

# **Understanding early Web history through three case studies**

**Methodological and technical  
challenges**

---

**Ian Milligan**  
Assistant Professor  
@ianmilligan1



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

# Why?

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



MATCH YOUR INTEREST TO A NEIGHBOR

FREE HOME PAGES AND E-MAIL	ARTS AUTOS BUSINESS COMPUTERS CULTURE	EDUCATION ENTERTAINMENT ENVIRONMENT FAMILY FASHION	FOOD GAMES GAY & LESBIAN GOVERNMENT HEALTH	KIDS MUSIC PEOPLE RECREATION SCIENCE FI
--	---	--	--	---



Your home on the web!

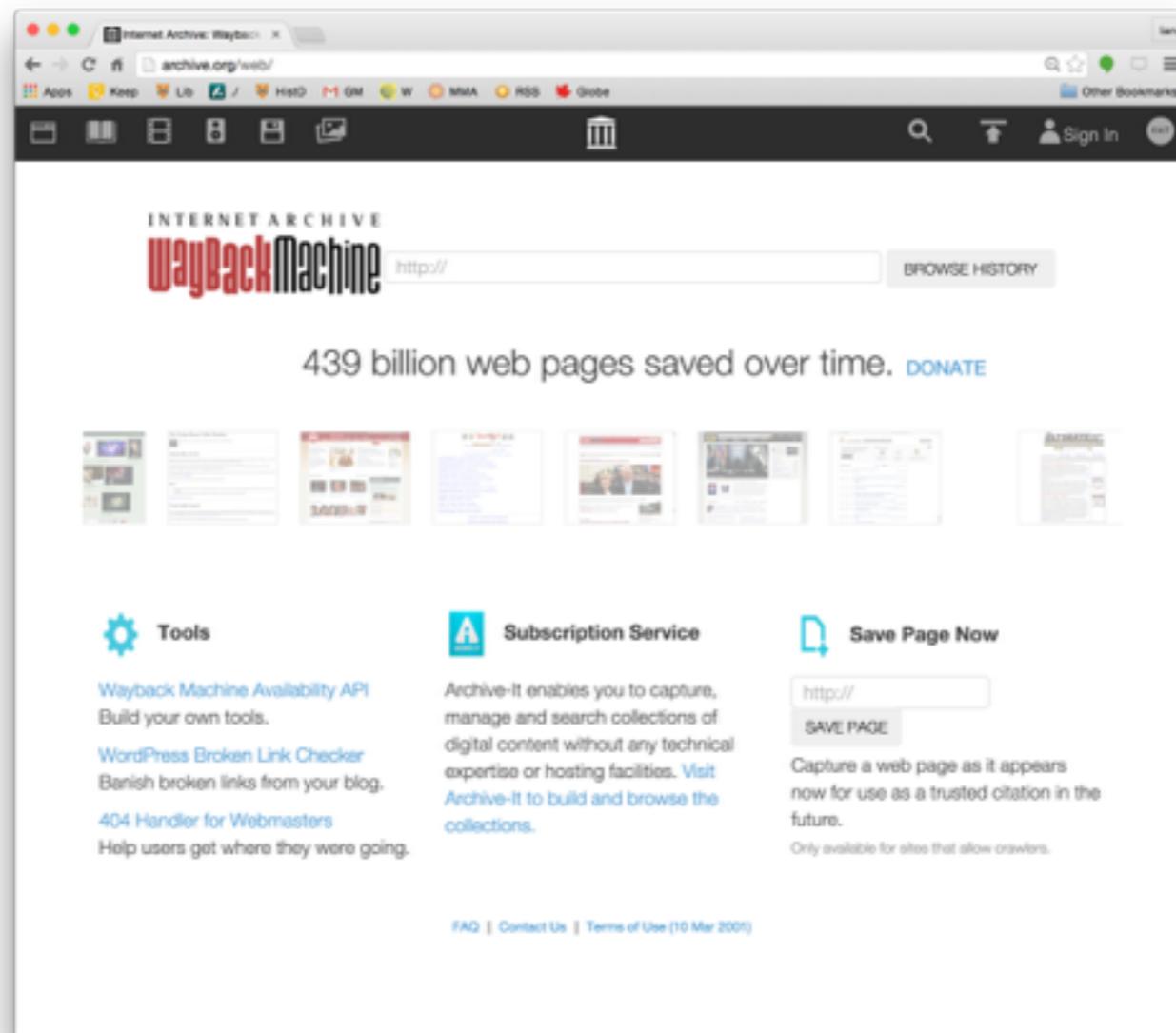


1990s

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



# Nightmare Scenario



This won't be enough!



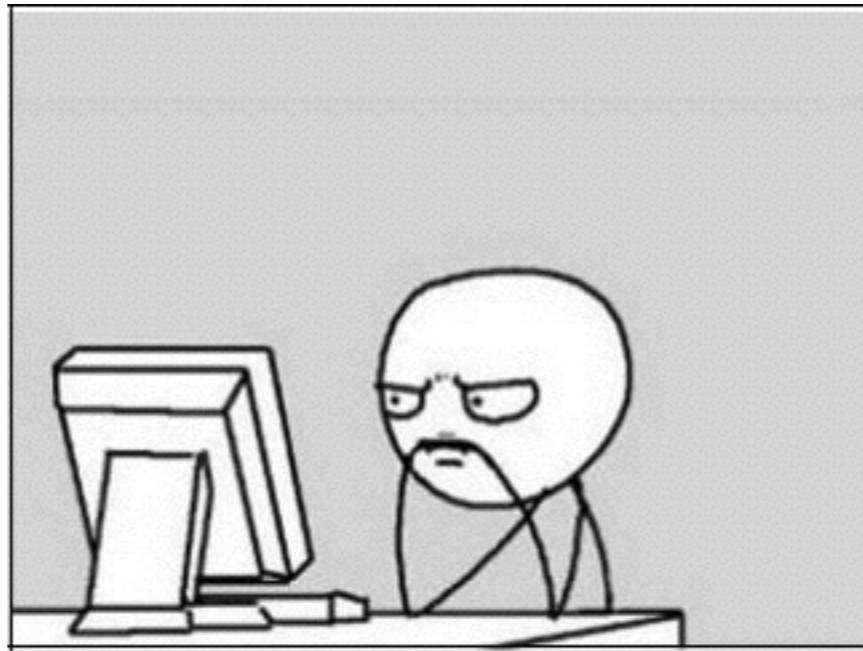
... but what will our  
search engines look like?

