

# The Record of our Recent Past

**Enabling Large-Scale Web Archive Data Mining**

---

**Ian Milligan**  
Assistant Professor



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

# Later Today

CFI-seminar: Web history: x Ian

cfi.au.dk/news/article/artikel/cfi-seminar-web-history-geocities-and-news-websites/    

AARHUS UNIVERSITY Research Talent development Knowledge exchange Education About AU FOR STAFF & ST

## CENTRE FOR INTERNET STUDIES

You are here: AU » Research » Research Centres » Centre for Internet Studies » News » Article

### CFI-SEMINAR: WEB HISTORY: GEOCITIES, AND NEWS WEBSITES

On 22 September CFI will host a seminar about web history with visiting scholar Ian Milligan and CFI-members Niels Brügger and Henrik Bødker

2015.08.25 | SIGRID NIELSEN SAABYE



The title of the seminar is "Web history: Geocities, and News Websites". Visiting scholar Ian Milligan and the two CFI-members Henrik Bødker and Niels Brügger will present their work giving the following presentations:

**Welcome to the GeoHood: Exploring Early Web History through the GeoCities Torrent, 1994-2009.**  
by Ian Milligan, Assistant Professor of history at the University of Waterloo, in Waterloo, Ontario, Canada.

#### Other news

- > [Inaugural lecture: Listen \(2015.08.31\)](#)
- > [New Publication in the CF Monograph Series \(2015.\)](#)
- > [New member: Henrik Bødker \(2015.08.20\)](#)
- > [Inaugural Lecture: Niels Brügger \(2015.08.13\)](#)
- > [Niels Brügger appointed Professor \(2015.06.03\)](#)

Some links/resources collected at:

<https://github.com/ianmlligan1/>

Aarhus-Netlab

# Why?

The sheer amount of social, cultural, and political information generated every day presents new opportunities for historians.



## MATCH YOUR INTEREST TO A NEIGHBOR

<a href="#">FREE HOME PAGES AND E-MAIL</a>	<a href="#">ARTS AUTOS BUSINESS COMPUTERS CULTURE</a>	<a href="#">EDUCATION ENTERTAINMENT ENVIRONMENT FAMILY FASHION</a>	<a href="#">FOOD GAMES GAY &amp; LESBIAN GOVERNMENT HEALTH</a>	<a href="#">KIDS MUSIC PEOPLE RECREATION SCIENCE FI</a>
--	---	--	--	---



[Your home on the web!](#)

A screenshot of a Netscape browser window titled "Cigardude's Smoking Room - Netscape". The address bar shows the URL "http://www.geocities.com/NapaValley/1070/". The main content area has a blue wavy background and displays the title "The Smoking Room". On the left, there's a sidebar with links: "Your Choices", "Cigars", "Wine", "Beer", "Links", and "Home". In the center, there's a cartoon illustration of a man smoking a cigar. To the right, text reads "Welcome to the Cigar Dude's Smoking Room". Below that, it says "You are visitor number" followed by a small icon. Further down, it says "since June 5, 1996". At the bottom, there's a paragraph about the purpose of the page and a "NETSCAPE Navigator 4.0" logo. The status bar at the bottom of the browser window shows "Welcome to my home page, devoted to some of the finer pleasures in life: good cigar ...".

1990s

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

**No.**

Historians need to do this now, or  
we're going to be left behind.

# Nightmare Scenario

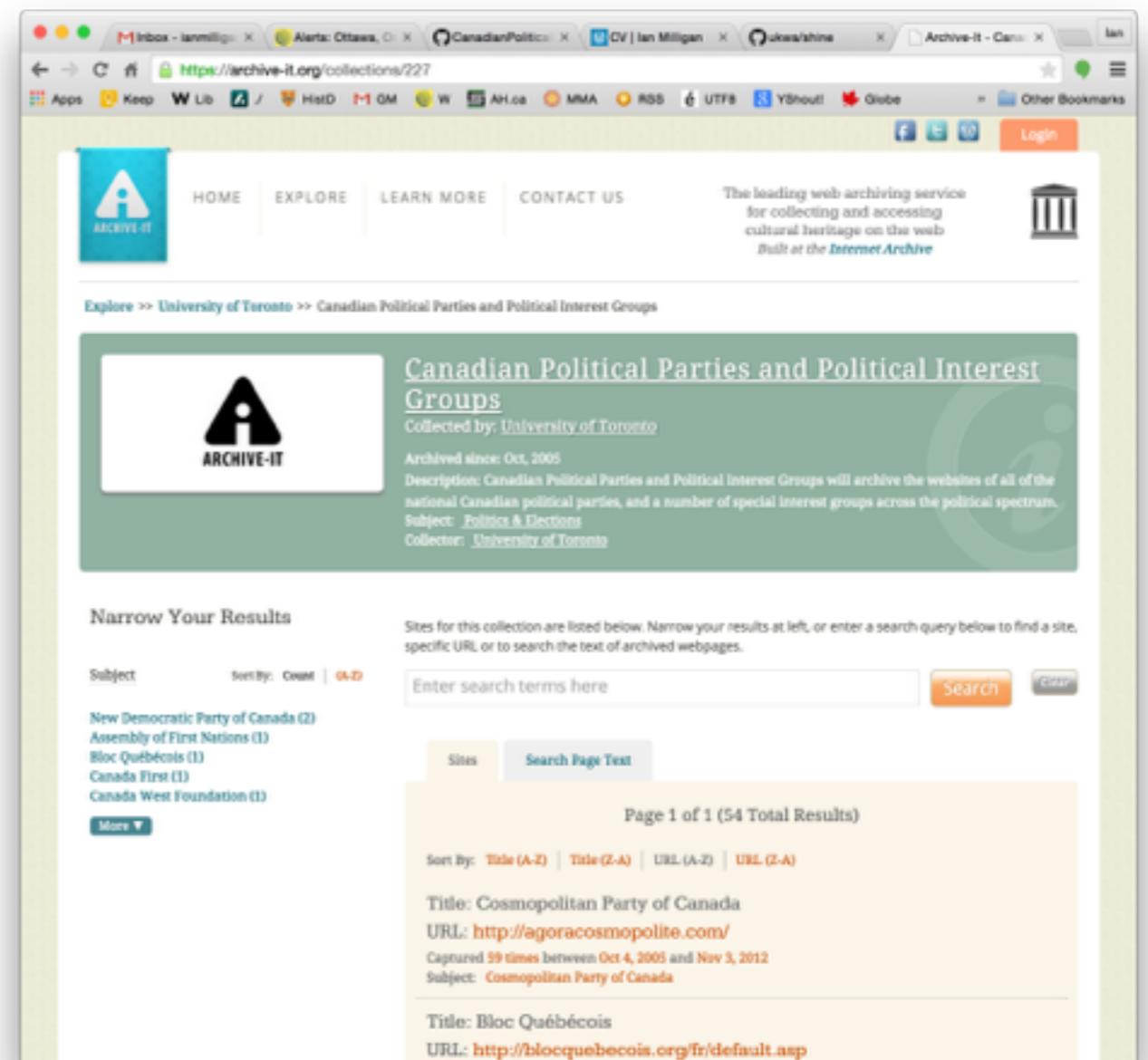
- Wayback Machine won't be enough. We won't use that.
- Historians rely uncritically on **date-ordered keyword search results**, putting them at mercy of search algorithms they do not understand;
- Historians are completely left out of post-1996 research, letting everybody else do the work (a la Culturomics project/*Nature* magazine article);
- Our profession gets left behind...

