

# **Big Data and History:**

## **Seeing the Past**

## **through a Macroscope**

---

**Ian Milligan, PhD**  
Assistant Professor  
**@ianmilligan1**



**UNIVERSITY OF WATERLOO**  
**FACULTY OF ARTS**  
Department of History

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

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

# Today's Talk

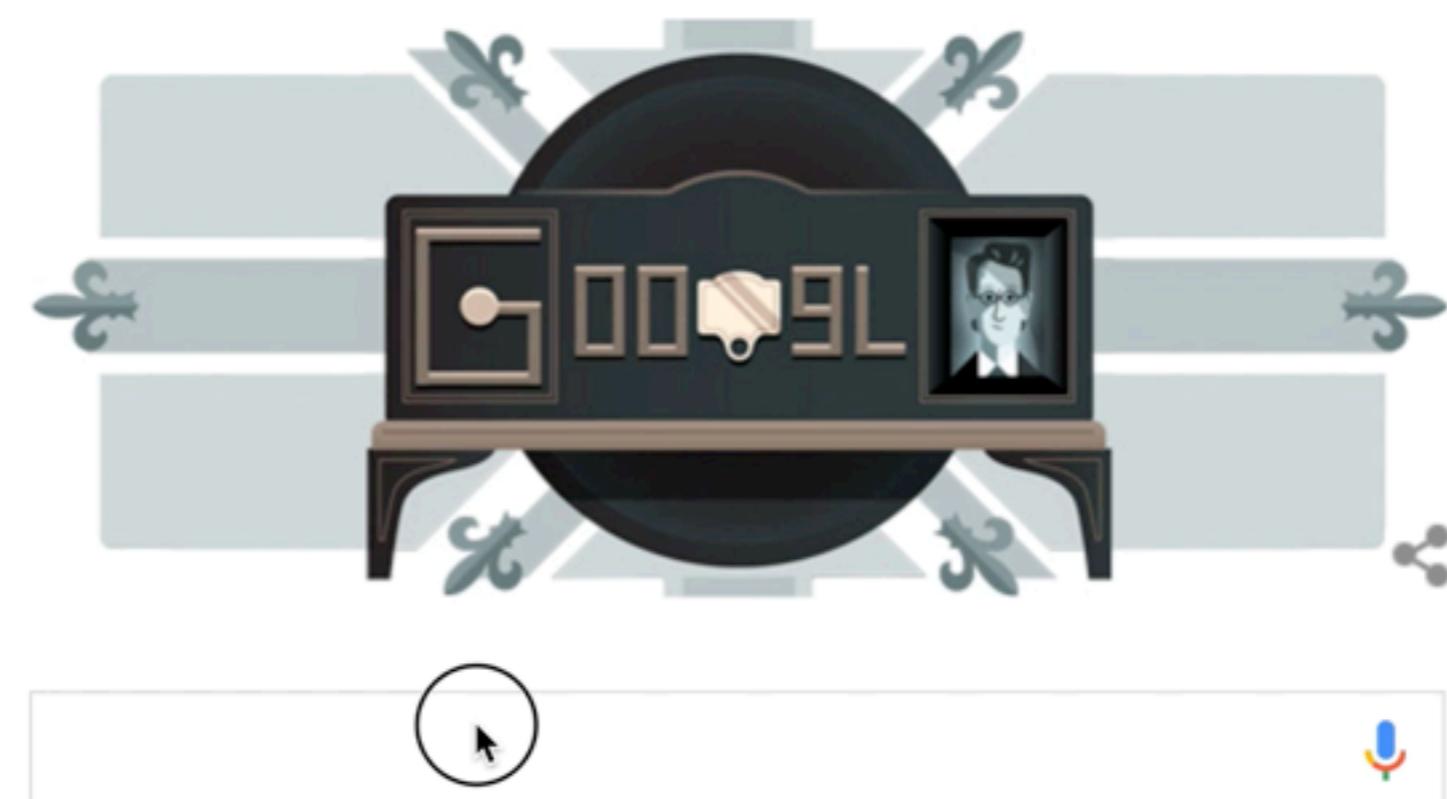
- **1. Prologue:** Big Data is everywhere
- **2. The Web Age:** Will accelerate this process
- **3. What can we do with big data?**

**A Prologue:  
Big Data is  
Everywhere**

# What do we mean by Big Data?

- Computational definition: the 5 Vs (Volume, Velocity, Variety, Veracity, and Value)
- “For us, as humanists, big is in the eye of the beholder. If it’s more data that you could conceivably read yourself in a reasonable amount of time, or that requires computational intervention to make new sense of it, it’s big enough!” (Shawn Graham, Ian Milligan, Scott Weingart, *Exploring Big Historical Data*)

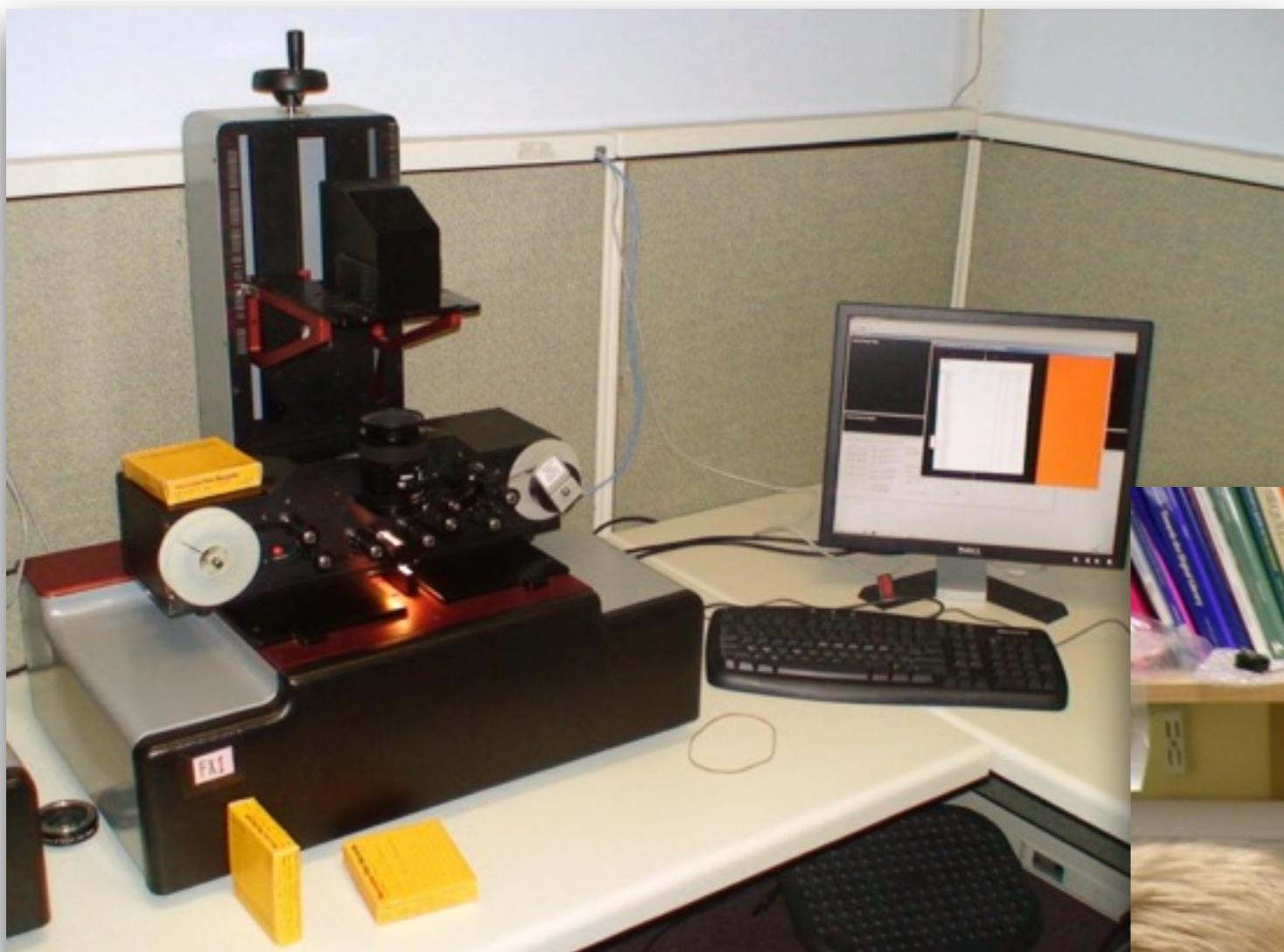
**Why is it  
everywhere?**



Google Search

I'm Feeling Lucky

Google.ca offered in: [Français](#)



Advanced Search - ProQuest X

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

0 Recent searches | 0 Selected items | My Research | Exit

< All databases | News & Newspapers databases

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

Publication date: All dates

Sort results by: Publication date (most recent first)

Items per page: 50

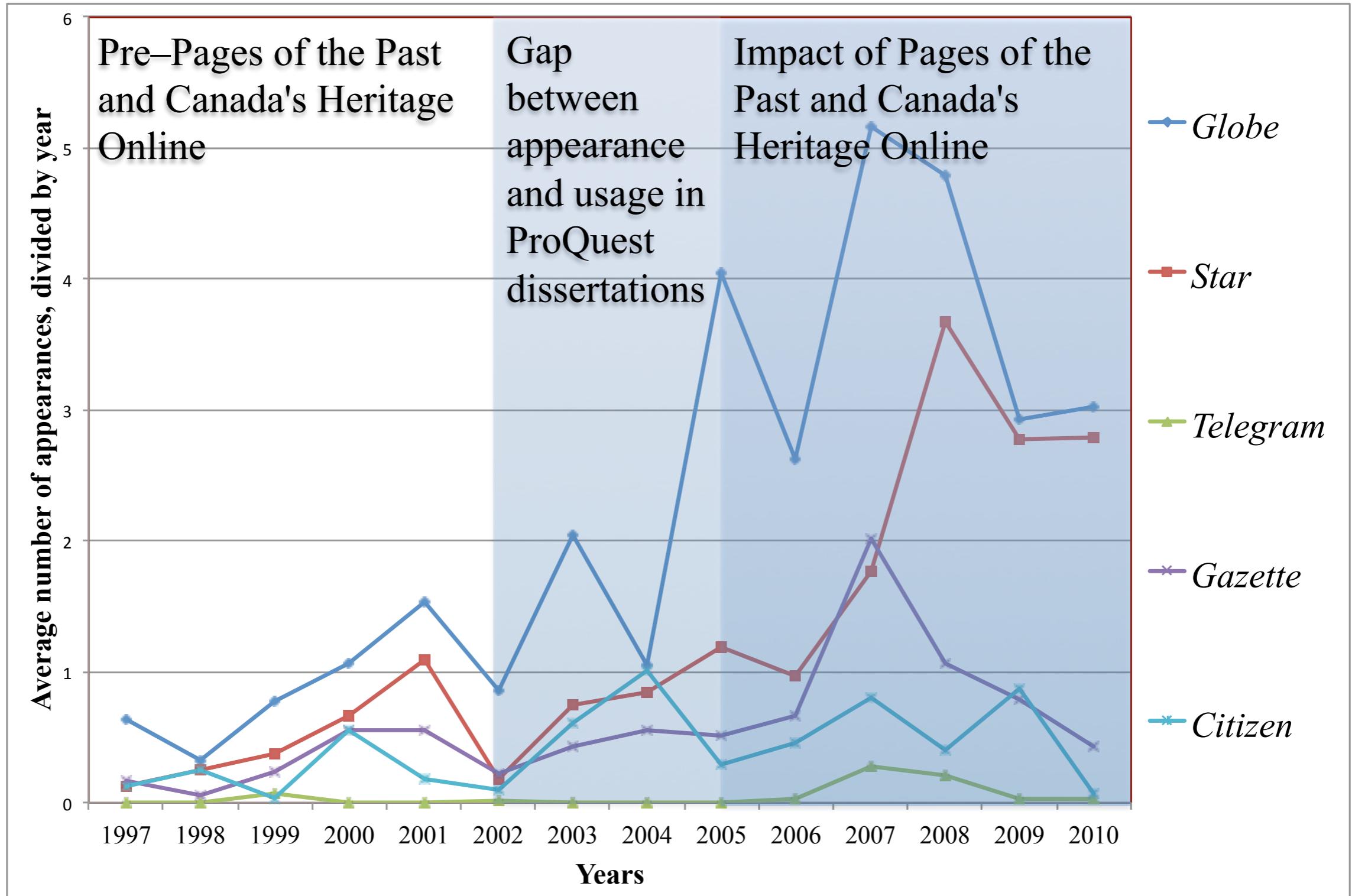
Duplicates:  Include duplicate documents

Search Clear form

Search subject areas

Use search forms customized for each subject.

	The Arts
	Business
	Dissertations & Theses
	Health & Medicine
	History
	Literature & Language
	News & Newspapers



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

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

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

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

A process that is only  
now beginning to  
accelerate.