# But what will web archives look like?

- Three Distinct Case Studies
  - **Wide Web Scrape**, March - December 2011 (Internet Archive) (sample of 80TB WARC collection);
  - **Canadian Political Parties & Interest Groups**, 2005-2015 (Archive-It/University of Toronto)
  - **GeoCities End-of-Life Torrent**, 2009 (Archive Team);

**Similarities -**  
Windows into the lives of  
everyday people.





**Differences -**  
Incredible range of technical  
skills/no common platform!

# Case Study One

- A handy introduction to WARCs and CDXs
- The **Wide Web Scrape** (~ 80TB)
- **85,570** WARC files, CDX metadata

The screenshot shows a web browser window with the URL <https://archive.org/details/wide00002&tab=about>. The page title is "Wide Crawl started March 2011". The page content includes a description of the crawl, statistics, and a sidebar with contributor information.

**DESCRIPTION**

Web wide crawl with initial seedlist and crawler configuration from March 2011. This uses the new HQ software for distributed crawling by Kenji Nagahashi.

**What's in the data set:**

- Crawl start date: 09 March, 2011
- Crawl end date: 23 December, 2011
- Number of captures: 2,713,676,341
- Number of unique URLs: 2,273,840,159
- Number of hosts: 29,032,069

The seed list for this crawl was a list of Alexa's top 1 million web sites, retrieved close to the crawl start date. We used Heritrix (3.1.1-SNAPSHOT) crawler software and respected robots.txt directives. The scope of the crawl was not limited except for a few manually excluded sites.

Created on October 5 2010

**ARossi**  
Archivist

**ADDITIONAL CONTRIBUTOR**

**brewster**  
Archivist

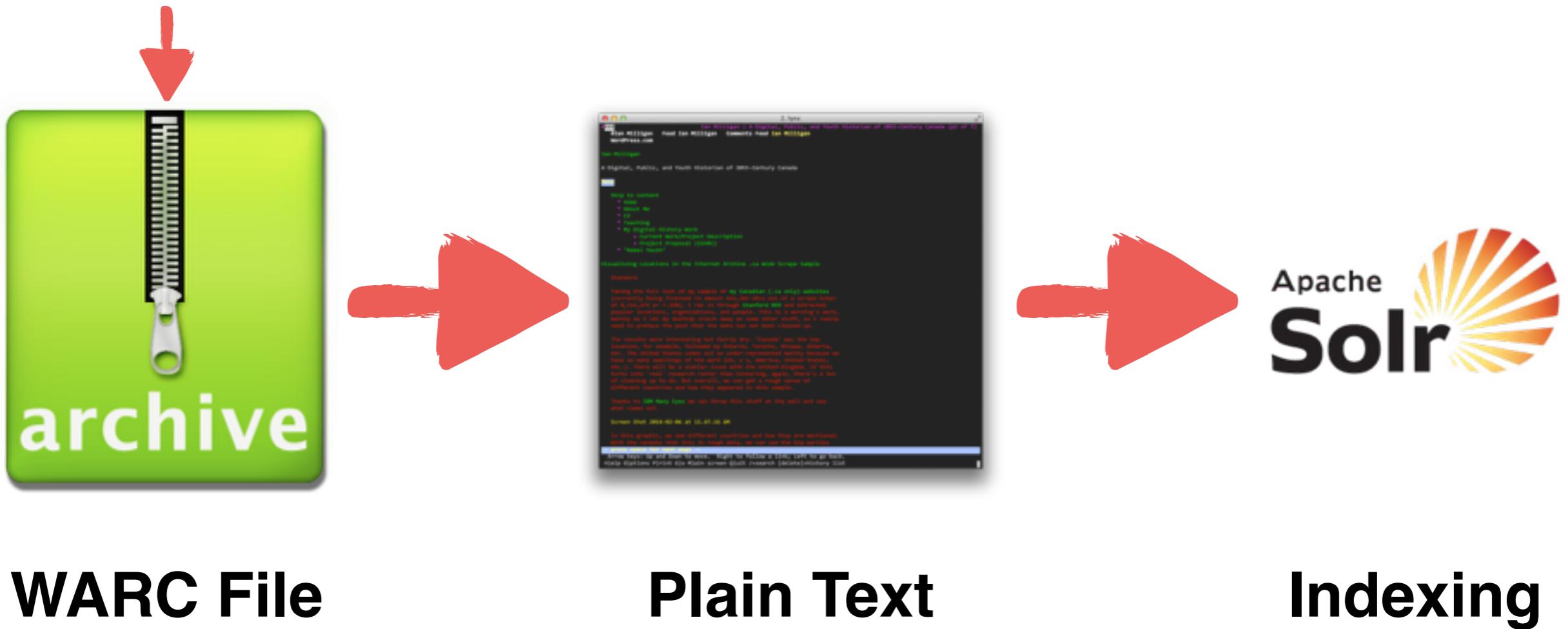
**kngenie**  
Archivist

VIEWS

ca,yorku,justlabour)/ 20110714073726  
<http://www.justlabour.yorku.ca/> text/html  
302 3I42H3S6NNFQ2MSVX7XZKYAYSCX5QBYJ  
[http://www.justlabour.yorku.ca/index.php?  
page=toc&volume=16](http://www.justlabour.yorku.ca/index.php?page=toc&volume=16) - 462 880654831  
WIDE-20110714062831-crawl416/  
WIDE-20110714070859-02373.warc.gz

