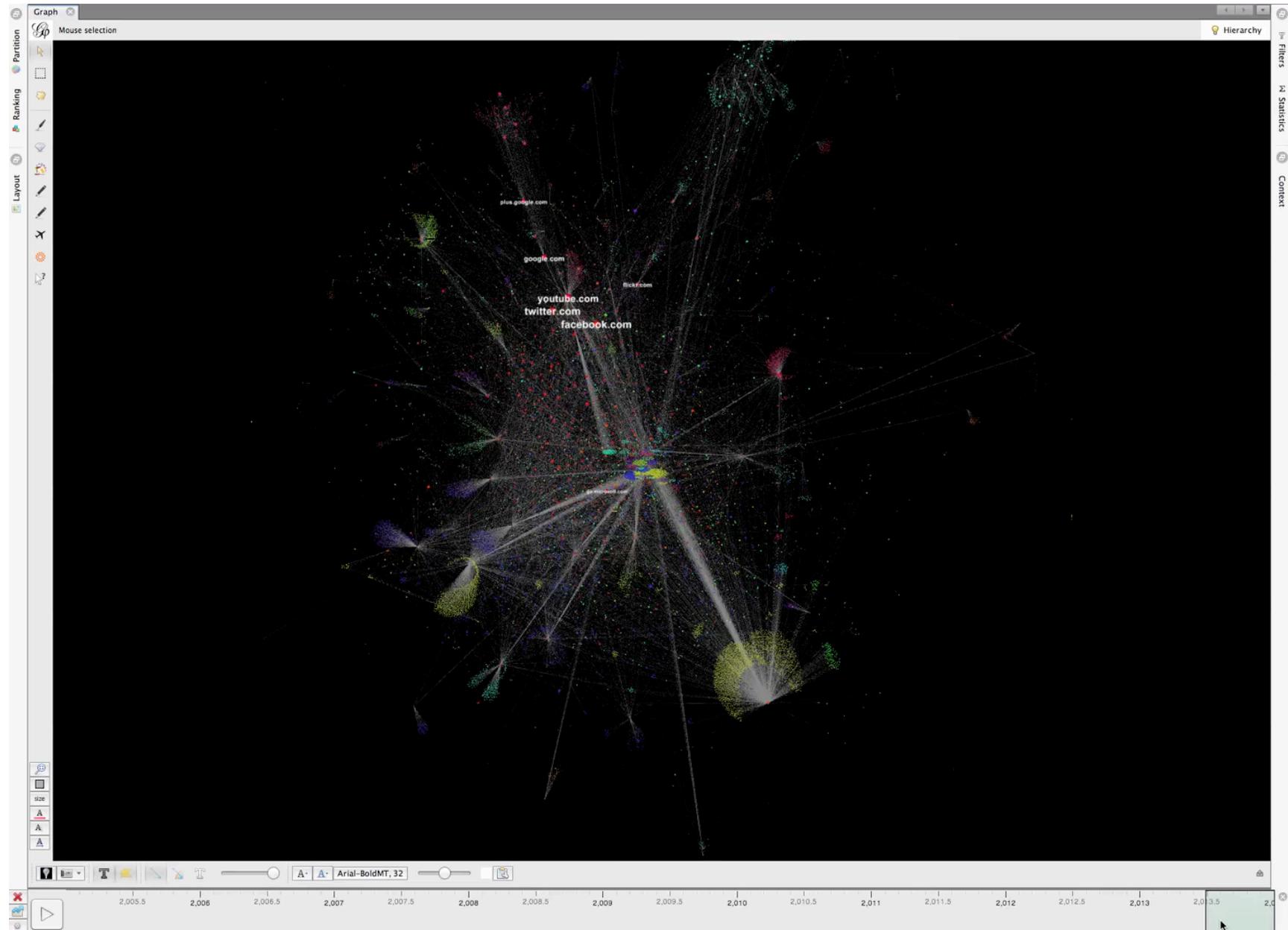
You should now see what I (nerdily) call a borg cube. That's good, because it means that the data is in there. We need to make it usable, however.

Click on the “Data Laboratory” tab at the top.

Click on “Nodes” above. When it is shaded behind it, that means that it is selected.