**First - more data than  
ever before being  
preserved;**

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

CONGRESSIONAL ARCHIVES

THOMAS P. O'NEILL PAPERS

Series I

CONGRESSIONAL ARCHIVES

THOMAS P. O'NEILL PAPERS

Series I

PARTY LEADERSHIP / ADMINISTRATIVE FILES

Democratic Party  
BOX 29

Subseries F Democratic Party  
BOX 29

JOHN J. BURNS LIBRARY  
BOSTON COLLEGE

374

JOHN J. BURNS LIBRARY  
BOSTON COLLEGE

375

CONGRESSIONAL ARCHIVES

THOMAS P. O'NEILL PAPERS

Series I  
PARTY LEADERSHIP / ADMINISTRATIVE FILES

F Democratic Party  
BOX 28

JOHN J. BURNS LIBRARY  
BOSTON COLLEGE

373

CONGRESSIONAL ARCHIVES

THOMAS P. O'NEILL PAPERS

Series I  
PARTY LEADERSHIP / ADMINISTRATIVE FILES

Democratic Party  
BOX 27

JOHN J. BURNS LIBRARY  
BOSTON COLLEGE

372

CONGRESSIONAL ARCHIVES

THOMAS P. O'NEILL PAPERS

Series I  
PARTY LEADERSHIP / ADMINISTRATIVE FILES

Democratic Party  
BOX 26

JOHN J. BURNS LIBRARY  
BOSTON COLLEGE

371

CONGRESSIONAL ARCHIVES

THOMAS P. O'NEILL PAPERS

Series I  
PARTY LEADERSHIP / ADMINISTRATIVE FILES

Democratic Party  
BOX 25

JOHN J. BURNS LIBRARY  
BOSTON COLLEGE

370

# Scarcity

CONGRESSIONAL ARCHIVES  
THOMAS P. O'NEILL PAPERS

370

CONGRESSIONAL ARCHIVES

THOMAS P. O'NEILL PAPERS

Subseries F Democratic Party  
BOX 29

371

CONGRESSIONAL ARCHIVES

THOMAS P. O'NEILL PAPERS

Subseries F Democratic Party  
BOX 29

372

CONGRESSIONAL ARCHIVES

THOMAS P. O'NEILL PAPERS

Series I  
PARTY LEADERSHIP / ADMINISTRATIVE FILES  
Subseries F Democratic Party  
BOX 29

373

JOHN J. BURNS LIBRARY  
BOSTON COLLEGE

CONGRESSIONAL ARCHIVES

THOMAS P. O'NEILL PAPERS

Series I  
PARTY LEADERSHIP / ADMINISTRATIVE FILES  
Subseries F Democratic Party  
BOX 29

374

JOHN J. BURNS LIBRARY  
BOSTON COLLEGE

CONGRESSIONAL ARCHIVES

THOMAS P. O'NEILL PAPERS

Series I  
PARTY LEADERSHIP / ADMINISTRATIVE FILES  
Subseries F Democratic Party  
BOX 30

JOHN J. BURNS LIBRARY  
BOSTON COLLEGE

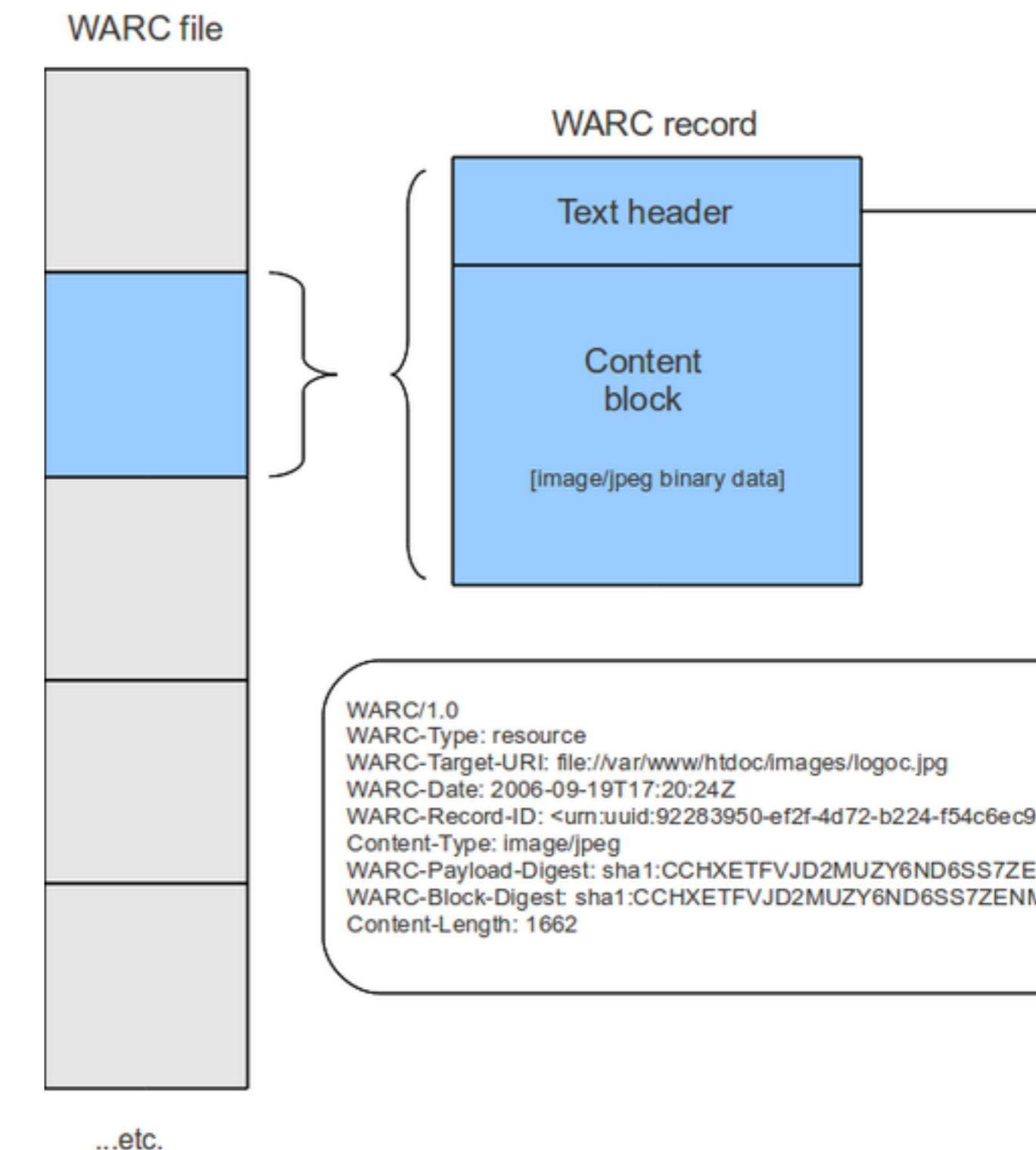
375



**WebARChive (WARC) File**

# ?????

- What does this really mean?



Danmarks Radio Online

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

INTERNET ARCHIVE  
WayBackMachine

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

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](#).

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](#).

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 NYC GA About Donate [Twitter](#) [YouTube](#) [Facebook](#) [RSS](#)

## 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 livelihoods 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 movement: the liberation of vacant bank-owned homes for those in need. The bank-owned

**General Inquiries:**  
[general@occupywallst.org](mailto:general@occupywallst.org)  
**Press Inquiries:**  
[press@occupywallst.org](mailto: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 NYC GA 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 NYC GA About Donate [Twitter](#) [YouTube](#) [Facebook](#) [RSS](#)

## 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 livelihoods 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 movement: the liberation of vacant bank-owned homes for those in need. The bank-owned

**General Inquiries:**  
[general@occupywallst.org](mailto:general@occupywallst.org)  
**Press Inquiries:**  
[press@occupywallst.org](mailto: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.

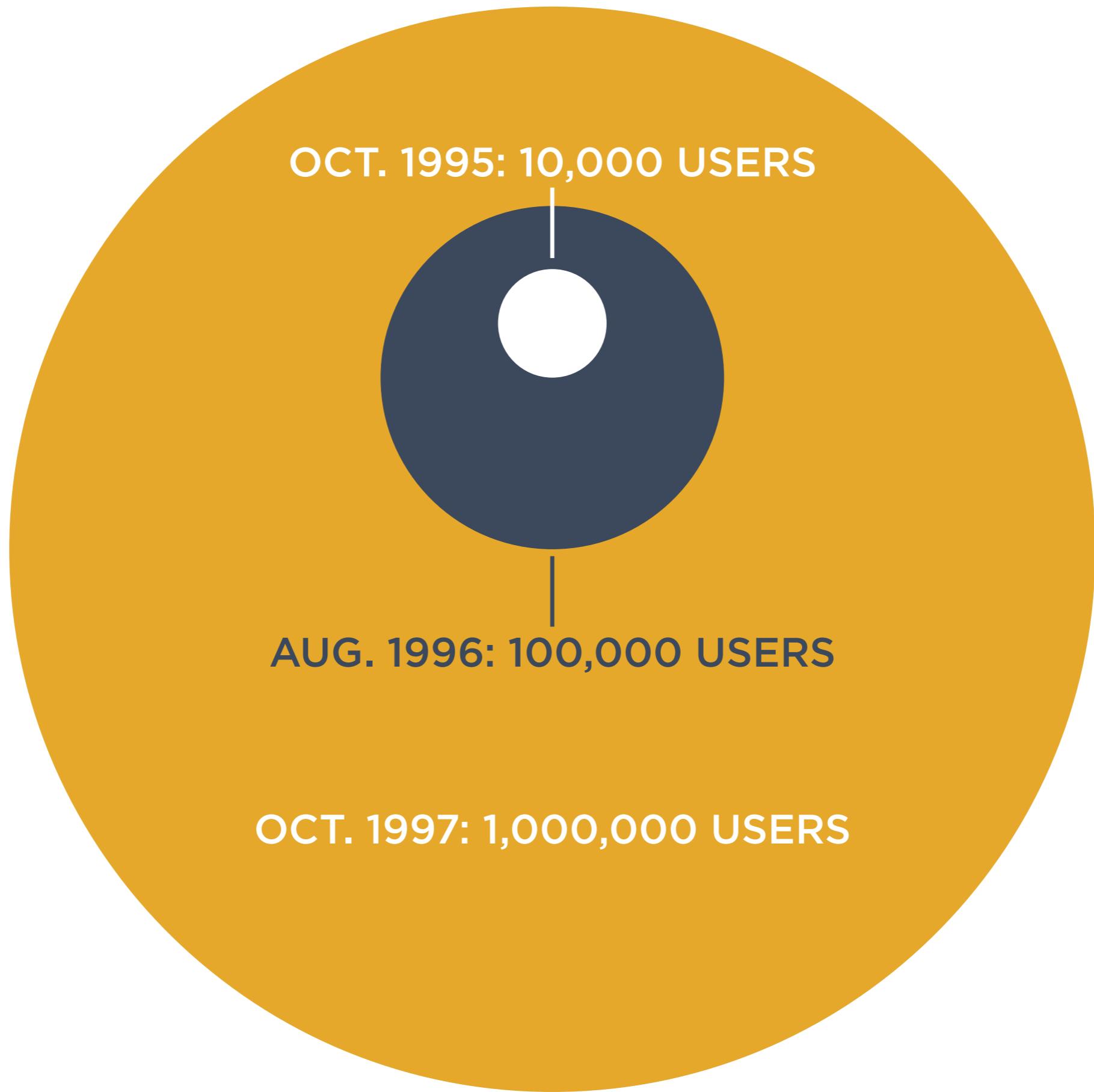
**the only solution is WorldRevolution**

[Click here](#) for NYC GA committee meeting times.