# But what will web archives look like?

- Three Distinct Case Studies
  - **Wide Web Scrape**, March - December 2011 (Internet Archive) (sample of 80TB WARC collection);
  - **GeoCities End-of-Life Torrent**, 2009 (Archive Team);
  - **Archive-It Longitudinal Collections, Canadian Political Parties & Labour Organizations**, 2005-2014 (Archive-It/University of Toronto)

# Example Dataset

- Archive-It Collection 227,  
**Canadian Political Parties and  
Political Interest Groups  
(University of Toronto)**
- October 2005 - Present
  - All major and minor political parties, as well as organized political interest groups (Council of Canadians, Coalition to Oppose the Arms Trade Assembly of First Nations, etc.)
- Started by now-retired librarian, hard to get details on seed list



# 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: "The leading web archiving service for collecting and accessing cultural heritage on the web. Built at the Internet Archive". Below the header, it says "Explore > University of Toronto > Canadian Political Parties and Political Interest Groups". The main content area features a banner for "Canadian Political Parties and Political Interest Groups" collected by the University of Toronto, with details about the collection's purpose and scope. A search bar at the bottom left contains the query "Stephen Harper". The results section displays a single result for "Stephen Harper | Facebook" with a link to the URL <http://www.facebook.com/pages/Stephen-Harper/9506562109>.

How to provide  
access?

# Two Main Approaches

- Warcbase
  - Link extraction and analytics
  - Full-text extraction and analytics
- Full-text faceted search
  - UK Web Archive's **Shine** solr front end