Top-Level Domain	Number of Distinct URLs Downloaded in Sample	Number of Overall URLs in Wide Web Scrape (selected domains)	Percentage of URLs Captured
.com	29,219,706	1,260,409,874	2.32%
.org	2,489,050	96,681,268	2.57%
.net	2,438,903	140,726,805	1.73%
.edu	350,482	6,620,283	5.29%
.gov	97,484	2,205,332	4.42%
.mil	10,268	103,507	9.92%
.ca	622,365	8,512,275	7.31%
.uk	464,991	21,870,821	2.13%
.fr	239,160	13,654,404	1.75%
.in	105,287	3,736,316	2.82%
.cn	5,499,593	133,105,864	4.13%
.ke	4883	37,871	12.89%
<b>TOTAL</b>	<b>41,542,172</b>	<b>1,687,664,620</b>	<b>2.46%</b>

# CDX Files (finding aids)



Carrot2 Workbench

Source: Solr  
Algorithm: Lingo

Basic

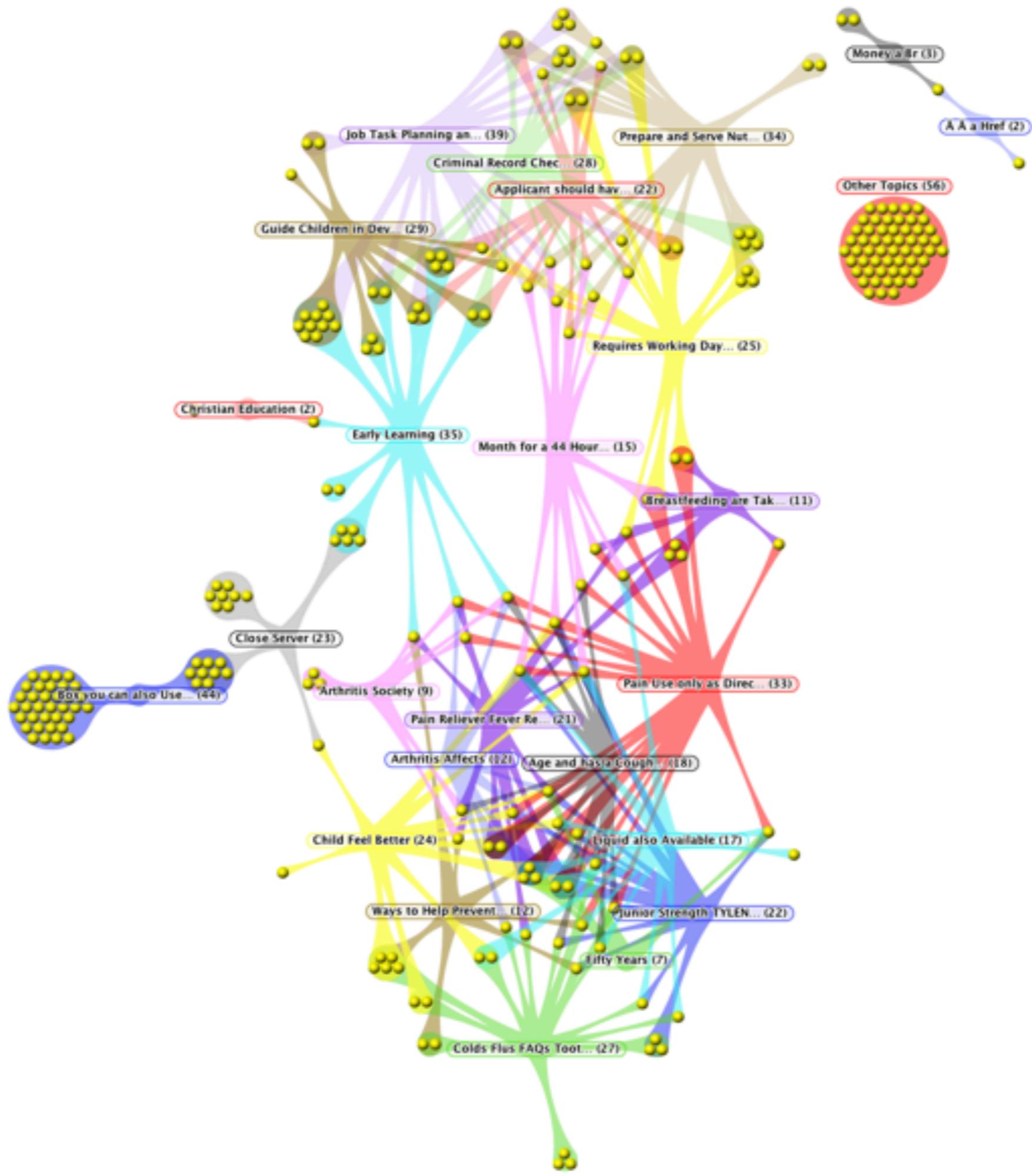
Query (required): Q: children  
 Read Solr clusters if present

Results: 1000

Aduna Cluster Map Visualization Circles Visualization FoamTree Visualization

http://www.tylenol.ca/children/children-6-11-years/cough-cold-flu/products  
text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Francais Search \_\_\_\_\_ Search \* Adult \* Children \* P...
- [1] <http://www.tylenol.ca/children/children-6-11-years/children-6-11-years>  
text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Francais Search \_\_\_\_\_ Search \* Adult \* Children \* Produ...
- [1] <http://www.tylenol.ca/children/children-3-5-years/children-3-5-years>  
text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Francais Search \_\_\_\_\_ Search \* Adult \* Children \* Product...
- [1] <http://www.tylenol.ca/children/children-products>  
text/html; charset=utf-8 For Adults For Children Tylenol logo Home | Contact us | Francais Search \_\_\_\_\_ Search \* Adult \* Children \* Products \* About Tylenol \* News & Promotions All Children's Pro...
- [1] <http://blogs.afortunecookie.ca/tag/children/fever/>  
text/html  
[1] <http://www.tylenol.ca/children/children-6-11-years/aches-pains/about-aches-pain>  
text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Francais Search \_\_\_\_\_ Search \* Adult \* Children...
- [1] <http://www.tylenol.ca/children/children-3-5-years/aches-pains/relieving-your-child-s-aches-pains>  
text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Francais Search \_\_\_\_\_ Search ...
- [1] <http://www.tylenol.ca/children/children-6-11-years/aches-pains/relieving-your-child-s-aches-pains>  
text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Francais Search \_\_\_\_\_ Search ...
- [1] <http://www.tylenol.ca/children/children-6-11-years/cough-cold-flu/relieving-your-child-s-coughcold-flu-symptoms>  
text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Francais Search \_\_\_\_\_ Search ...
- [1] <http://www.tylenol.ca/children/children-6-11-years/aches-pains/relieving-your-child-s-aches-pains>  
text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Francais Search \_\_\_\_\_ Search ...