# Scarcity Abundance



# GEOCITIES USERS:



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



RIC'S GRILL  
STEAK SEAFOOD  
& CHOP HOUSE

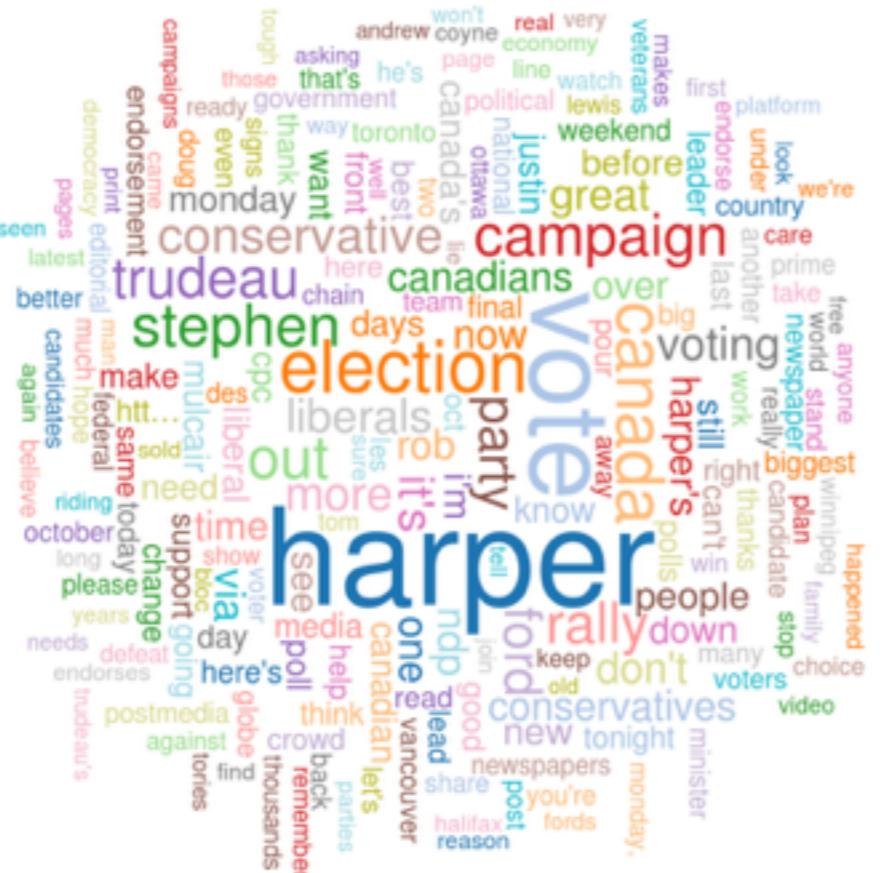
Wharf 800  
Kings Rd

Crown Royal  
Centrum

81

IDLE NO MORE

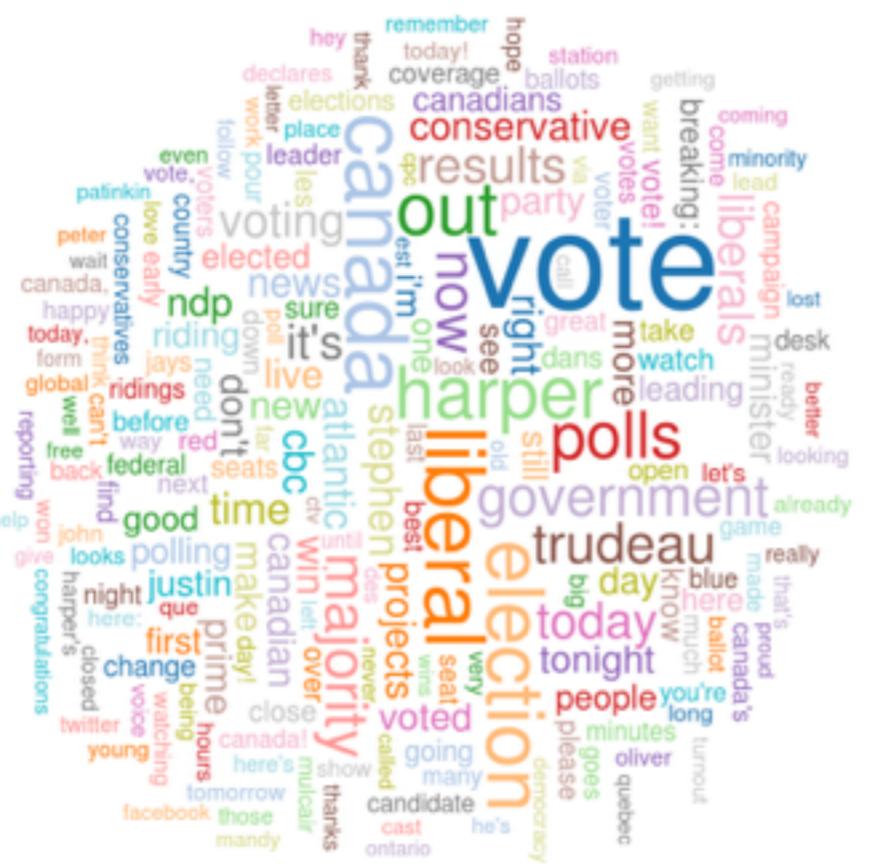
October 17, 2015



October 18, 2015



October 19, 2015

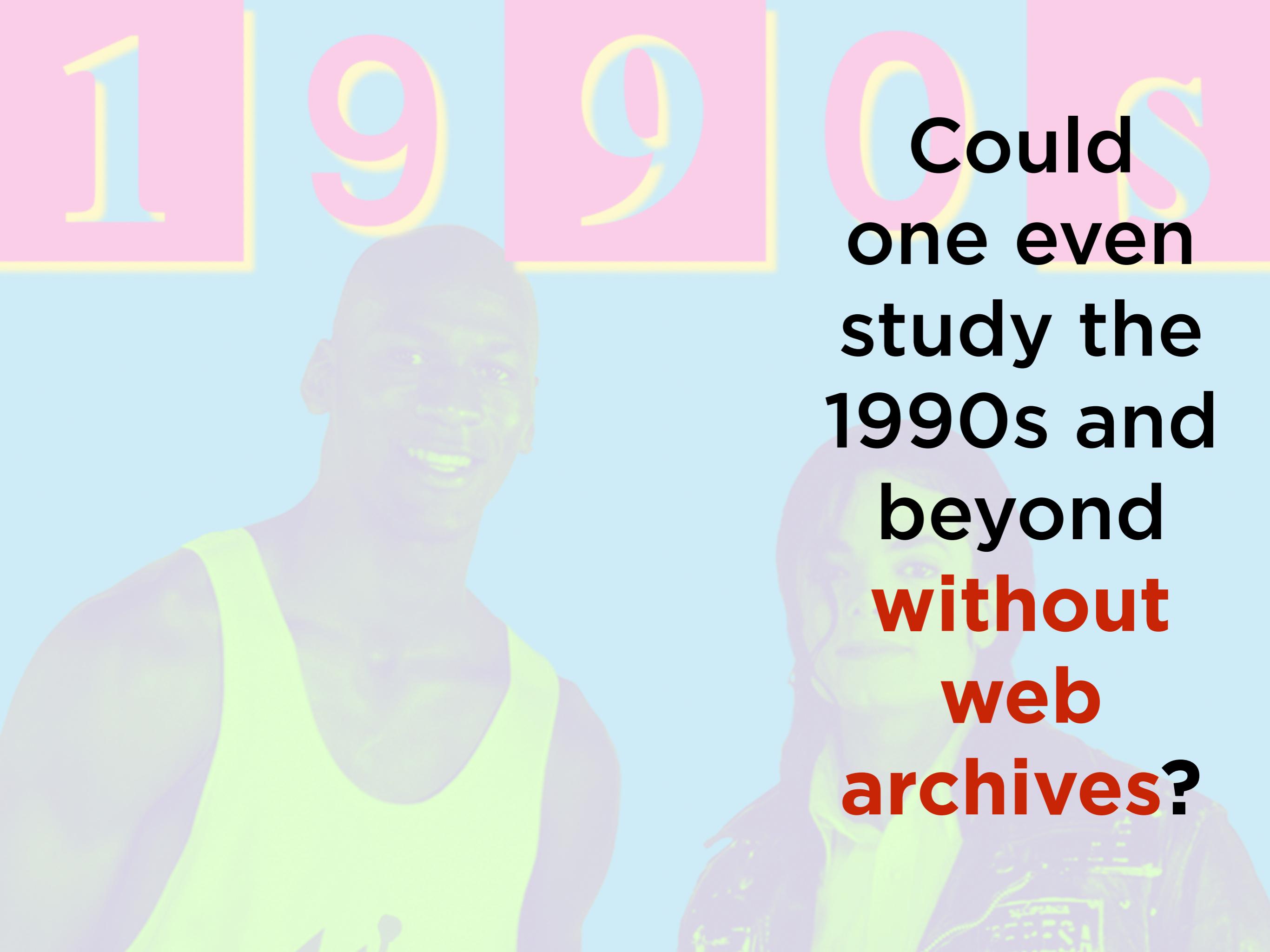


October 20, 2015



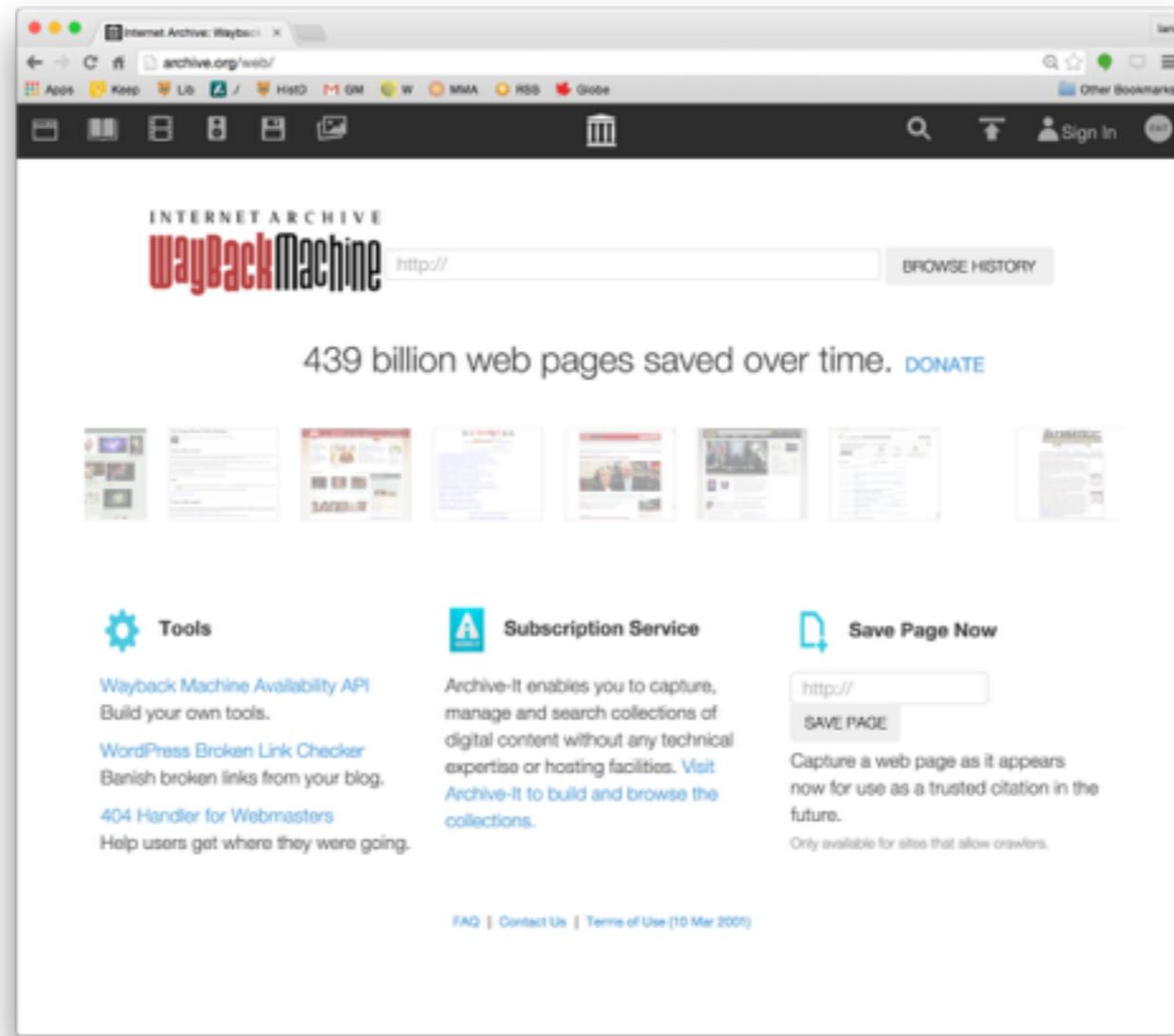
**“.... [n]ow expectations have inverted. Everything may be recorded and preserved, at least potentially.”**