WAT files  
vs.  
ARC/WARC files

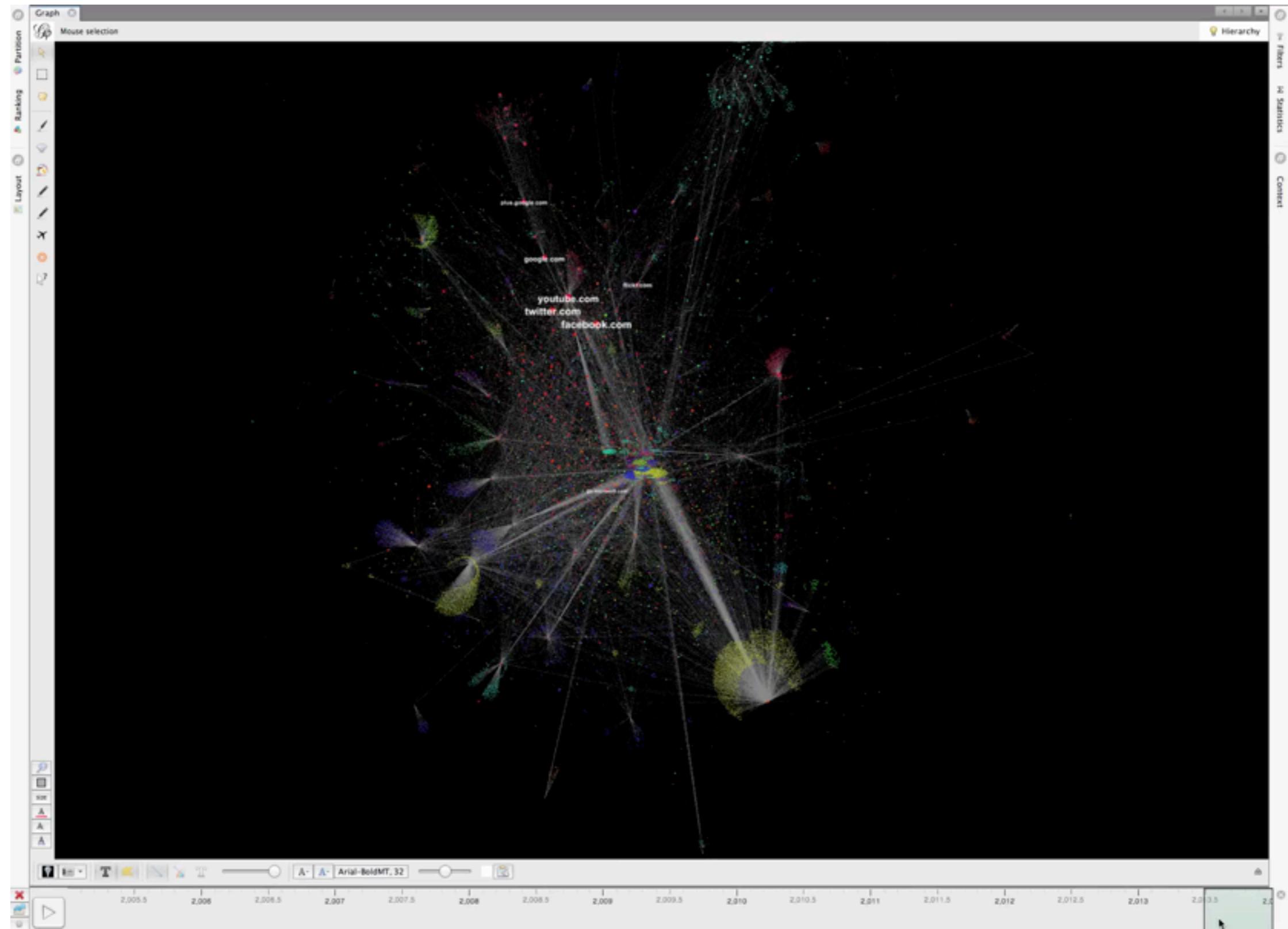
## **Problem One:**

Historians want content, but we  
can only locally work with  
metadata

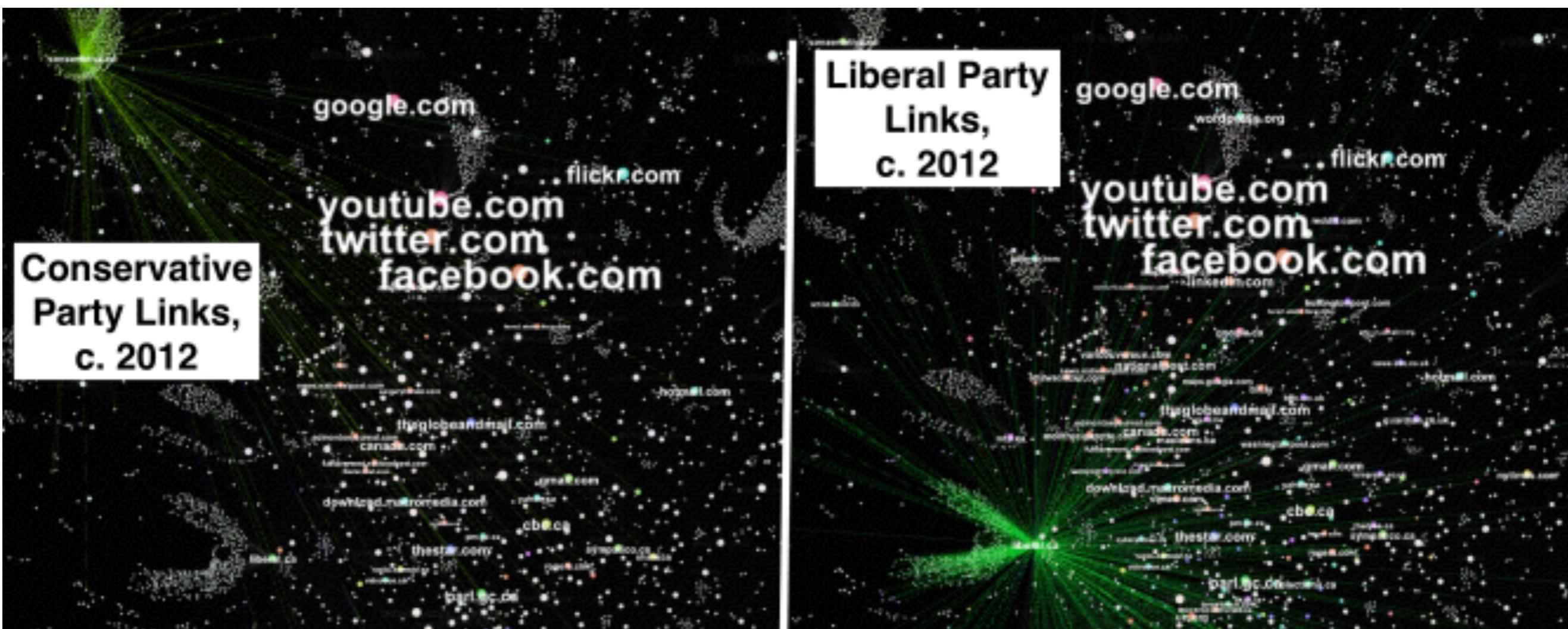
**Do we want metadata  
or content analysis?**

**Historians NEED content,  
but metadata can help us  
find and contextualize it**

# Metadata Extraction



# Metadata Extraction



# 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

- 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.

As well as short stories..