Click on “Copy Data to another Column,” select ID, and then select “label” on the drop

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

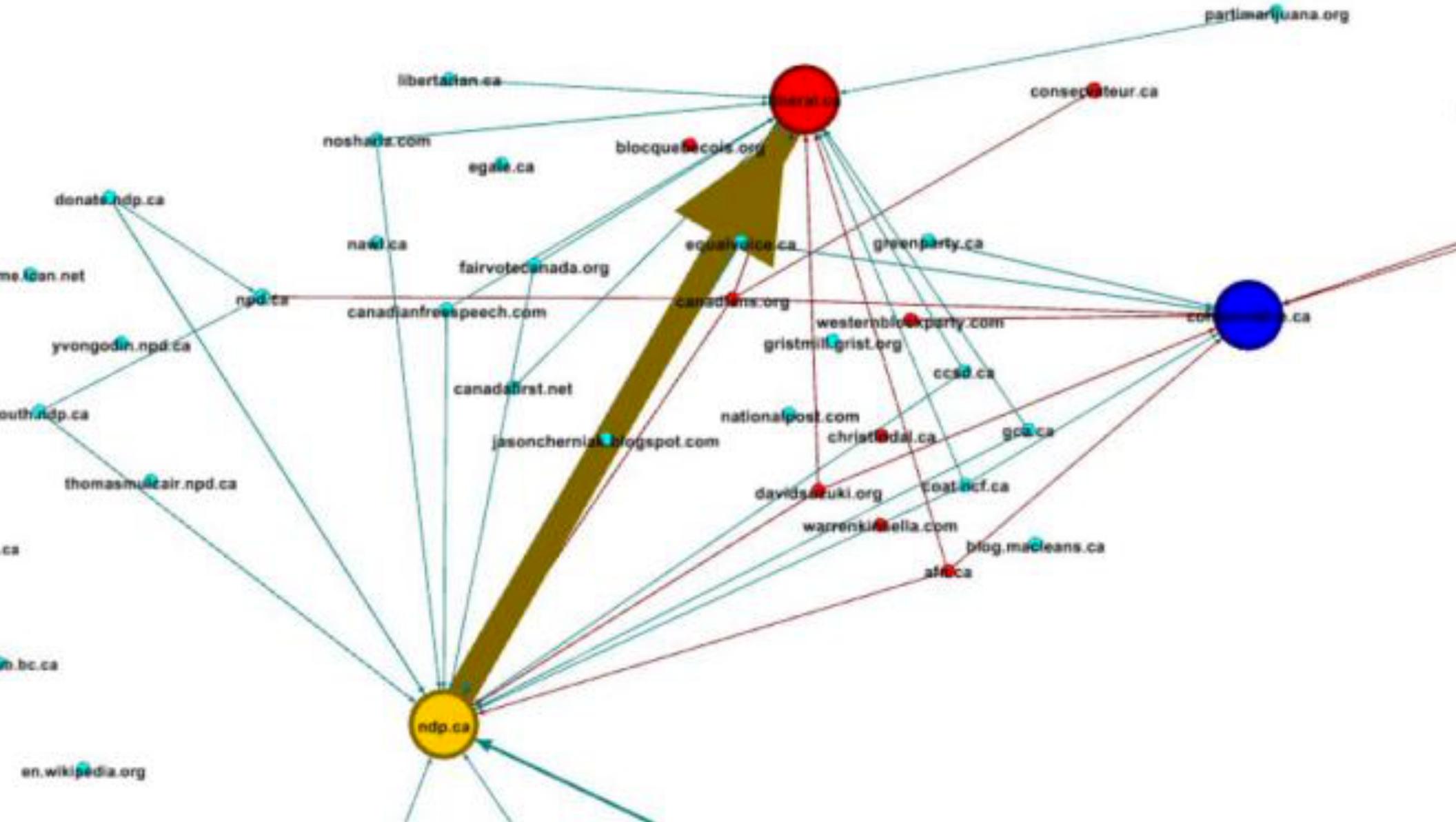


December 2008

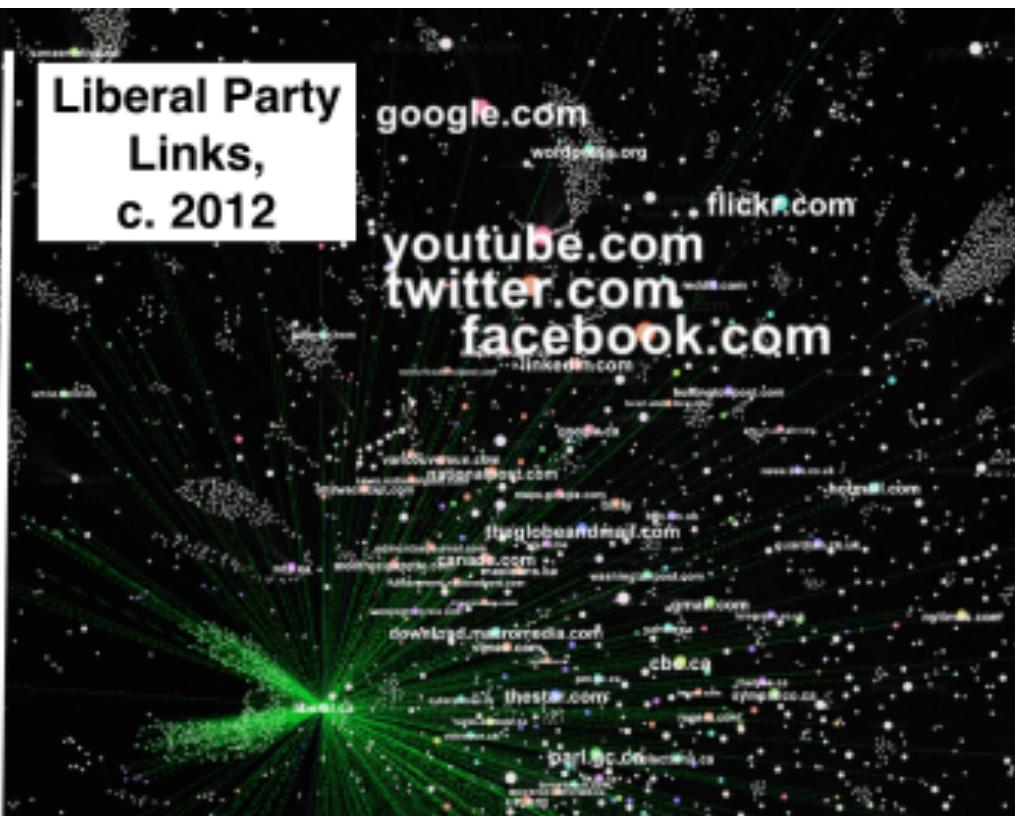
Election campaign Ends; Attacking Harper on Anti-American Grounds (bushharper)