**- James Gleick**



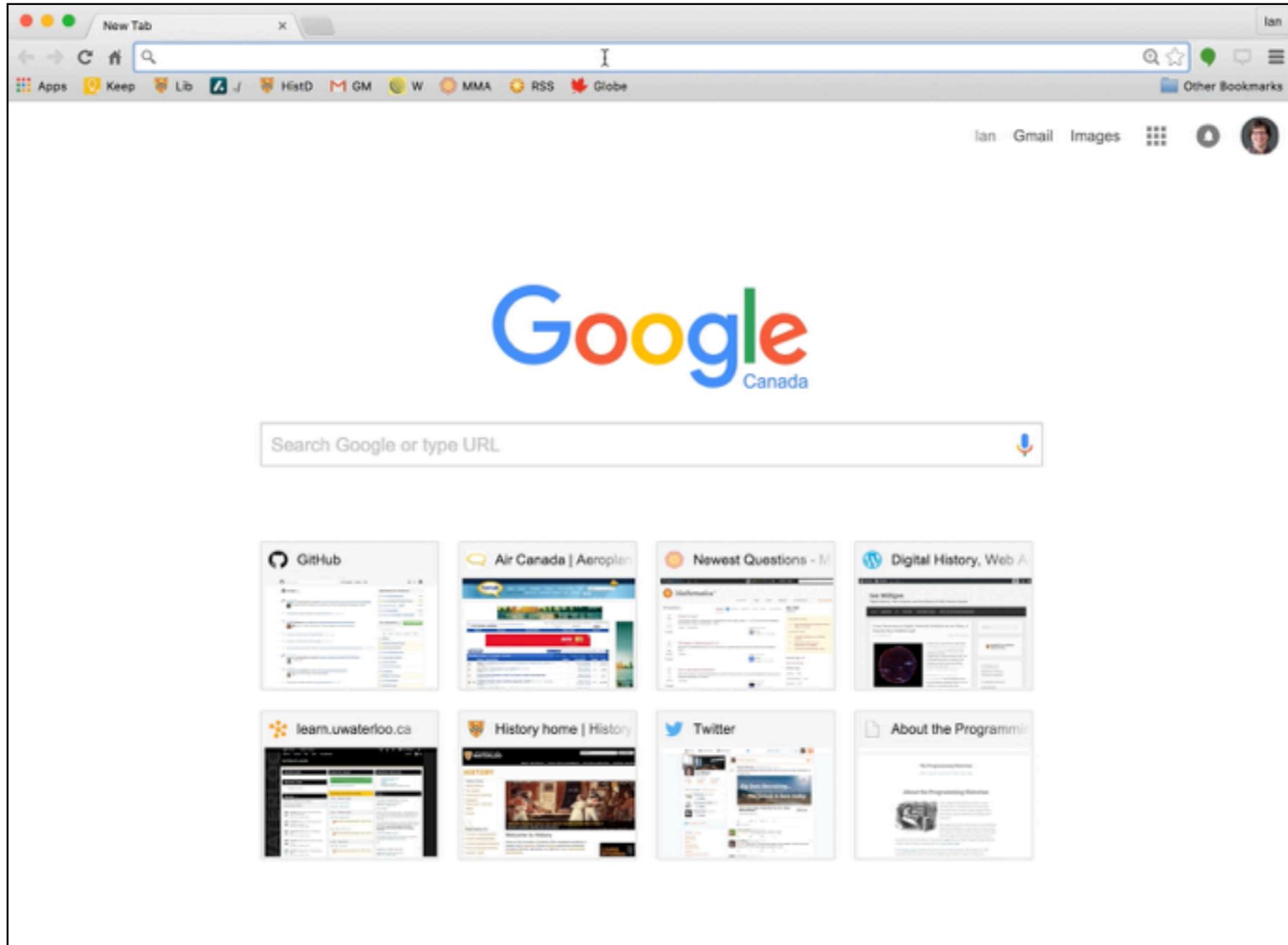
Could  
one even  
study the  
1990s and  
beyond  
**without**  
**web**  
**archives?**

# Nightmare Scenario



This won't be enough!

# Nightmare Scenario



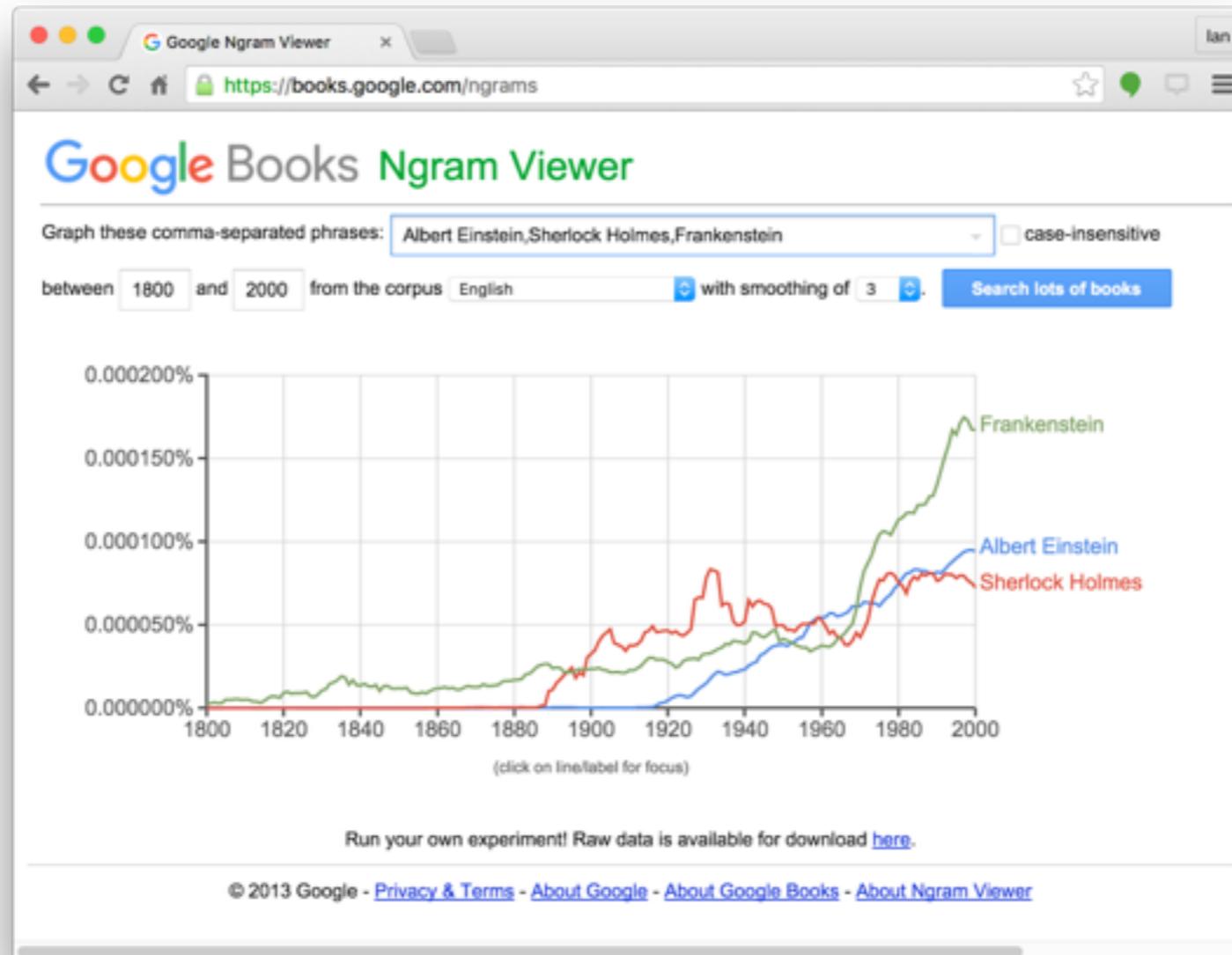
This won't be enough!



... but what will our  
search engines look  
like?

# Nightmare Scenario

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



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

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

- Erez Lieberman Aiden and Jean-Baptiste Michel

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

# Building Portals

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

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", with details like "Archived since: Oct, 2005" and "Description: Canadian Political Parties and Political Interest Groups". A large green button labeled "ARCHIVE-IT" is prominent. To the right, there's a section titled "Narrow Your Results" with a search bar and buttons for "Sites" and "Search Page Text". A list of subjects is provided, including "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 links for "Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | UF" and "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 displaying the URL <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 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". A green banner for "Canadian Political Parties and Political Interest Groups" is displayed, mentioning it was collected by the University of Toronto since Oct, 2005, and includes subjects like Politics & Elections. The main content area shows a search bar with "Stephen Harper" and a "Search" button. It also features sections for "Advanced Search" with fields for "Contains all of:", "Exact phrase:", "Not containing:", and "From the Host:" (with an example of "ex. www.archive-it.org"). On the right, there's a "Search Page Text" section, a message about 1,213,132 total results, and a link to "Stephen Harper | Facebook". At the bottom, there's a summary of the captured content, including URLs and dates.

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 / +](#)

File	Commit Message	Date
GilHoggarth	Added trailing slash to web archive url	Latest commit 11ace26 on Sep 18
python	jisc ssd ~506k ~optimized to 7 segments	2 months ago
shine	Added trailing slash to web archive url	2 months ago
.gitattributes	gitattribute file	2 years ago
.gitignore	Prevent cache being versioned.	a year ago
.gitmodules	Initial "check-in" of the bootstrap submodule	2 years ago
.travis.yml	Looks like it's simpler than I expected.	a year ago
README.md	Added Travis-CI status and a brief outline.	2 years ago

README.md

# Shine

A prototype web archives exploration UI, based on a Solr back-end that has been populated using the [warc-discovery](#) indexer.

build passing

Code

- Issues 40
- Pull requests 0
- Wiki

Pulse

Graphs

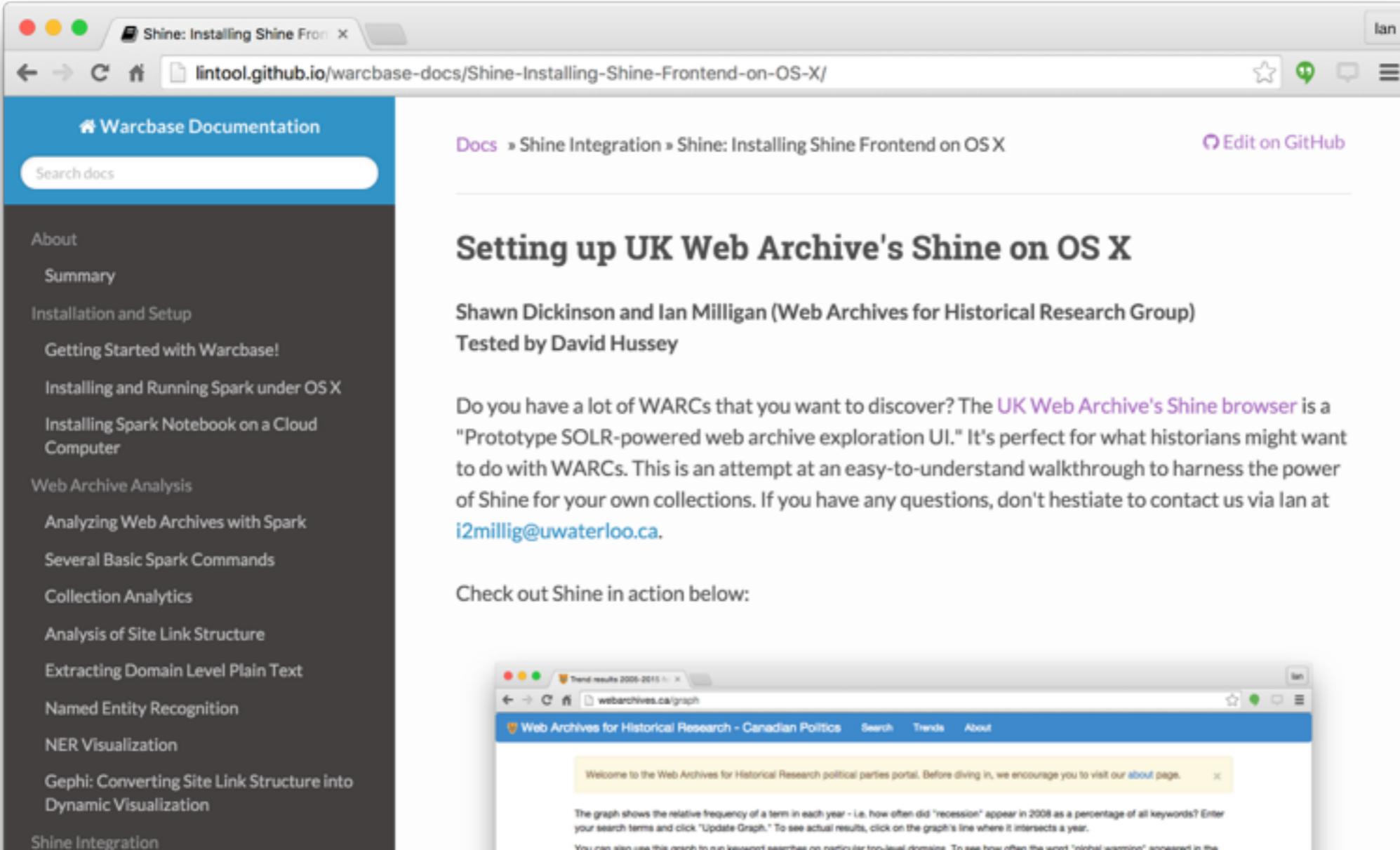
HTTPS clone URL <https://github.com>

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

Clone in Desktop

Download ZIP

# Walkthroughs at: [http://lintool.github.io/ warcbase-docs/Shine-Installing- Shine-Frontend-on-OS-X/](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 Frontend on OS X
- Address Bar:** lintool.github.io/warcbase-docs/Shine-Installing-Shine-Frontend-on-OS-X/
- Page Content:**
  - Header:** Docs > Shine Integration > Shine: Installing Shine Frontend on OS X
  - Section Title:** 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](mailto:i2millig@uwaterloo.ca).
  - Text:** Check out Shine in action below:
- Left Sidebar:** Warcbase Documentation (Search docs)
  - About
  - Summary
  - Installation and Setup
  - Getting Started with Warcbase!
  - Installing and Running Spark under OS X
  - Installing Spark Notebook on a Cloud Computer
  - Web Archive Analysis
  - Analyzing Web Archives with Spark
  - Several Basic Spark Commands
  - Collection Analytics
  - Analysis of Site Link Structure
  - Extracting Domain Level Plain Text
  - Named Entity Recognition
  - NER Visualization
  - Gephi: Converting Site Link Structure into Dynamic Visualization
  - Shine Integration