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

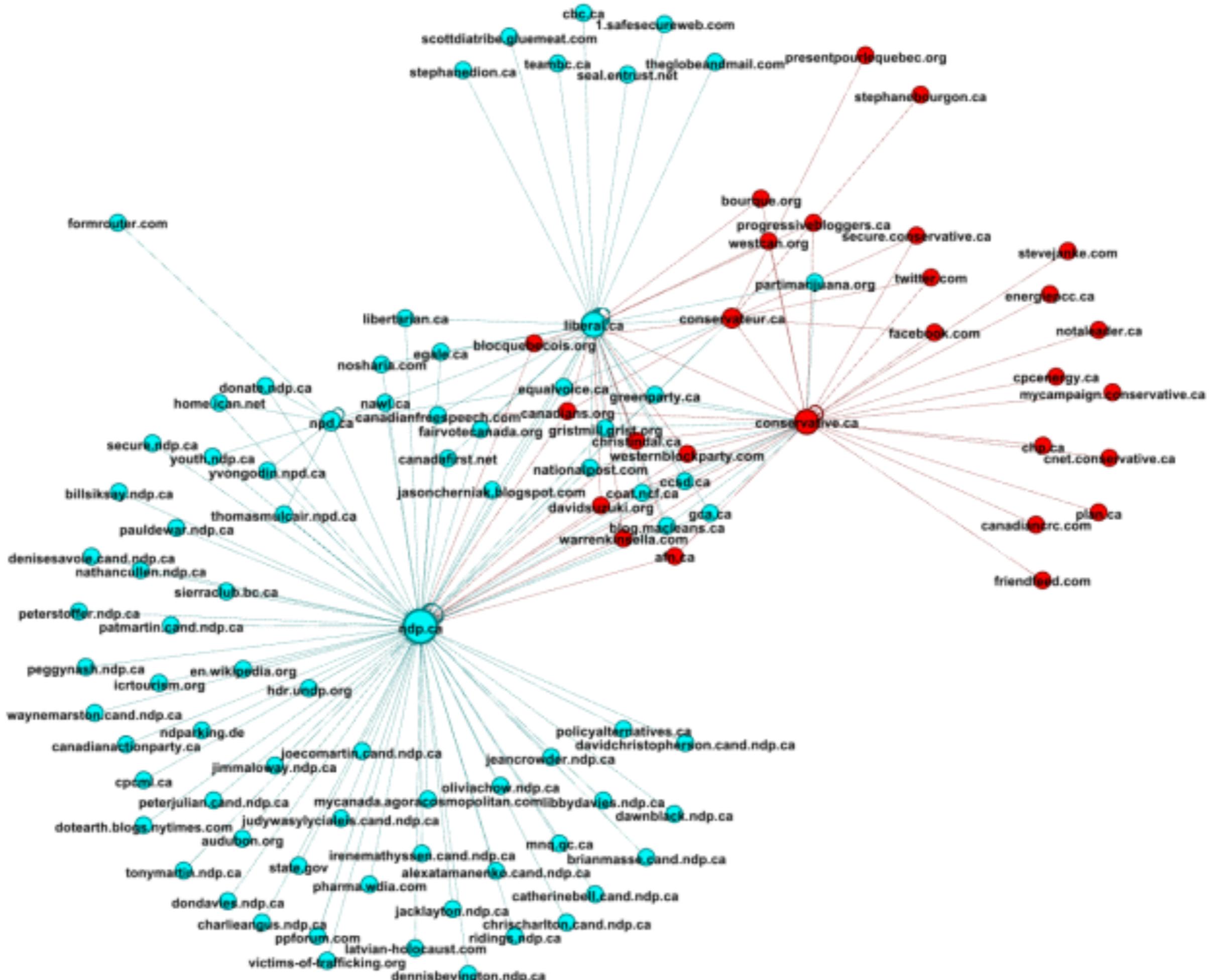
## Election Campaign - Advertisement Sites