2005 Canadian Federal Election

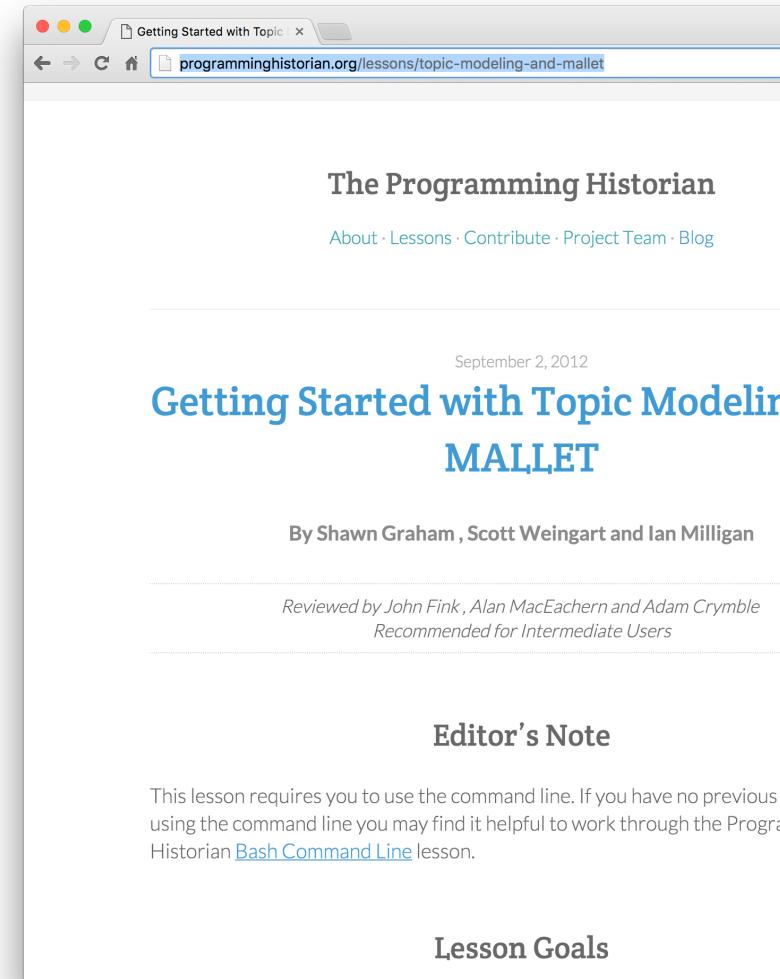


Metadata Extraction



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

Walkthrough:
[http://
programminghistorian.org/
lessons/topic-modeling-
and-mallet](http://programminghistorian.org/lessons/topic-modeling-and-mallet)



The screenshot shows a web browser window titled "Getting Started with Topic Modeling and Mallet". The URL in the address bar is "programminghistorian.org/lessons/topic-modeling-and-mallet". The page content includes the title "The Programming Historian", navigation links "About · Lessons · Contribute · Project Team · Blog", a date "September 2, 2012", the main heading "Getting Started with Topic Modeling and Mallet", author information "By Shawn Graham, Scott Weingart and Ian Milligan", and a review section with "Reviewed by John Fink, Alan MacEachern and Adam Crymble" and "Recommended for Intermediate Users". Below the page content, there is an "Editor's Note" section stating: "This lesson requires you to use the command line. If you have no previous experience using the command line you may find it helpful to work through the Programming Historian [Bash Command Line](#) lesson." At the bottom right, there is a "Lesson Goals" section.

The Programming Historian

About · Lessons · Contribute · Project Team · Blog

September 2, 2012

Getting Started with Topic Modeling and Mallet

By Shawn Graham, Scott Weingart and Ian Milligan

Reviewed by John Fink, Alan MacEachern and Adam Crymble
Recommended for Intermediate Users

Editor's Note

This lesson requires you to use the command line. If you have no previous experience using the command line you may find it helpful to work through the Programming Historian [Bash Command Line](#) lesson.

Lesson Goals

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

colinbarriemp.ca	12
colinbarriemp.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...

The screenshot shows a GitHub wiki page for the repository 'lintool/warcbase'. The title bar indicates the page is at 'Home · lintool/warcbase · GitHub, Inc. [US]'. The main content area is titled 'Home' and shows the following text:

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

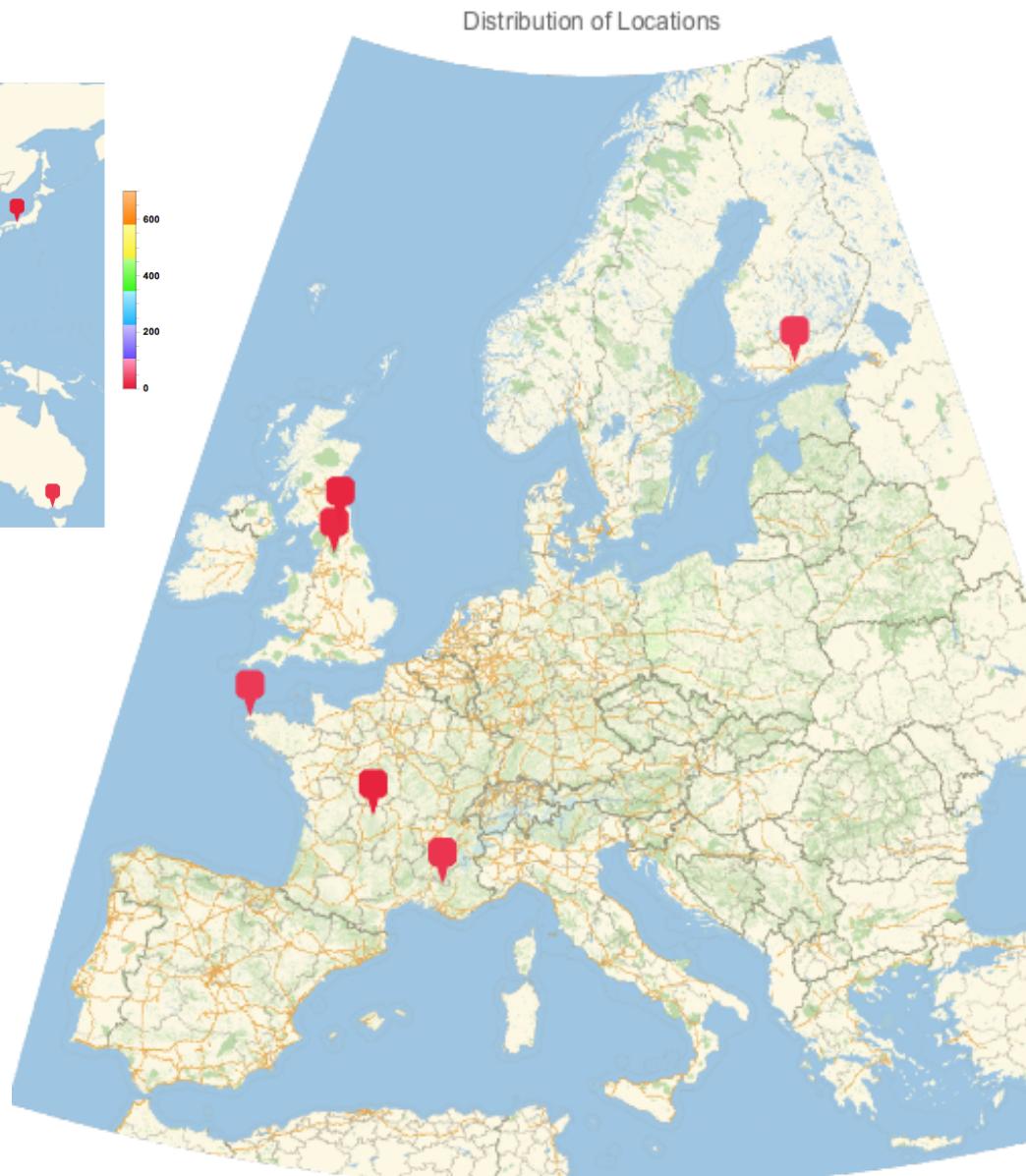
Extract all Plain Text

```
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



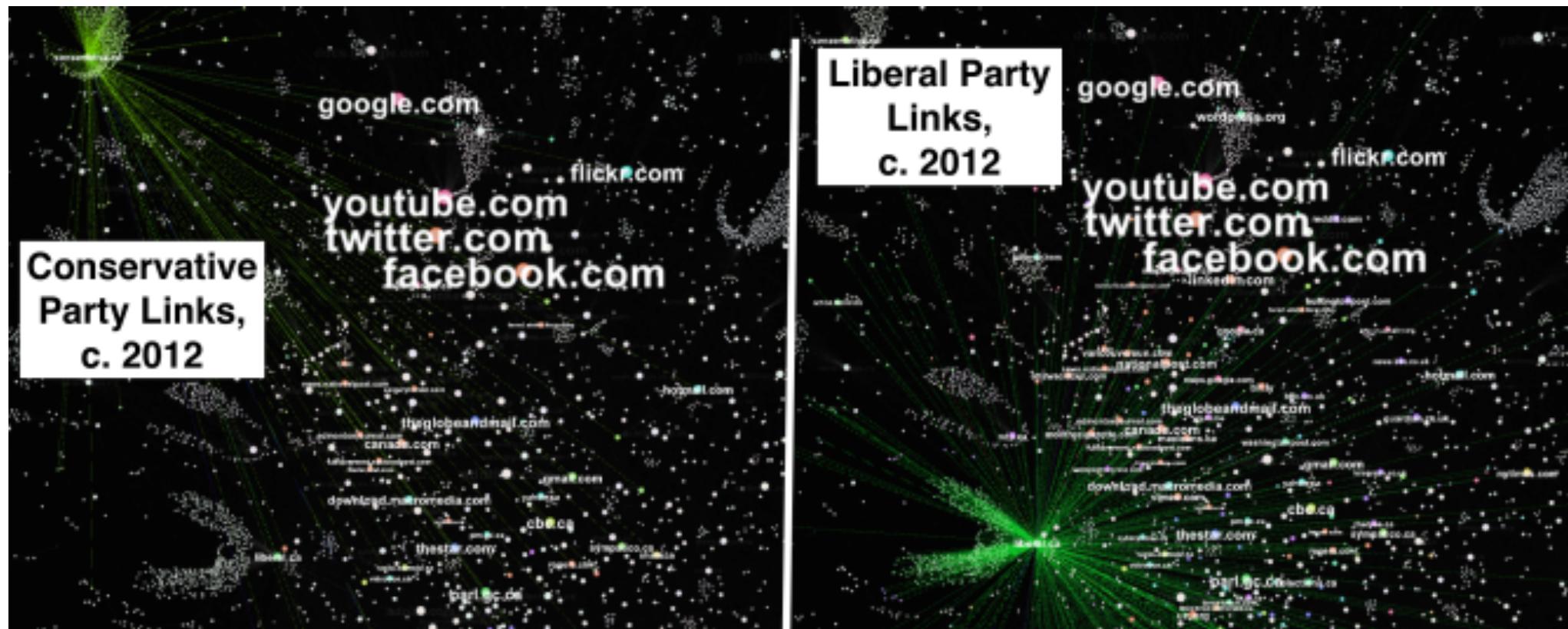
```
In[95]:= int = SemanticInterpretation[#] & /@ processedfreq[[All, 1]]  
Out[95]= {Canada, Calgary, Colombia, Montreal, $Failed, Ontario, Canada, Afghanistan,  
Ottawa, Manitoba, Canada, British Columbia, Canada, Toronto, Nova Scotia, Canada,  
Saskatchewan, Canada, Quebec, Canada, Alberta, Canada,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