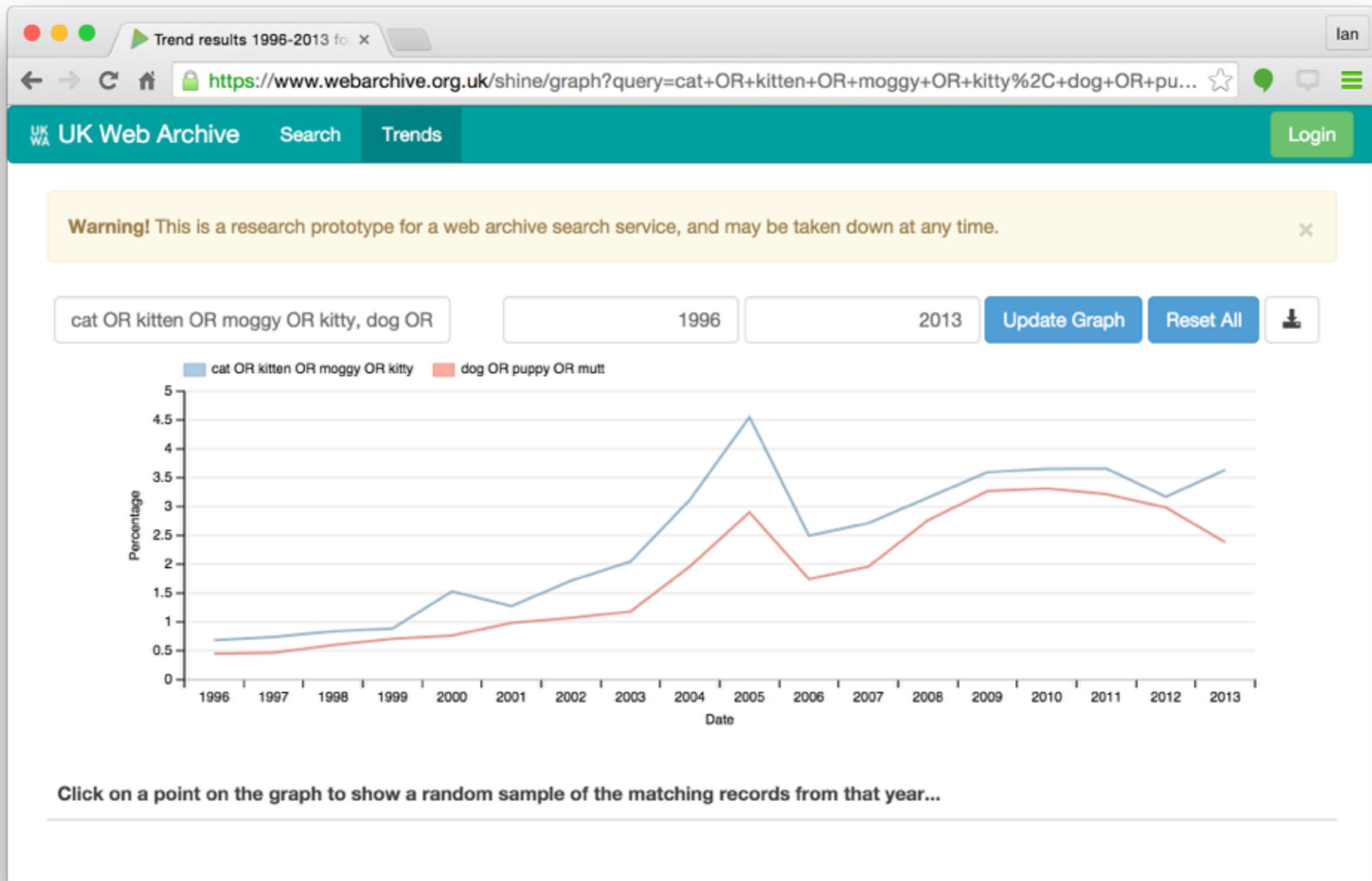
**Great research question that our  
contemporary historians were  
studying (Canada changing)**

**+**

**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

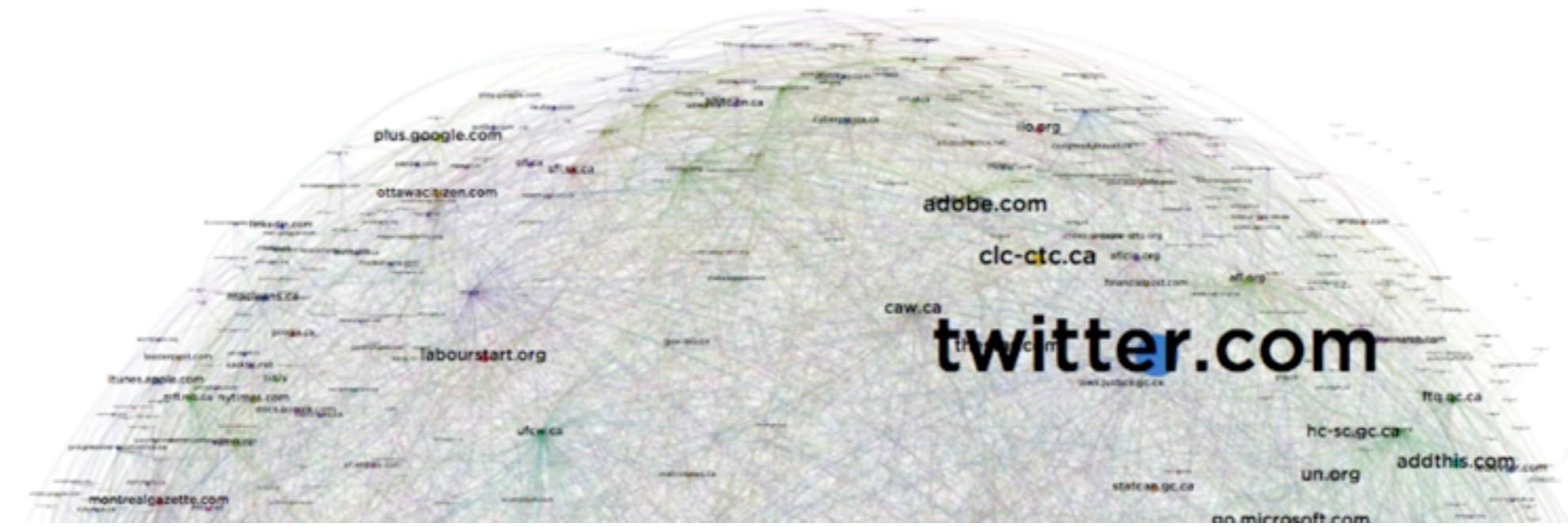
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

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

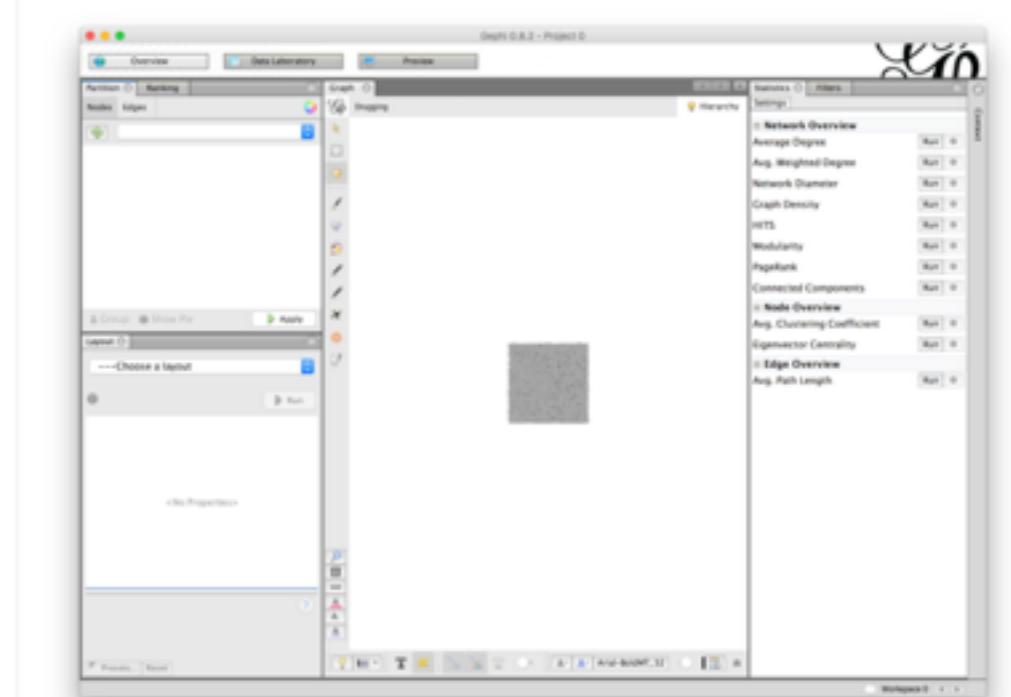
# Good for public engagement - but limited for scholarship....

The screenshot shows a web browser window for the "Web Archives for Historical Research - Canadian Politics" portal. The URL in the address bar is `webarchives.ca/search?query=stephen+harper&tab=results&action=search`. The page features a blue header with links for "Search", "Trends", and "About". A yellow banner at the top says, "Welcome to the Web Archives for Historical Research political parties portal. Before diving in, we encourage you to visit our [about](#) page." Below the banner are two search options: "Search" and "Advanced Search". The "Search" section includes a "Sample Mode" dropdown set to "stephen harper", a "Search" button, and a "Reset" button. To the left of the search bar is a "General Content Type" filter with six categories: html (1,085,201), other (71,691), pdf (3,947), audio (341), text (106), and image (14). Below this is a "Crawl Years" filter for the years 2008, 2010, 2007, 2006, 2011, and 2014. At the bottom of the search area, it says "Results 1 to 10 of 1,161,300" and has buttons for "CSV" and "Asc". The main content area below the search bar is currently empty.

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

# Gephi 0.9 (<http://gephi.github.io/>)

Walkthrough at  
[ianmilligan.ca](http://ianmilligan.ca): “From  
Dataverse to Gephi” -  
try it on this data!



From Dataverse to Gephi: ianmilligan.ca/2015/12/11/from-dataverse-to-gephi-network-analysis

My Sites Reader

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.

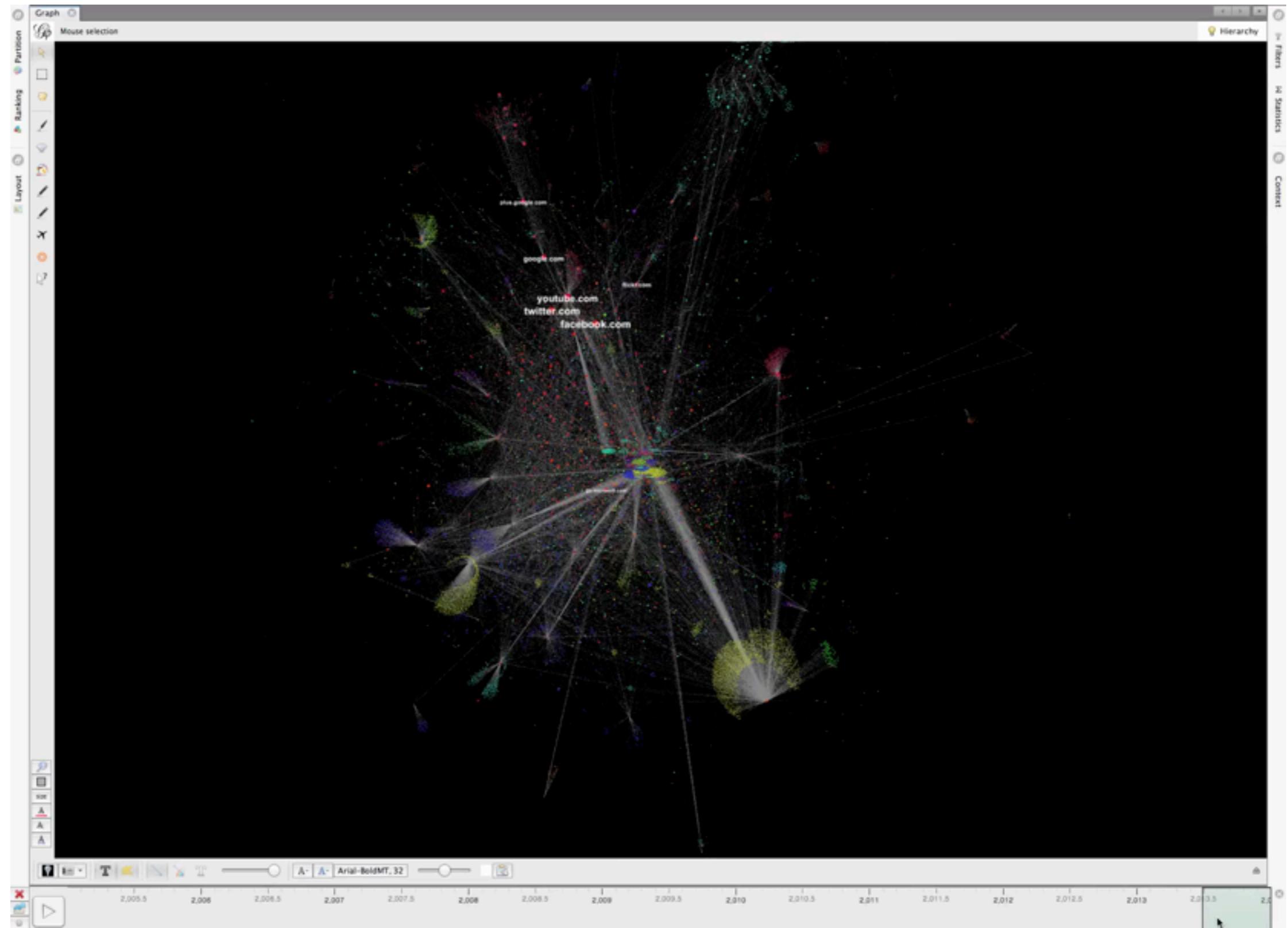
You should now see what I (nerdily) call a borg cube. That’s good, because it means that the data is in there. We need to make it usable, however.