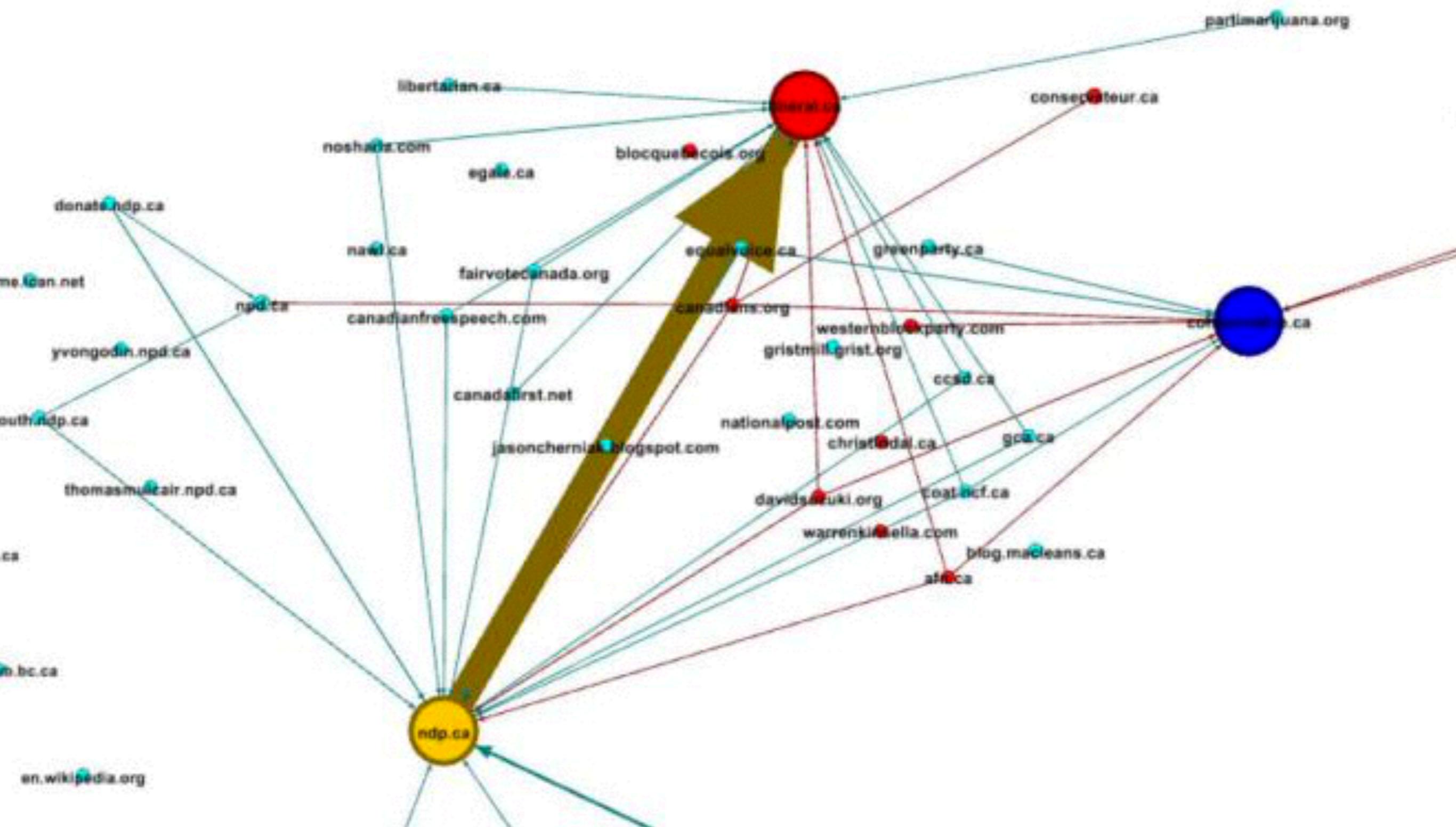
December 2008

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





# 2005 Canadian Federal Election



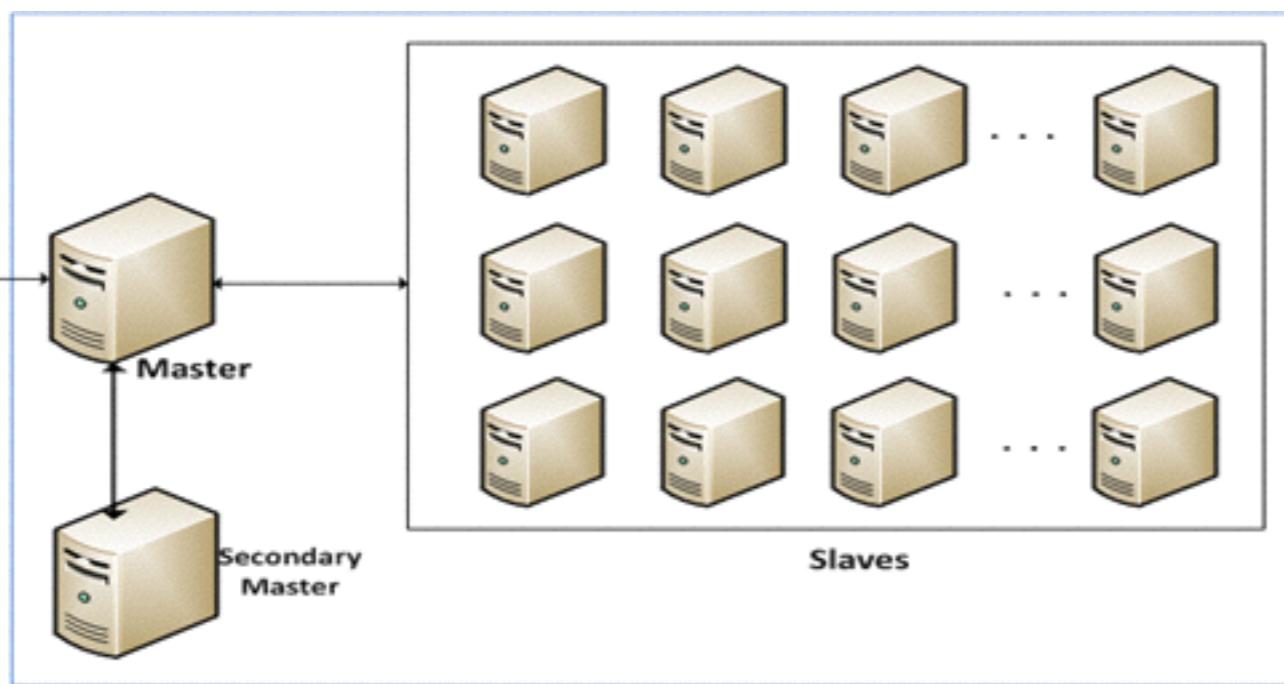


WATs help us find the files  
we need to use - and to  
contextualize them



## Problem Two:

You can do amazing things with the content (WARCs), but you need a cluster or powerful computer.



Hadoop Cluster

# WARC Analysis

- **2005-2009:** 244 GB of content; 2.9 GB of plain text
- 10,606,822 websites
- On a local powerful node (3 Ghz 8-Core Intel Xeon E5/64 GB RAM, data on SSD), about three to four hours per query
- On a cluster, about ~10-20 minutes per query, depending on traffic



# Large-Scale Text Analysis

- With Hadoop about 15-20 minutes to extract all plain-text from any specified queries:  
i.e. all pages belonging to Green Party, Liberal Party, Conservative Party, Council of Canadians, etc.
- Compared to “out of memory”/ go home for an extended weekend on a local node

The screenshot shows a Mac OS X terminal window with several tabs open. The tabs are labeled 'java', 'java', 'ssh', and 'ssh'. The content of the tabs is mostly illegible due to the small font size, but it appears to be command-line output or log files related to Java and SSH operations. Below the tabs, there is a file viewer window titled '200703.txt' which displays a large amount of text. The text is a collection of web page snippets, likely crawled by a search engine or爬虫 (spider). The snippets are in French and English, mentioning political parties like 'Parti Vert du Canada', 'Green Party', and 'Liberal Party'. They also mention various political figures and events, such as Elizabeth May's campaign, the Canadian election of 2008, and the war in Afghanistan. The text is heavily redacted with black bars, obscuring many of the specific details.