131MB of 4094MB



children (250 documents from Solr, 26 clusters from Lingo)

**Clusters**

- Box you can also Use it Program
- Job Task Planning and Organizati
- Early Learning (35)
- Prepare and Serve Nutritious Me
- Pain Use only as Directed (33)

**Documents**

- [190] <http://www.lutheranchurch.ca/missions.php?s=nicaragua&p=6&print=yes>

**Services**

- Open Link
- Open Link in New Window
- Download Linked File
- Copy Link

**WaybackMachine**

New TextWrangler Document with Selection

EasyFind: Find Selection...

Add to iTunes as a Spoken Track

Open URL

Add to Reading List

- Age and has a Cough or Cold (1)
- Liquid also Available (17)
- Month for a 44 Hour Week (15)
- Arthritis Affects (12)
- Ways to Help Prevent Earaches (
- Breastfeeding are Taking (11)
- Arthritis Society (9)
- Fifty Years (7)
- Money a Br (3)
- Christian Education (2)
- Ã Ä a Href (2)
- Other Topics (56)

Lutheran Church-Canada

web.archive.org/web/20110714130258/http://www.lutheranchurch.ca/news.php?id=158&print=yes

INTERNET ARCHIVE Wayback Machine 3 captures 5 Dec 10 – 14 Jul 11 DEC JUL 14 2010

LUTHERAN CHURCH-CANADA ÉGLISE LUTHÉRIENNE du CANADA

CLWR funds Nicaraguan medical and dental clinic, scholarships

Friday, January 22, 2010

WINNIPEG – Canadian Lutheran World Relief (CLWR) has announced \$36,500 in funding for two Lutheran Church-Canada (LCC) programs in Nicaragua this year.

The announcement was made as Iglesia Luterana Sinodo de Nicaragua (ILSN) prepares for its first biennial convention and includes new money for a medical and dental clinic and increased school scholarships.

The medical clinic, which began operations in May 2009, is open every Thursday beginning at 8 a.m. and remains open until all patients have been seen.

The clinic is staffed by a doctor and a dentist, who see an average of 40-45 patients each week, and provides common medications because many patients are too poor to purchase them.

CLWR will continue supporting the Christian Children Education Program. The program, conducted in all 23 congregations of ILSN, provides an average of 25 scholarships in each community to the neediest children. The scholarships include the required school uniforms, shoes, backpacks and school supplies.

Each child is also enrolled in the tutoring and Christian-education class held five days a week when children are not in school (Children after school in the morning or in the afternoon).

These classes, held in the churches and led by teachers and deaconesses, provide tutoring and homework support for the children in math, Spanish and other subjects. A portion of the time is also set aside for Christian education and cultural activities.

More than 750 children are enrolled in the program. CLWR has provided support for about 250 children.

Since 1999, CLWR has partnered with LCC to support community-development projects.

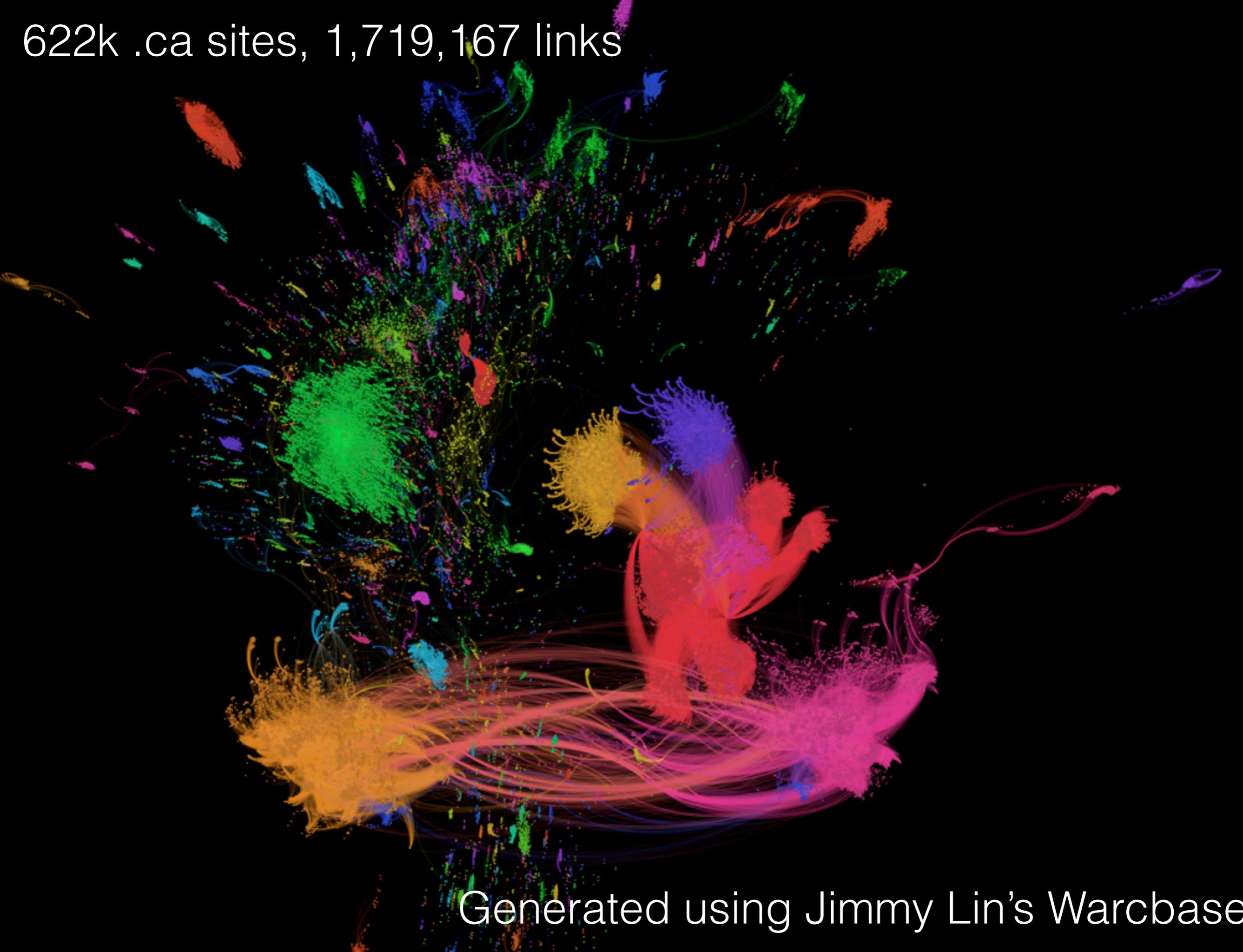
Robert Granke, executive director of CLWR, visited congregations of the ILSN in November. You can read more about his visit at [www.Iccontheroad.ca](http://www.Iccontheroad.ca). The Canadian Lutheran or in the forthcoming issue of CLWR's Partnership newsletter due out in early February.