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

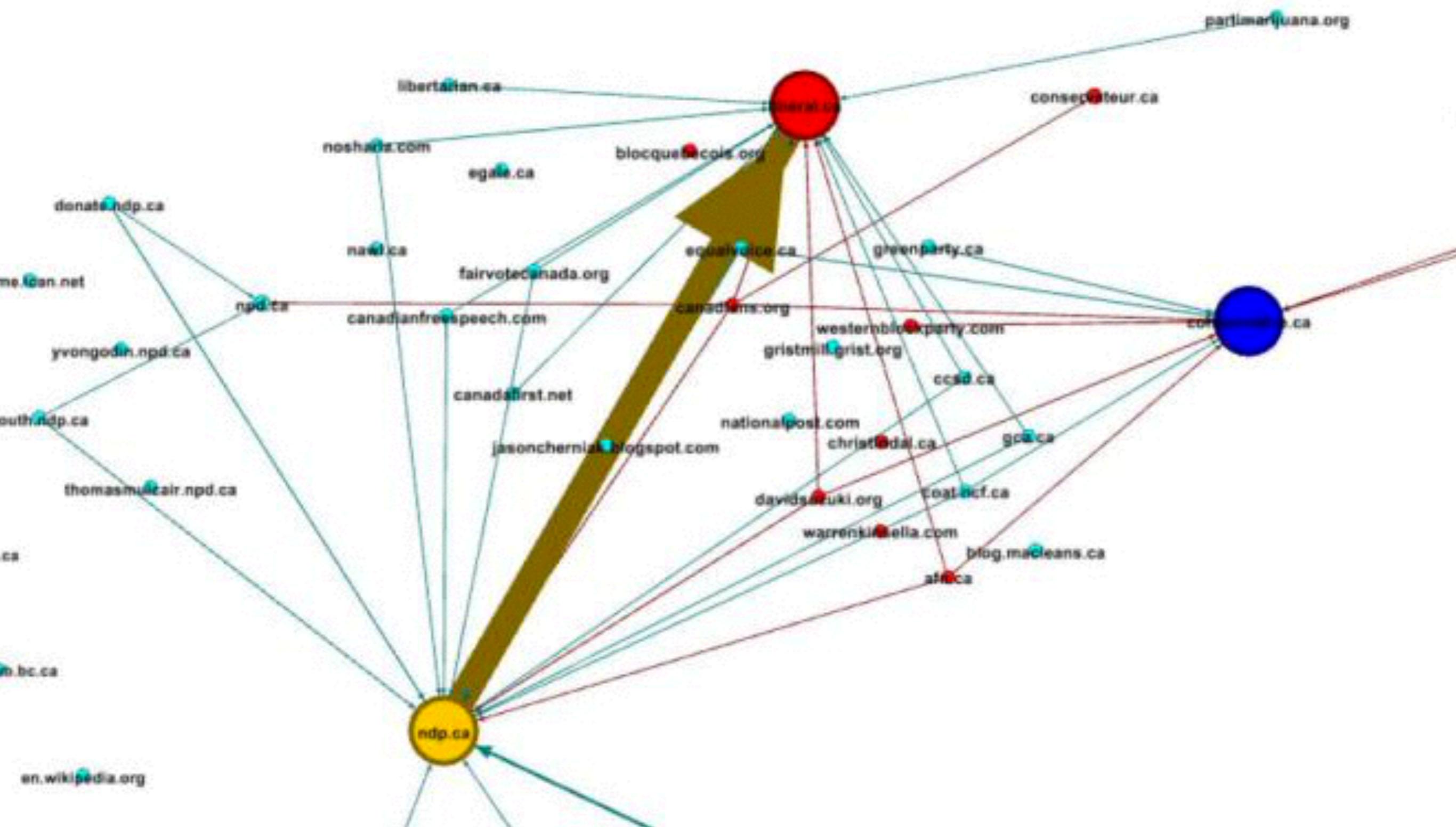


December 2008

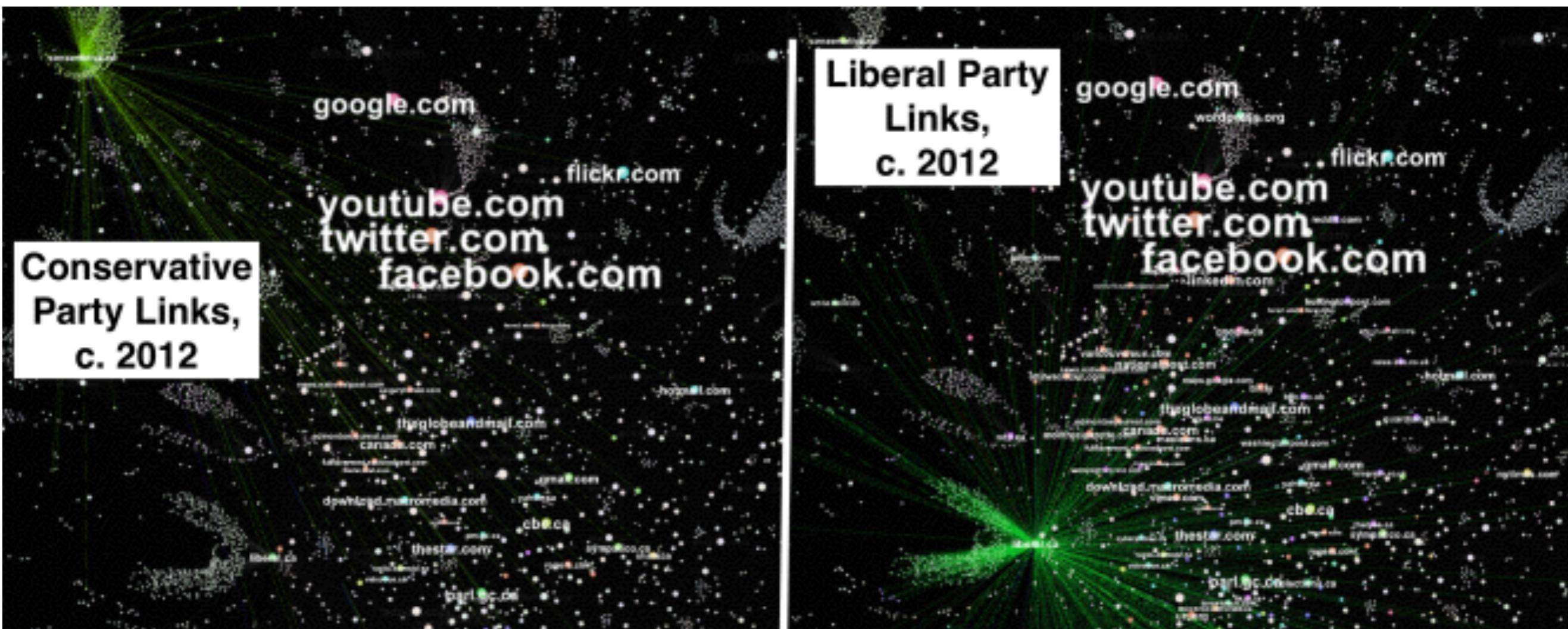
## Election campaign Ends; Attacking Harper on Anti-American Grounds (bushharper)



# 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)



# Metadata Extraction

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

<b>colincarriemp.ca</b>	12
<b>colincarriemp.ca&amp;lang=fr</b>	12
<b>colinmayes.ca</b>	12
<b>colinmayes.ca&amp;lang=fr</b>	12
<b>congrespcc.ca</b>	12
<b>conservateur.ca</b>	12
<b>conservateur.us5.list-manage.com</b>	12
<b>conservative.ca</b>	12
<b>conservative.us5.list-manage.com</b>	12
<b>consumersfirst.ca</b>	12
<b>corneliuchisu.ca</b>	12
<b>corneliuchisu.ca&amp;lang=fr</b>	12
<b>costasmenegakis.ca</b>	12
<b>costasmenegakis.ca&amp;lang=fr</b>	12
<b>cpcconvention.ca</b>	12

# Metadata Extraction

- **Conservative themes (2014):** economic development, family, immigration, legislation, women's issues, senior issues, Ukrainians, constituency offices, some prominent (and not-so-prominent) MPs, and of course, our economic action plan.
- **Liberal themes (2014):** Justin Trudeau (the new leader), cuts to social programs, child poverty, mental health, municipal issues, labour, workers, Stop the Cuts, and housing.