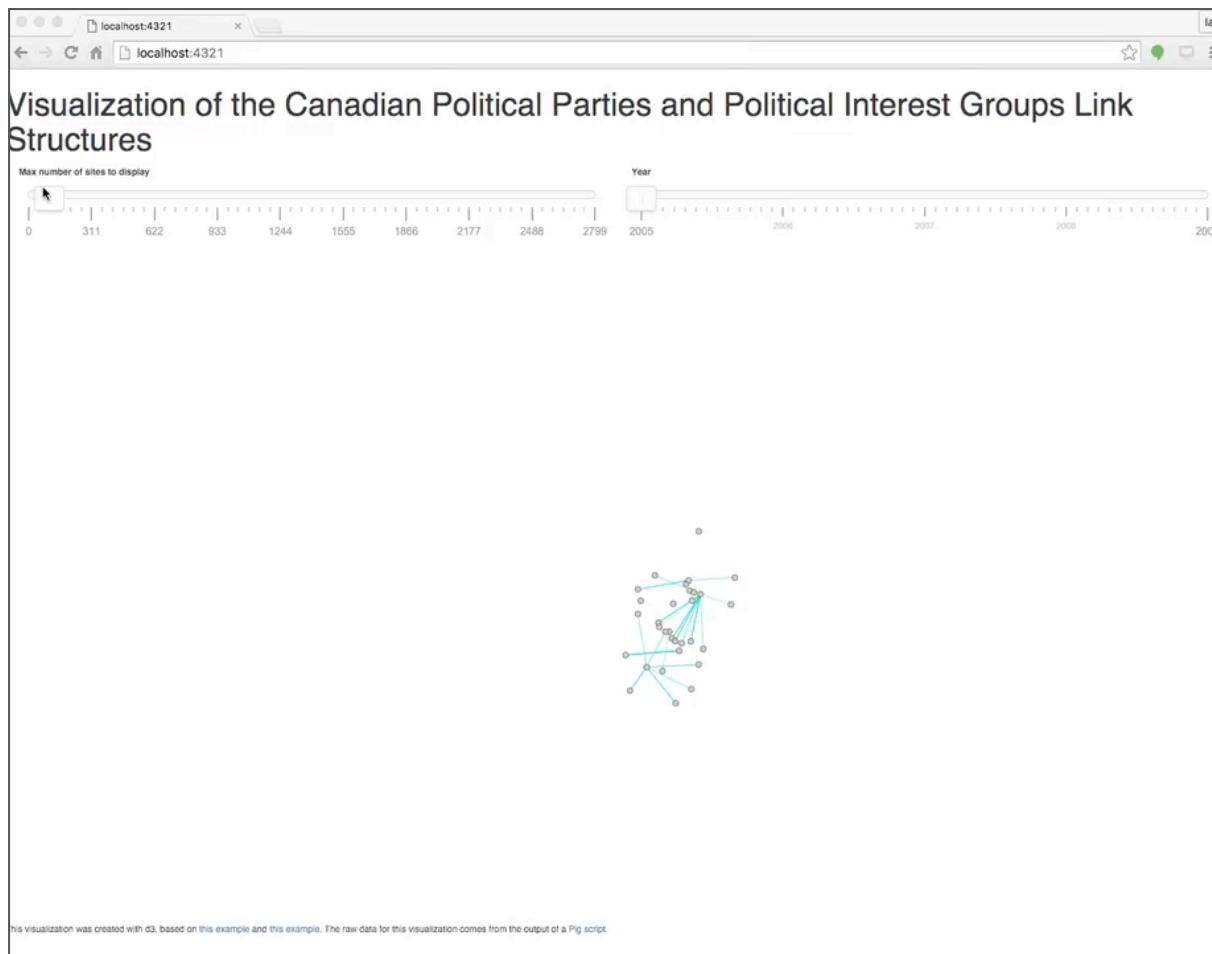
Extract Entities



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

[..](#)[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?



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 about the project.

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.

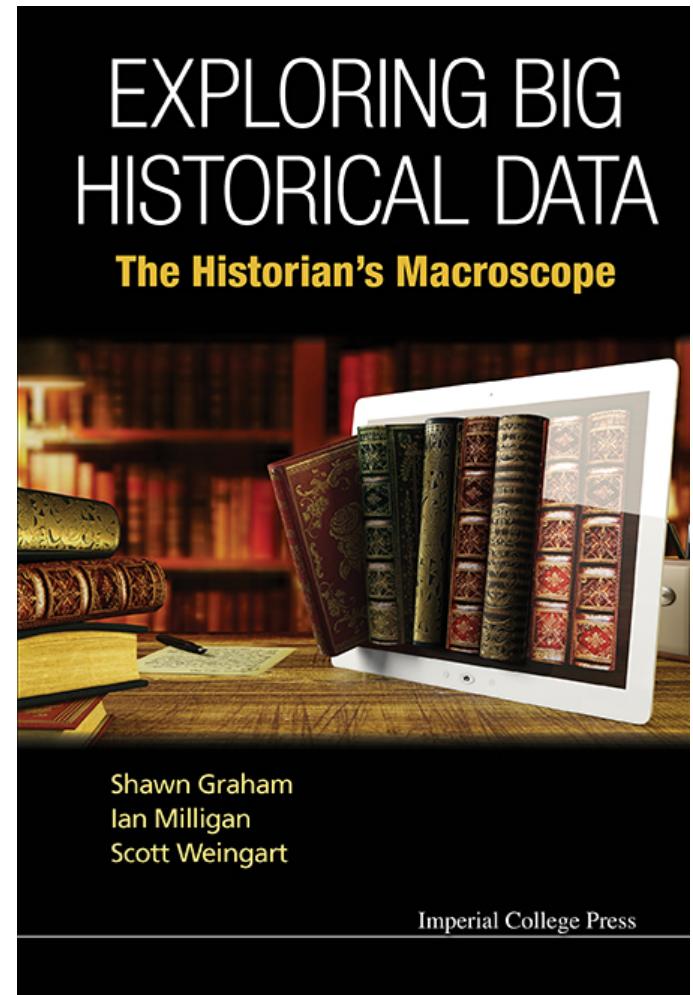
Our [Project Team](#) and peer reviewers work collaboratively with authors to craft tutorials that illustrate fundamental digital and programming principles and

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
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