A medical clinic in Nicaragua.

Problem is.. you need to  
know what you're looking  
for!

622k .ca sites, 1,719,167 links



Generated using Jimmy Lin's Warcbase

<http://www.jobs-open.ca/>

<http://www.jobs-open.ca/info-about.php>  
<http://www.jobs-open.ca/about-us.php>

mailto:[webmaster@jobs-open.ca](mailto:webmaster@jobs-open.ca)  
<http://www.jobs-open.ca/info-about.php>

<http://nova-scotia.jobs-open.ca/>

<http://www.uottawa.ca/cartes>

<http://www.biblio.uottawa.ca/index-f.php>

<http://www.uottawa.ca/info/cotain>

<http://www.uottawa.ca/bjenvvenue.html>

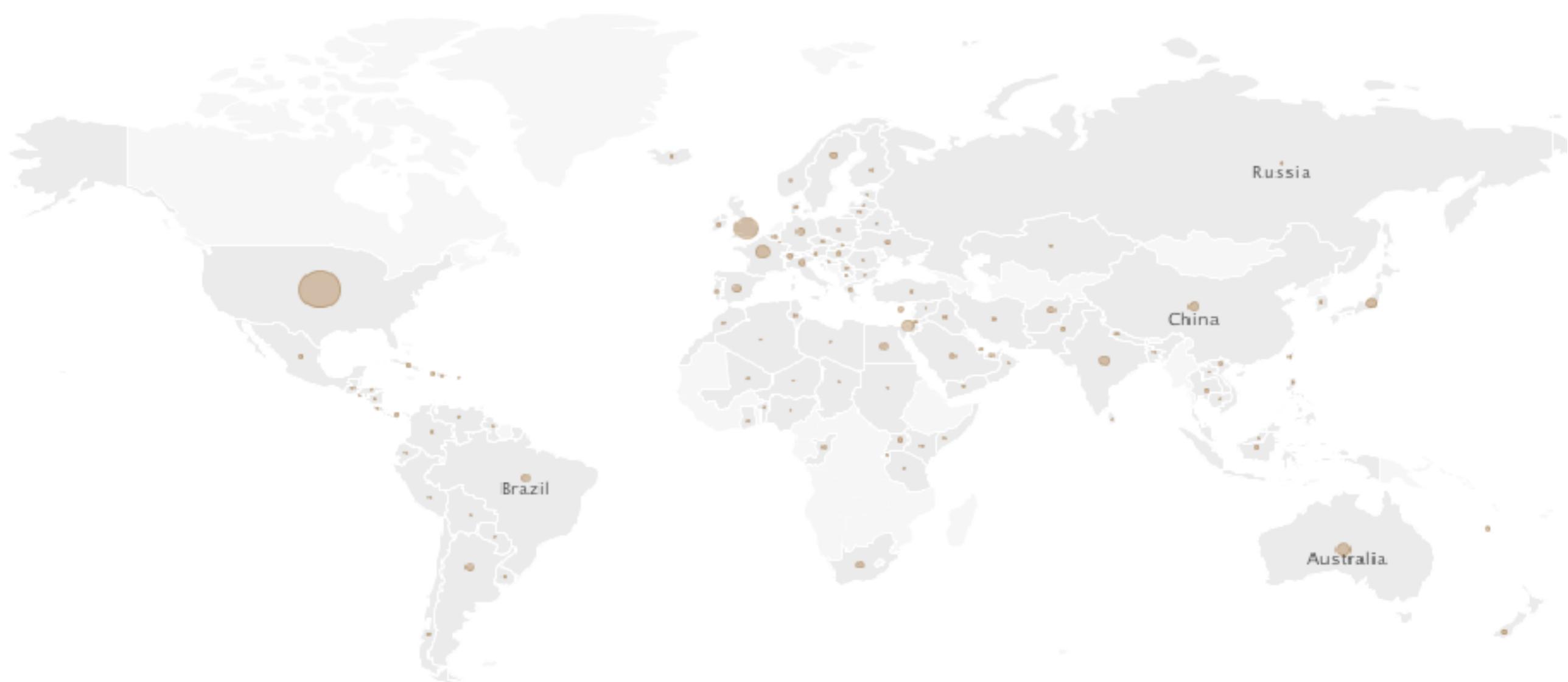
<http://www.ressourcesfinancieres.uottawa.ca/etudiant/payment-university-fees-fr.php>      <http://www.uottawa.ca/academicinfo/registration/programmes-hon-2565>