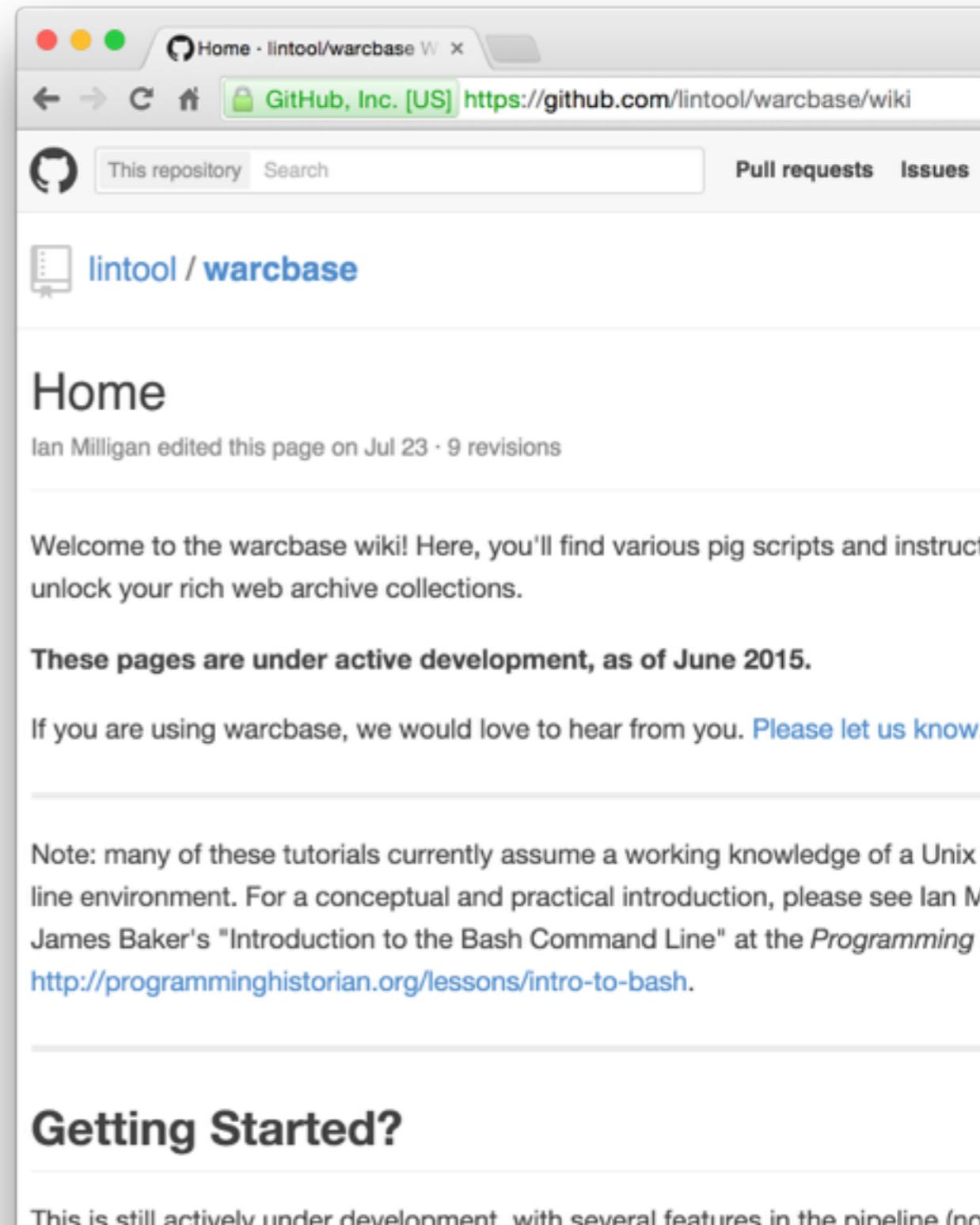
# Metadata Extraction

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

# **Interdisciplinary**

# Warcbase

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



# Extract all Plain Text

```
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
el 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

200606  
Andrew Lewis  
Bill  
Bill Hulet  
Brown  
Bruce Abel  
Bush  
Camille Labchuk  
Chandler  
Cherfi  
Chernushenko  
David

David Chernushenko

David Chernushenko

David Kay  
Derek Pinto  
Ed Broadbent

Elizabeth May  
Eric Walton  
Fannon  
Gomery  
Green

Harper

Harris

Jim  
Jim Fannon

Jim Harris  
Jim Harris Speech  
John

Julie Baribeau  
Junker  
Kevin Colton  
Labchuk  
Layton

Leonardo DiCaprio  
Manley  
Mark Brooks  
Mark MacGillivray  
Martin  
Michael Robinson  
Milliken  
Paul Martin  
Peter Martin

200607  
Adrienne Carr  
Andrew Lewis  
Bill  
Bill Hulet  
Brown  
Bruce Abel  
Bush  
Camille Labchuk  
Chandler  
Cherfi  
Chernushenko  
David

David Chernushenko

David Kay  
Derek Pinto

Dietrich  
Ed Broadbent

Elizabeth May  
Eric Walton  
Fannon  
Gomery

Green  
Harper

Harris

Jim  
Jim Fannon

Jim Harris  
Jim Harris Speech  
John  
Julie Baribeau  
Junker  
Kevin Colton  
Labchuk  
Layton  
Manley

200608  
Adrienne Carr  
Allan Gribbin  
Amélie Gingras  
Andrew Lewis  
Bill  
Bill Hulet  
Brown  
Bruce Abel  
Bush  
Camille Labchuk  
Chandler  
Cherfi  
Chernushenko  
Clements Verhoeven  
David

David Chernushenko

David Kay

Derek Pinto

Dietrich

Ed Broadbent

Elizabeth May

Eric Walton

Fannon

Gomery

Green

Harper

Harris

Jim

Jim Harris  
Jim Harris Speech  
John  
Julie Baribeau  
Junker  
Kevin Colton  
Kootenay-Columbia Jo...  
Labchuk  
Lawrence Redfern  
Layton  
Manley  
Mark Brooks

200609  
Adrienne Carr  
Amélie Gingras  
Brown  
Bruce Abel  
Bush  
Cameron Wigmore  
Chandler  
Cherfi  
Chernushenko  
Chretien  
David

David Chernushenko

David Kay  
Derek Pinto

Dietrich  
Dion  
Elizabeth

Elizabeth May

Elizabeth May  
10 mentions

Elizabeth Peloza

Eric Walton

Gomery

Green

Harper

Harris

Jasper

Jim

Jim Harris

Jim Harris Speech

John

Labchuk

Lougheed

Mackenzie

Manley

Martin

May

Mona Elaine Adlman ...

Paul Martin

Peter Foster

Pierre Pettigrew

Schiller

200610  
Ambrose  
Andrew Lewis  
Bill  
Bridget Doherty  
Bush  
Carol Gudz  
Catharine Johannson  
Chandler  
Cherfi  
Chernushenko  
Daphne Wysham  
David

David Chernushenko

David Kay  
Derek  
Derek Pinto

Dundas

Elizabeth

Elizabeth Goes  
Elizabeth May

Elizabeth May Say

Eric Walton

Gagnon

Gomery

Green

Grenon

Halton

Harper

Harris

Jim

Jim Harris

John

Jude Larkin

Judith

Kyle Grice

Labchuk

Manley