# Large-Scale Text Analysis

- **NER/LDA/Keyword Frequency broken down by scrape date:** i.e. scrape carried out 2005-10, see change over time;
  - Downside: not everything is optimized for parallel environment; if not, it crawls (there goes a day)
  - Downside: scrape date != creation date, requiring temporal analysis



# atlanta cooks

125 Recipes From 25 Top Atlanta Chefs

Andrew Friedman and Maria Pollio

A Return to Cooking



LEMONGRASS

charlie Trotter's Chicago Kitchen

EFFORTLESS ELEGANCE WITH LEMONGRASS AND LIME

One&Only Palmilla



COLIN COWIE

Charlie Trotter's Desserts



SPATULACUISINE by Charlie Trotter

Cuba Cocina

SAM BEALL

THE BLACKBERRY FARM COOKBOOK



TEN SPEED

PRESS

RED SAGE

MICHAEL MINA'S MEAT & COOKIES



TEN SPEED PRESS

CAKES AND MAZZOLA

Boulevard



TEN SPEED

My Beverly Hills Kitchen

ALICE HITZ

FRANK STITT'S BOTTEGA FAVORITA

michel richard happy in the kitchen

Recipe book:

[https://github.com/lintool/  
warcbase/wiki](https://github.com/lintool/warcbase/wiki)

# NER

October 2005

62476 Stephen Harper

30234 Michael Chong

30109 Gwynne Dyer

28011 ami Entrez

26238 Paul Martin

22303 Harper

# NER

November 2008

3188 Stéphane Dion

2557 Stephen Harper

2471 Stephen HarperLaureen

2410 Dion

2356 Harper

# Visualizing Interface Next Step?

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

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  
Secondo DiGangi

Leonardo DiCaprio  
Manley  
Mark Brooks  
Mark MacGillivray  
Martin  
Michael Robinson  
Milliken  
Paul Martin  
Pete Marin

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

David Chernushenko  
David Chernushenko E...  
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  
Adrianne Carr  
Allan Gribbin  
Amélie Gingras  
Andrew Lewis  
Bill  
Bill Hulet  
Brown  
Bruce Abel  
Bush  
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

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 Chernush  
David Kay  
Derek Binto

O

Derek Pinto  
Dietrich  
Dion  
Elizabeth

# Elizabeth May

Elizabeth May  
0 mentions

Elizabeth Peloza  
Eric Walton  
Gomery  
Green  
Harper  
**Harris**  
Jasper  
Jim

## Jim Harris

Jim Harris Speech  
John  
Labchuk  
Lougheed  
Mackenzie  
Mapley

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  
Cherfl  
Chernushenko  
Daphne Wysham  
David  
David Chernushenko  
David Kay  
Derek  
Derek Pinto

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

Laibchuk

Manley

Mark MacGillivray

Martin  
May  
Melanie Ransom  
Michael Grayson  
Michele  
Paul Martin  
Richard Reble  
Sharon Labchuk  
Stuart Weller

200611  
Ambrose  
Andrew Lewis  
Bill  
Bill Clinton  
Bush  
Chandler  
Cherfi  
Chernushenko  
Chris Alders  
Daphne Wysham  
David  
David Chernushenko  
**David Cox**  
David Kay  
David Suzuki  
Dawda