<http://www.admission.uottawa.ca/Default.aspx?tabid=2548&sewreadd>

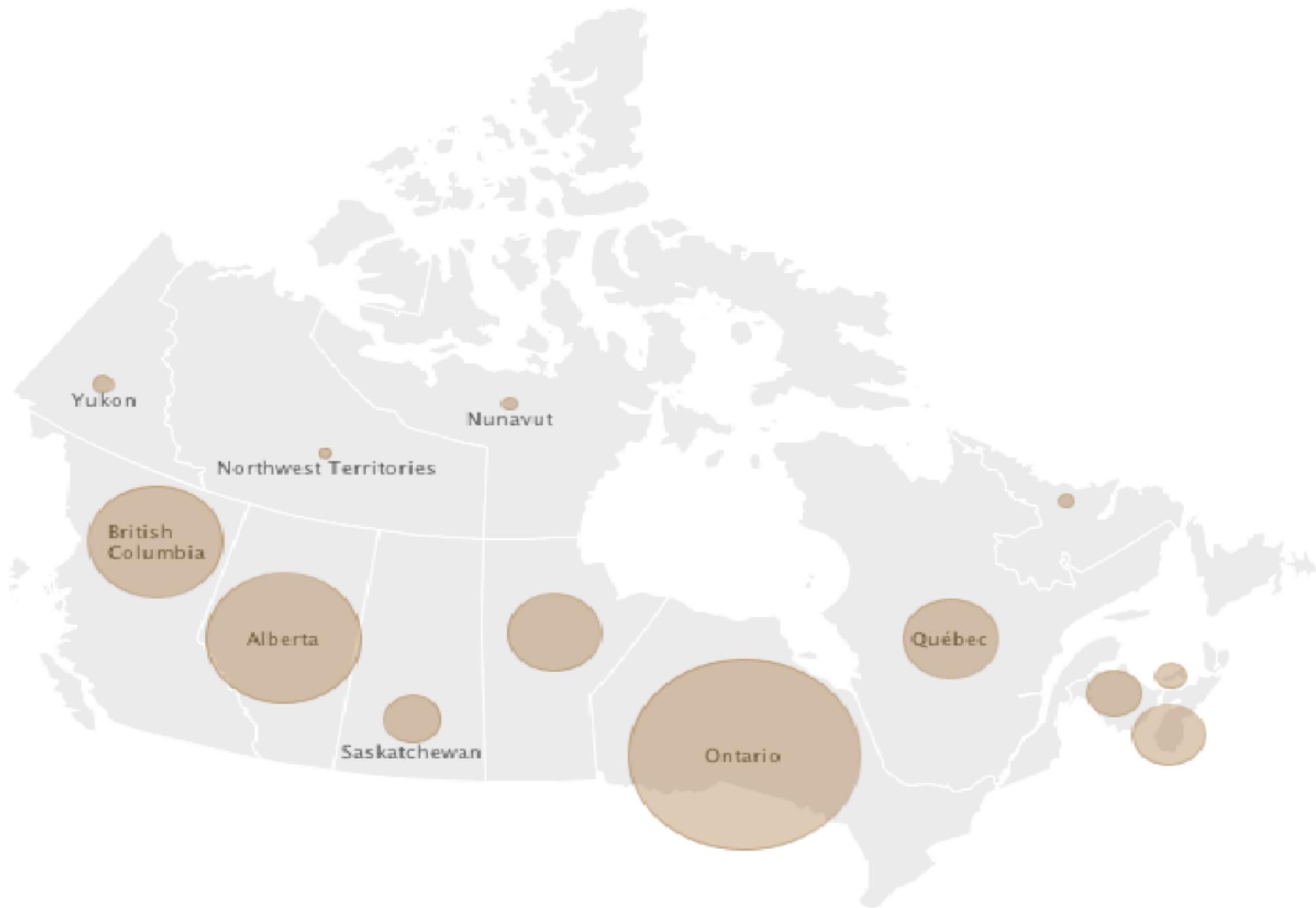
<https://web9.uottawa.ca/services/stores/evaluations/studentguides/francais/section/bourses/>

<http://www.uottawa.ca/icone-recherche/>

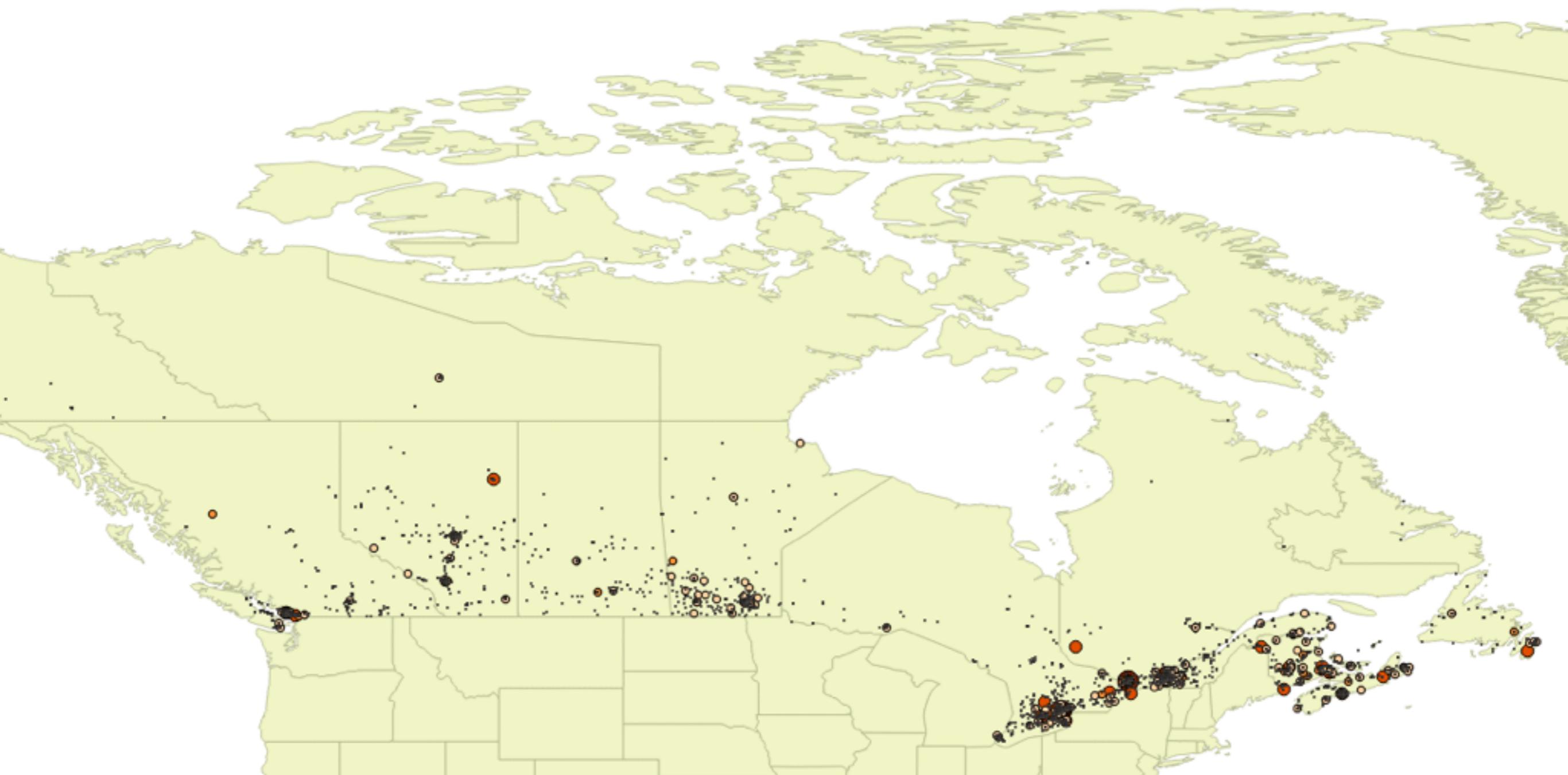
# Countries Mentioned in .ca TLD (excluding Canada)