Mark MacGillivray

Martin

May

Melanie Ransom

Michael Grayson

Michele

Paul Martin

Richard Reble

Sharon Labchuk

Stefanie Moore

200611  
Ambrose  
Andrew Lewis  
Bill  
Bill Clinton  
Bush  
Chandler  
Cherfi  
Chernushenko  
Chris Alders  
Daphne Wysham  
David

David Chernushenko

David Cox

David Kay  
David Suzuki  
Derek

Derek Pinto  
Dundas

Edward Burtynsky

Elizabeth

Eric Walton

Garth Turner

Gomery

Green

Halton

Harper

Harris

Jim

Jim Harris

Jim Harris Speech

John

Julie Baribeau

Labchuk

Manley

Margaret

Mark MacGillivray

Martin

May

Paul

Paul Martin

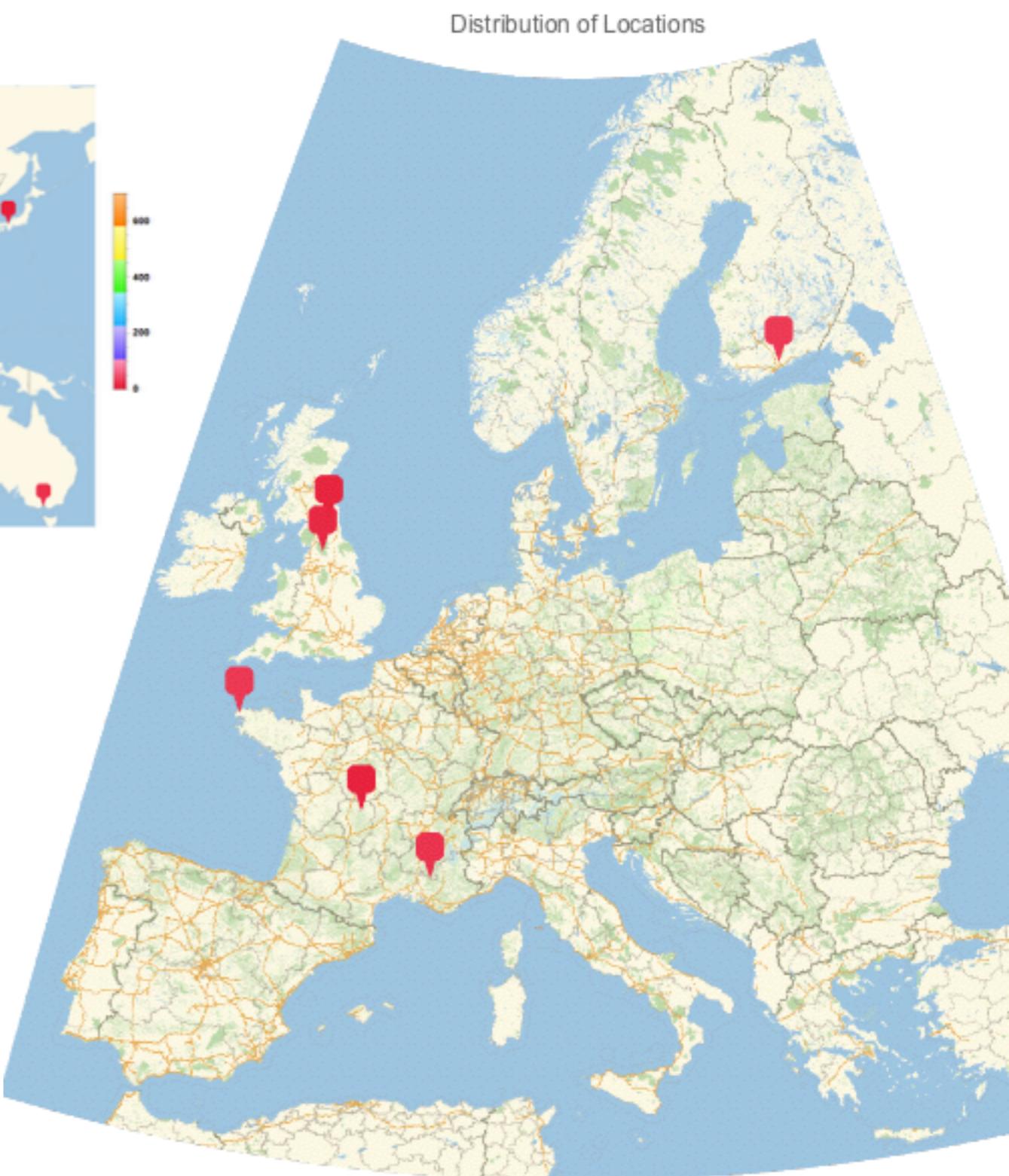
Ross

Sharon Labchuk

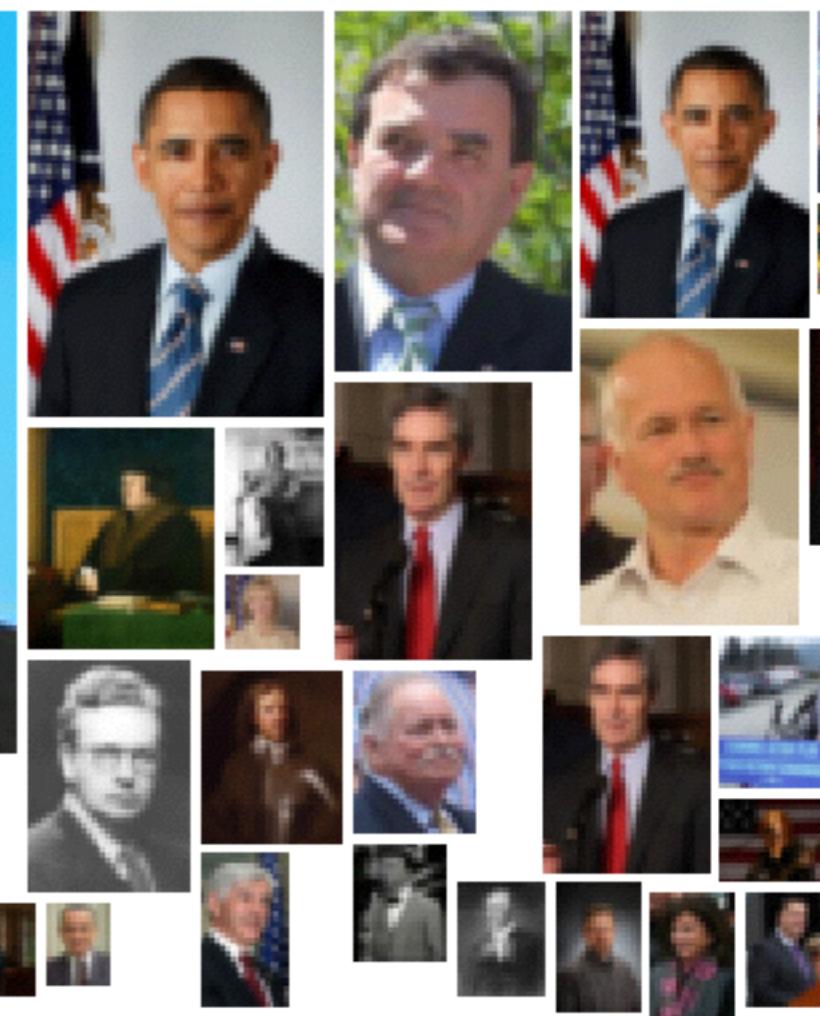
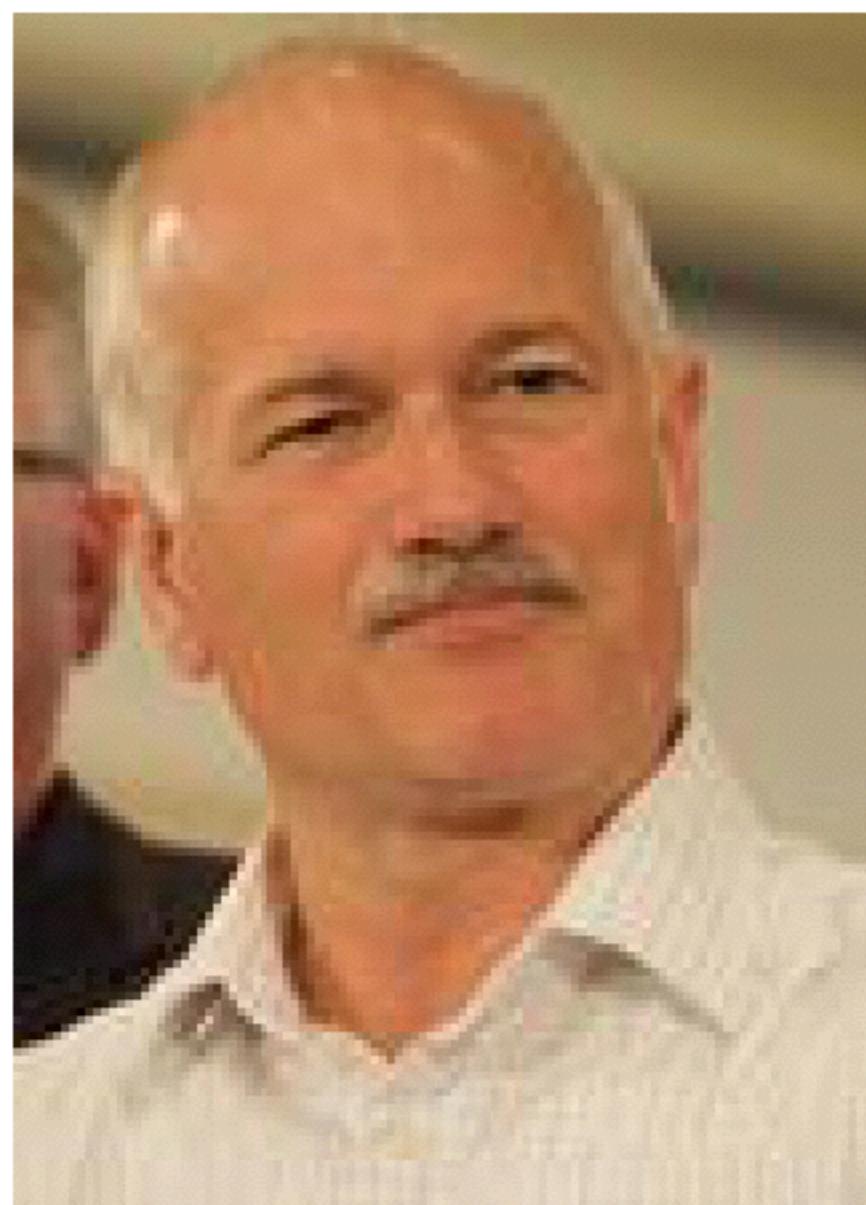
# Extract Entities



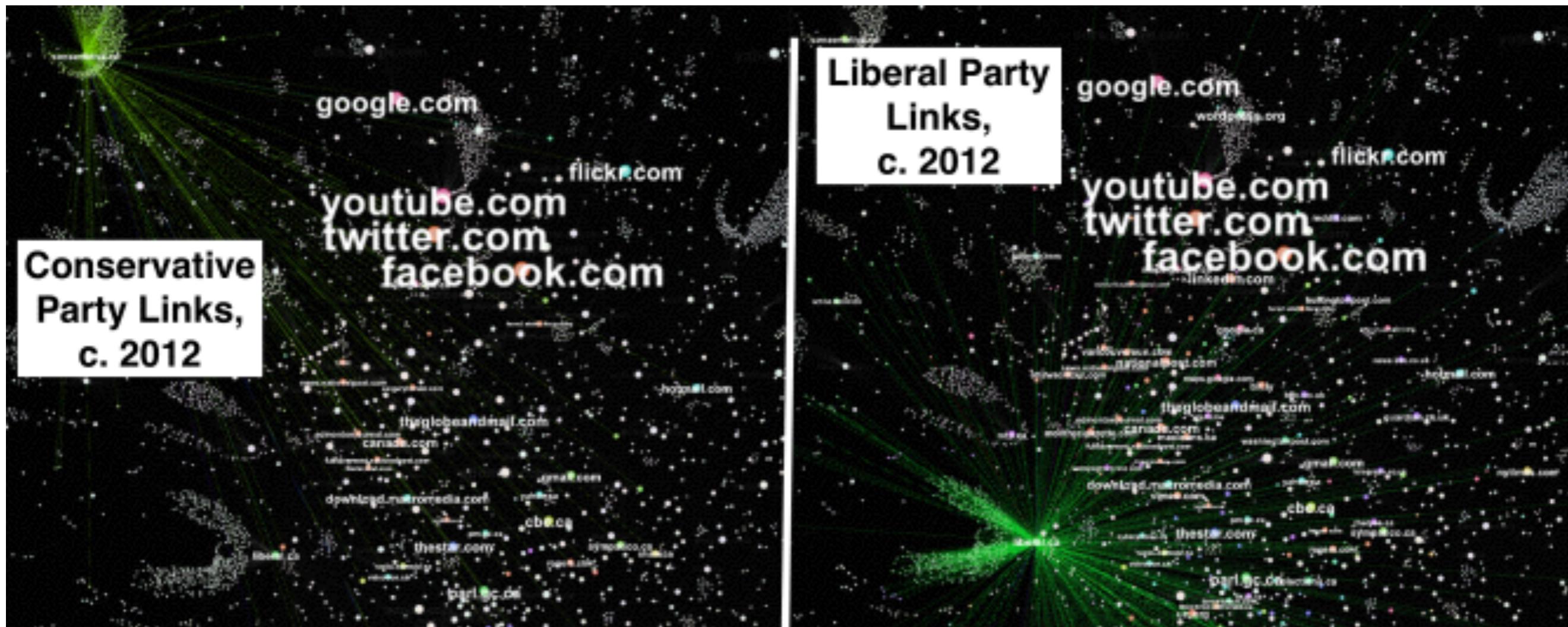
```
In[26]:= int = SemanticInterpretation[#] & /@ processedfreq[[All, 1]]  
Out[26]= {Canada, Calgary, Colombia, Montreal, $Failed, Ontario, Canada, Afghanistan,  
Ottawa, Manitoba, Canada, British Columbia, Canada, Toronto, Nova Scotia, Canada,  
Saskatchewan, Canada, Quebec, Canada, Alberta, Canada, 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,
```



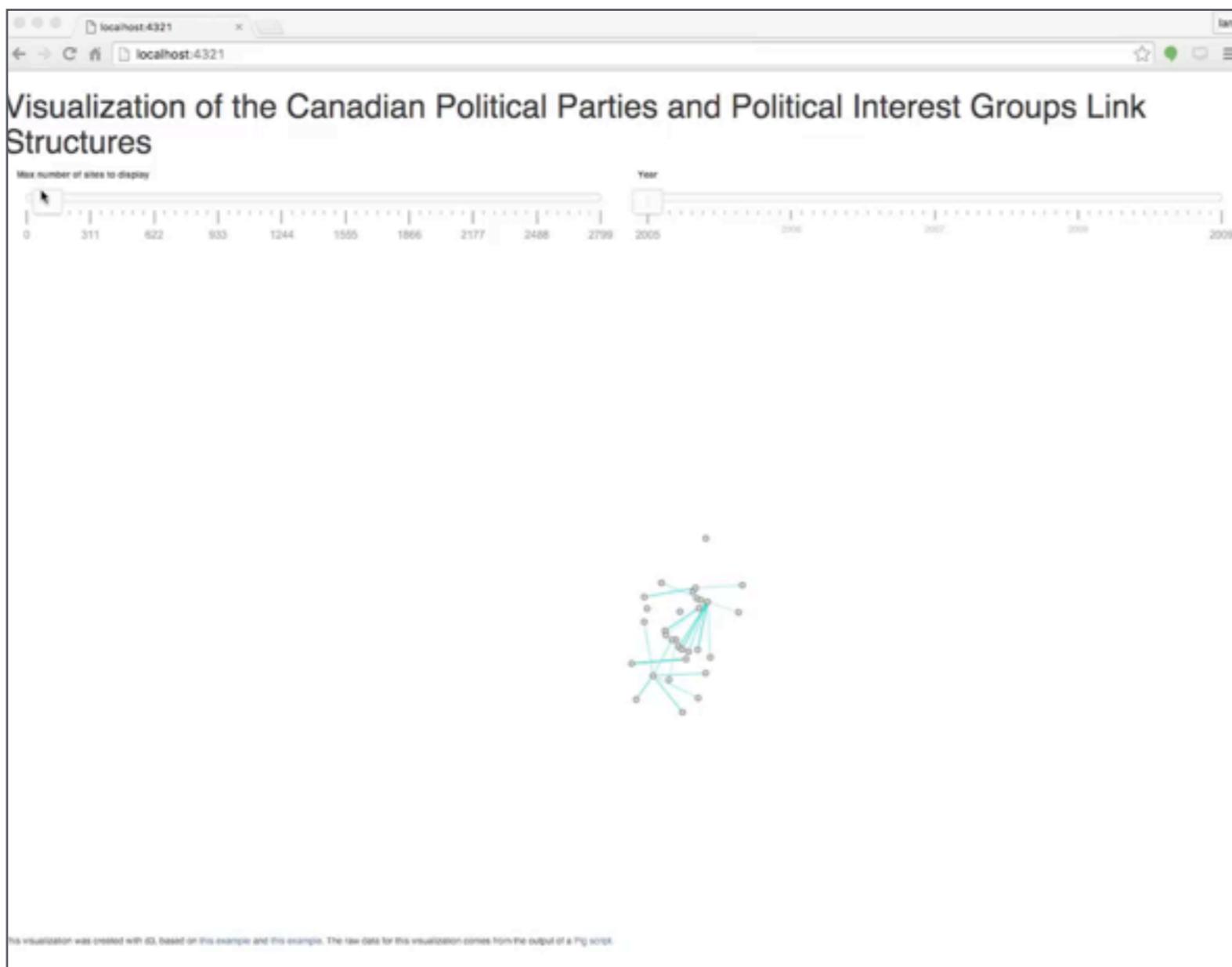
# Extract Entities



# Extract Links/Gephi Connector



# Or D3.js link networks in browser



All walkthroughs at:  
[docs.warcbase.org](https://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 URL [programminghistorian.org](http://programminghistorian.org) in the address bar. The page title is "The Programming Historian". Below the title are navigation links: "About · Lessons · Contribute · Project Team · Blog". The main section features a large heading "About the Programming Historian" and a black and white illustration of a person working at a computer terminal. To the right of the illustration is a descriptive text block. At the bottom, there is a call to action for reviewers and contributors.

## 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.

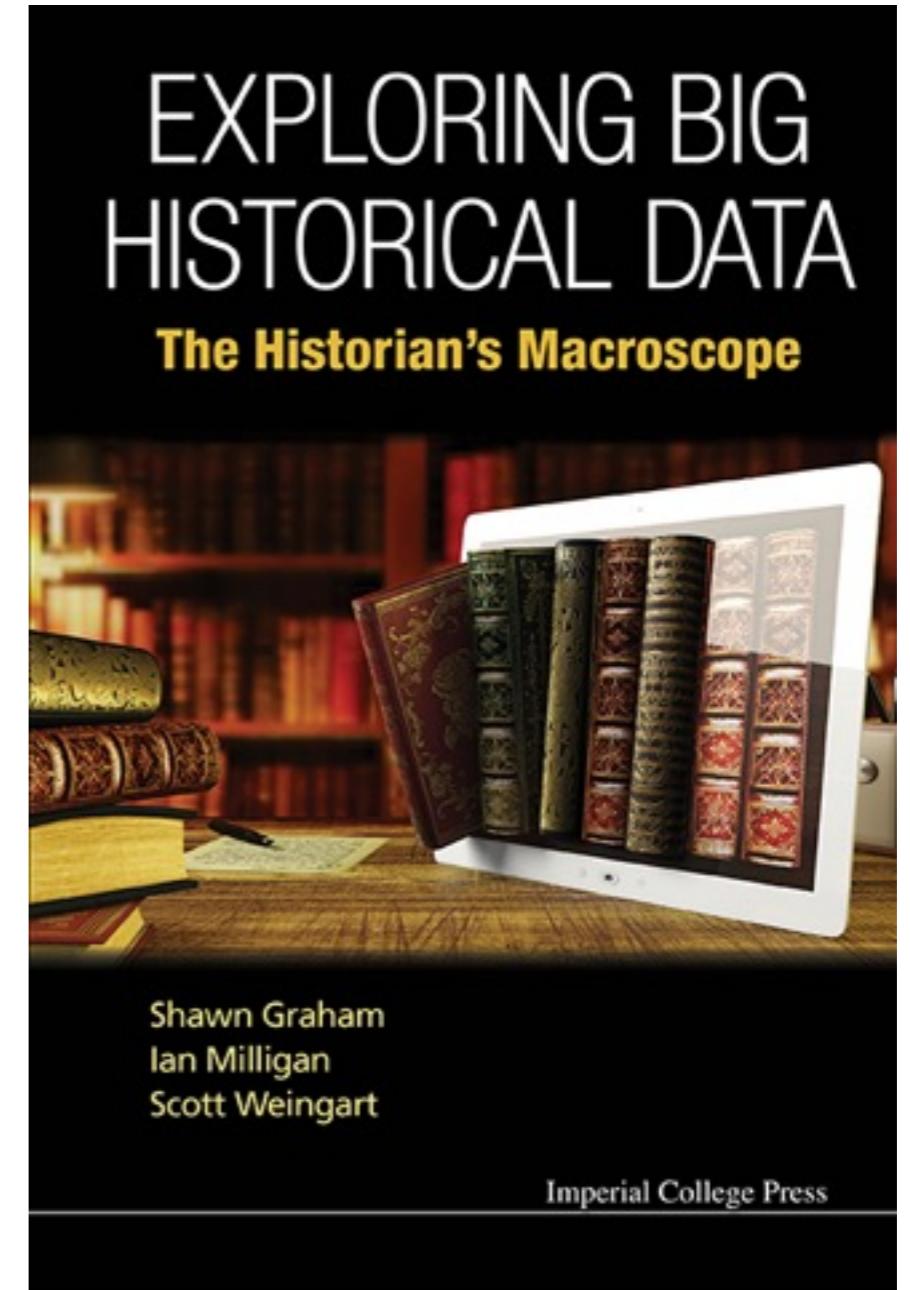
Our [Project Team](#) and peer reviewers work collaboratively with authors to craft tutorials that illustrate fundamental digital and programming principles and

# Programming Historian

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

# Exploring Big Historical Data

- Check out our draft at [macroscope.org](http://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.**



Social Sciences and Humanities  
Research Council of Canada

Conseil de recherches en  
sciences humaines du Canada

Canada



compute \* calcul  
CANADA



UNIVERSITY OF  
**WATERLOO**

# Thanks very much!

## Questions?

---

Ian Milligan  
Assistant Professor  
@ianmilligan1



UNIVERSITY OF WATERLOO  
FACULTY OF ARTS  
Department of History