Derek  
Derek Pinto  
Dundas

**Edward Burtynsky**

Elizabeth

**Elizabeth May**

Eric Walton  
Garth Turner  
Gormery  
Green

**Halton**

Harper  
Harris

Jim

**Jim Harris**

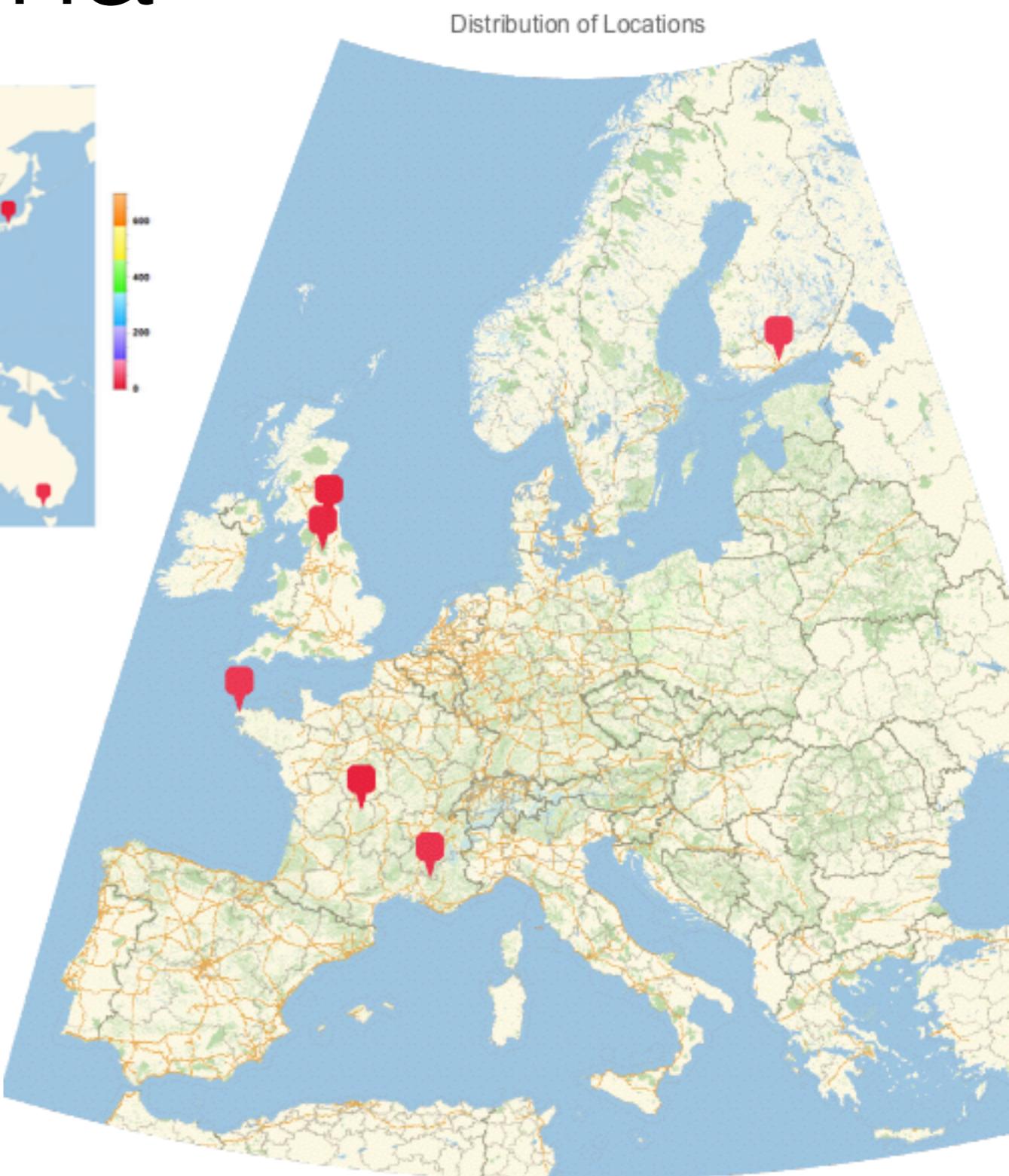
Jim Harris Speech  
John  
Julie Baribeau  
Labchuk  
Manley

Margaret  
Mark MacGillivray  
Martin  
May  
Paul  
Paul Martin  
Ross  
Sharon Labchuk

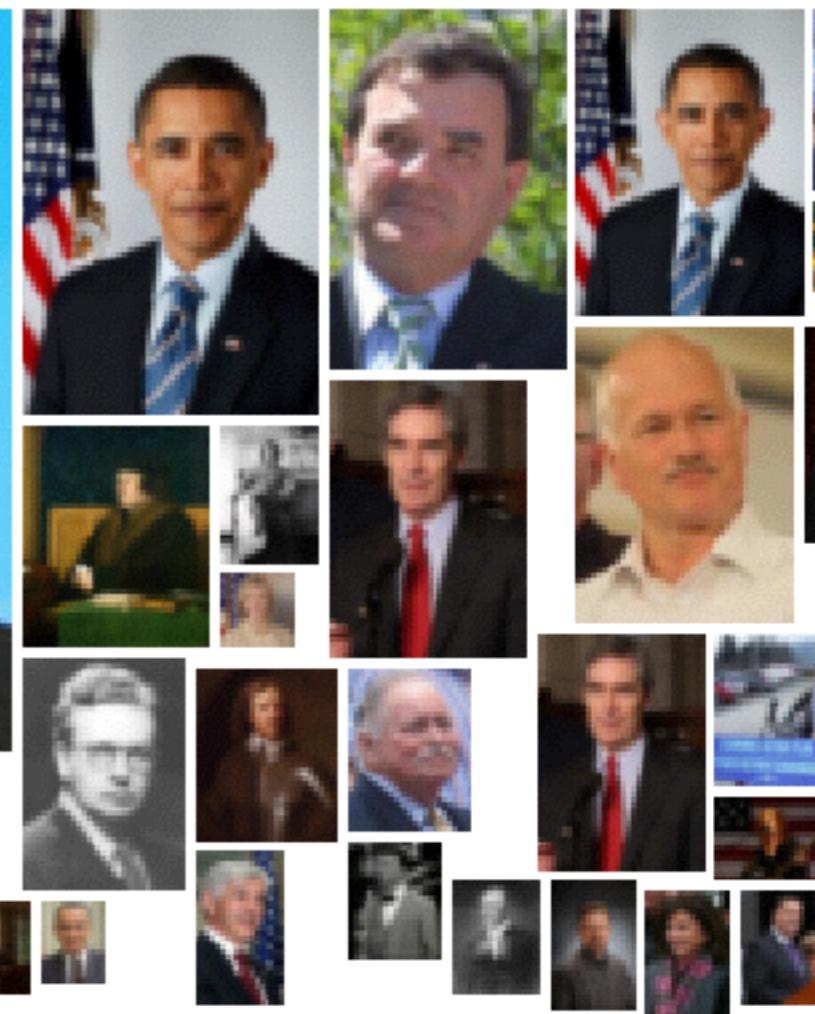
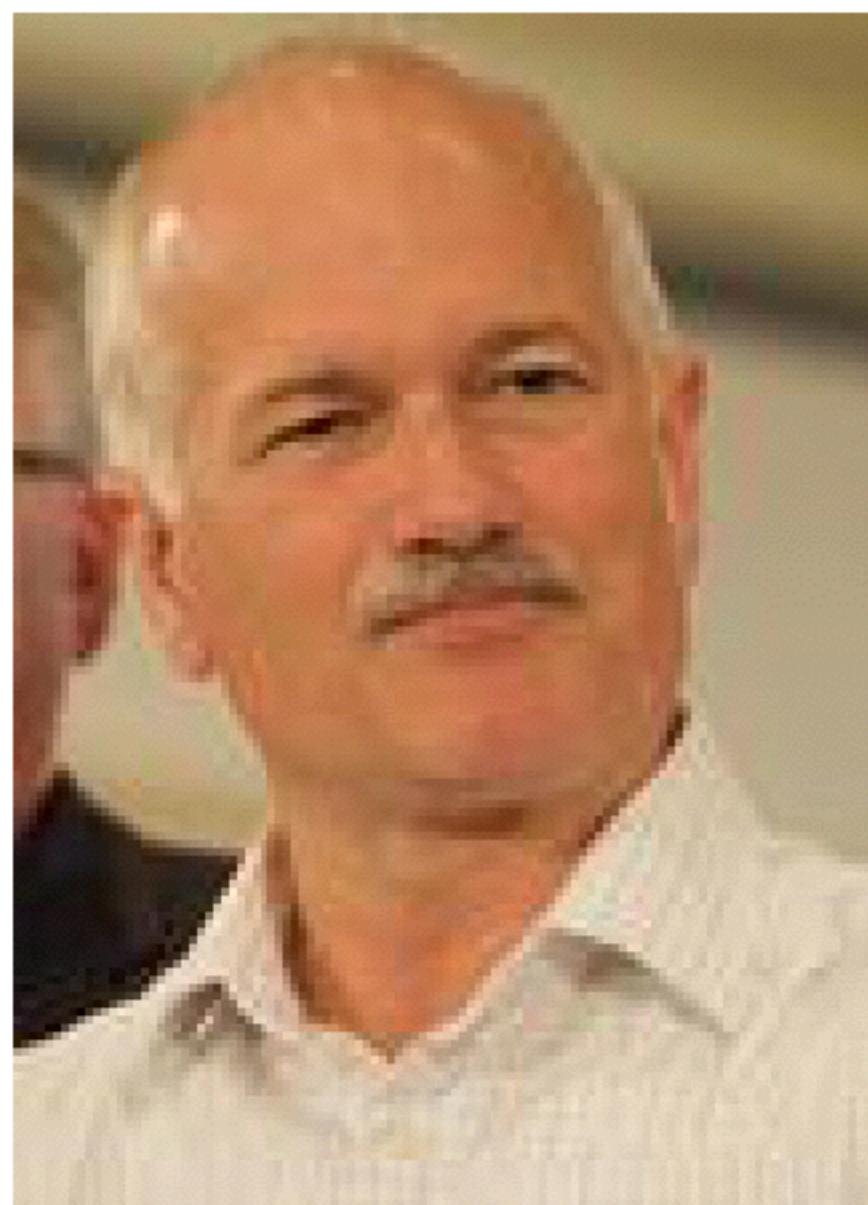
# Integration with Wolfram|Alpha



```
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,
```



# Integration with Wolfram| Alpha



# Shine/WebArchives.ca

- UK Web Archive's Shine (<https://github.com/ukwa/shine>)
- Indexing as bottleneck
  - ~ 250GB of WARCs takes ~ 5 days on a single machine
  - Hadoop indexer available if data in HFDS
- ~ 90GB index size



The screenshot shows a web browser window titled "The Canadian Political Party". The address bar contains "webarchives.ca". The main content area has a blue header bar with the text "Web Archives for Historical Research - Canadian Politics". Below this, a yellow box contains the text: "Welcome to the Web Archives for Historical Research political parties portal. Before diving in, we encourage you to visit our [about](#) page." To the right of this text, there is a large, bold heading: "The Canadian Political Parties and Political Interest Groups Portal". Below the heading, a paragraph of text reads: "On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015." Further down, another paragraph says: "Curious how the Liberal Party of Canada responded to the 2008 financial crisis ([a search for "recession 2008, liberal.ca"](#))? How the Canadian Centre for Policy Alternatives [reacted to Michael Ignatieff?](#) Now you can check it all out." Below this, a section titled "Options include:" lists three items:

- [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 climate change within 25 words of each other]

At the bottom of the page, a note states: "Below, here are all of the links for the entire time period, visualized below." Below this note, there is a complex network visualization consisting of numerous small, semi-transparent nodes connected by thin lines, forming a dense web-like structure.

# Demo

# Five Things I've 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

# Shine

- **Advantages:** accessible to the general public, easy to use, interactive trend diagram allows digging down for context, can move down to level of document itself.
- **Disadvantage:** keyword searching requires you know what to look for; random sampling misleading when tens of thousands of records; etc.
- Doesn't take advantage of what makes web sources so powerful: hyperlinks

Building connections  
between Warcbase and  
Shine

End-user tools and co-operation with CS colleagues is key.

The screenshot shows a GitHub repository page for 'lintool/warcbase'. The repository has 449 commits, 4 branches, and 0 releases. The master branch is selected. A list of recent commits includes:

- .settings: Tweaked settings.
- src: Added option to change MAX\_CONTENT\_SIZE in IngestFiles, Issues #112
- .gitignore: Added .iml files
- README.md: Error in README
- pom.xml: Updated versions of some artifacts.

The page also features a 'Warbase' section with a brief description and a 'Getting Started' section.

Warcbase is an open-source platform for managing web archives built on Hadoop and HBase. The platform provides a flexible data model for storing and managing raw content as well as extracted knowledge. Tight integration with Hadoop provides powerful tools for analysis and data processing.

**Getting Started**

Clone the repo:

But the shared  
promise...



**More voices, more  
people, the promise of  
social history achieved.**

# Thank you!

**@ianmilligan1**  
**ianmilligan1@gmail.com**

---

**Ian Milligan**  
**Assistant Professor**



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