# Provinces Mentioned in .ca TLD



# Canadian Postal Codes visualized



**Need longitudinal, but the  
size/intensity = extreme.**

# **Wide Web Scraps and** **the Dream of Social** **History.**

# Case Study Two

- **Archive-It Research Services:** “Canadian Political Parties and Political Interest Groups” collection
- 2005 - 2015
- WARC and WAT Files

The screenshot shows a web browser window with the URL <https://archive-it.org/collections/227>. The page title is "Canadian Political Parties and Political Interest Groups". The header includes the Archive-It logo, navigation links for HOME, EXPLORE, LEARN MORE, and CONTACT US, and a tagline "The leading provider of digital preservation services for cultural institutions". Below the header, breadcrumb navigation shows "Explore > University of Toronto > Canadian Political Parties and Political Interest Groups". A large green sidebar on the right contains the collection's logo, the title "Canadian Political Parties and Political Interest Groups", and details: "Collected by: University of Toronto", "Archived since: Oct, 2005", "Description: Canadian Political Parties and Political Interest Groups", "Subject: Politics & Elections", and "Collector: University of Toronto". The main content area features a section titled "Narrow Your Results" with a search bar and a list of subjects: New Democratic Party of Canada (2), Assembly of First Nations (1), Bloc Québécois (1), Canada First (1), and Canada West Foundation (1). There are also buttons for "Sites" and "Search Page Text". At the bottom, it says "Page 1 of 1 (54 Total)" and "Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)".

# 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, a banner for the "Canadian Political Parties and Political Interest Groups" collection is displayed, noting it was collected by the University of Toronto since Oct, 2005. A search bar at the top right contains the query "Stephen Harper". The main content area shows search results for this term, with a total of 1,213,132 matches across 60,657 pages. One result is highlighted: "Stephen Harper | Facebook" with a link to <http://www.facebook.com/pages/Stephen-Harper/9506562109>. The page also includes filters for search terms, file formats, and results per host.

The Canadian Political Party x webarchives.ca

Web Archives for Historical Research - Canadian Politics Search Trends About

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

# The 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.

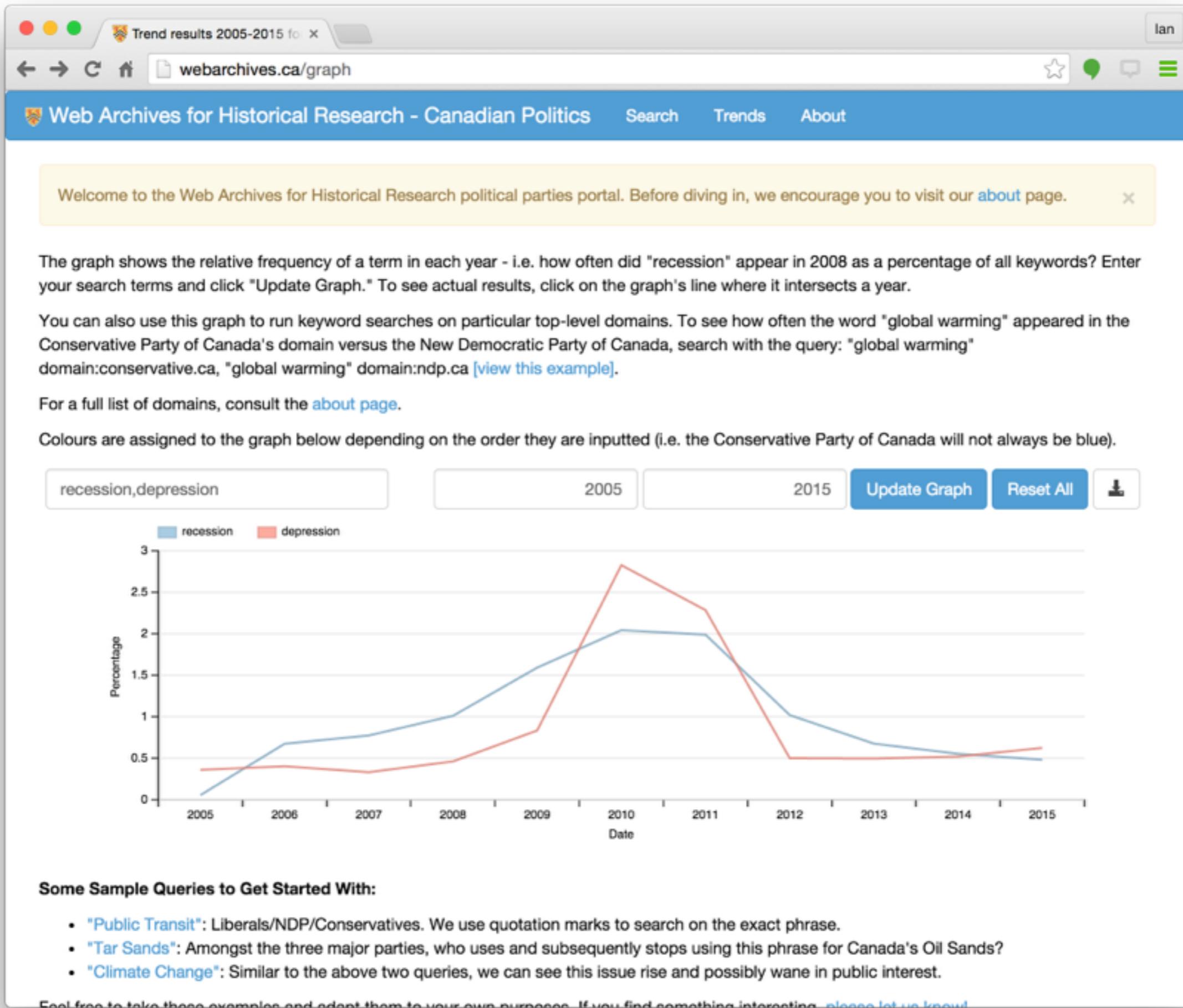
Options include:

- [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.



With Nick Ruest (York University), Nich Worby (University of Toronto), Jimmy Lin (University of Waterloo)





# 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 URL [webarchives.ca/search?query=stephen+harper&tab=results&action=search](http://webarchives.ca/search?query=stephen+harper&tab=results&action=search). The page title is "Web Archives for Historical Research - Canadian Politics". A welcome message encourages users to visit the [about](#) page. Below the message are two search options: "Search" and "Advanced Search".

**General Content Type** (6) [Settings](#)

- html 1,085,201
- other 71,691
- pdf 3,947
- audio 341
- text 106
- image 14

**Crawl Years** (10) [Settings](#)

- 2008 443,448
- 2010 142,609
- 2007 109,236
- 2006 104,564
- 2011 83,910
- 2014 70,740

**Sample Mode** [Search](#) [Reset](#)

Search Term(s): stephen harper

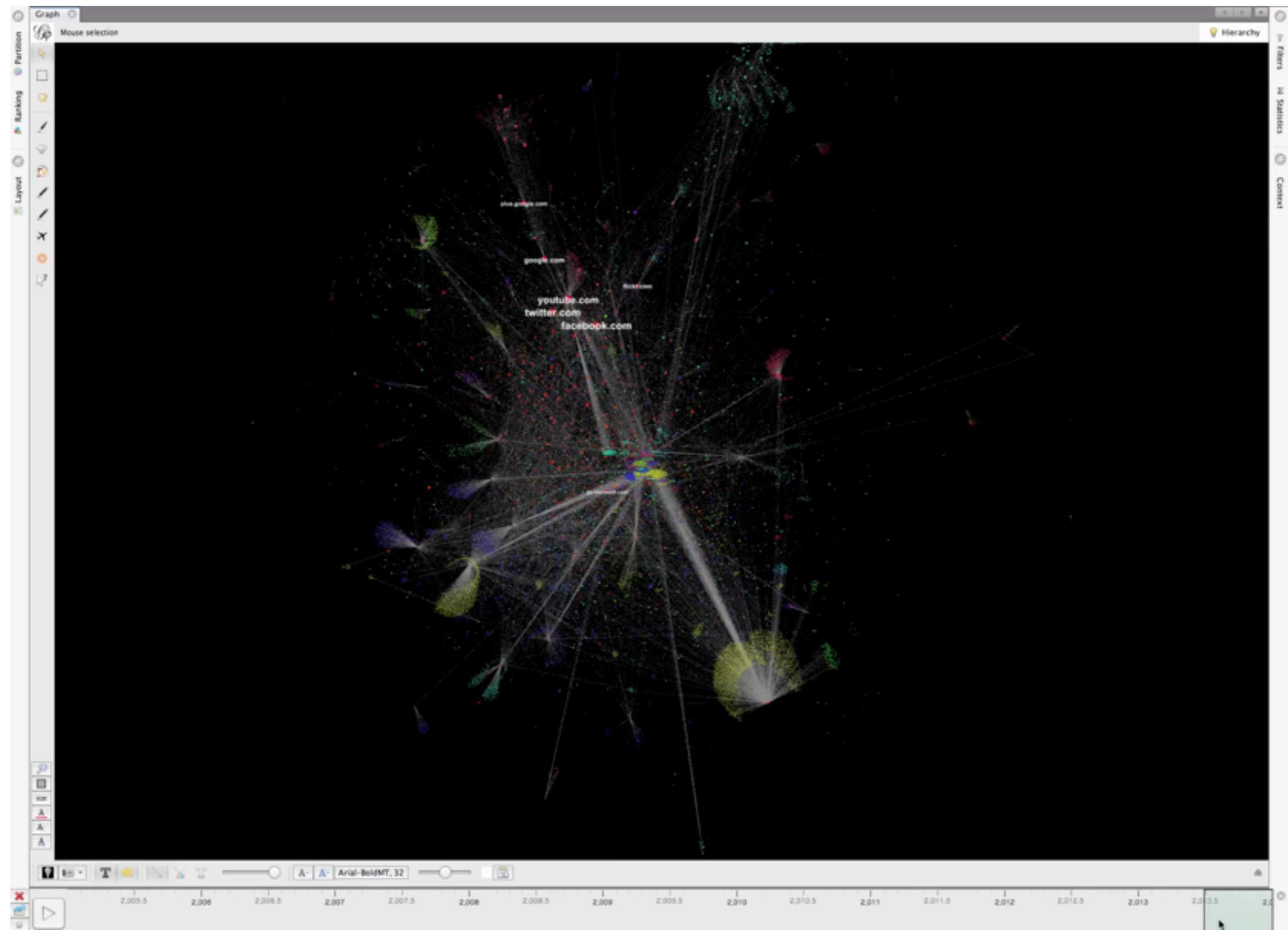
Results Concordance

Results 1 to 10 of 1,161,300 [CSV](#) [Asc](#)

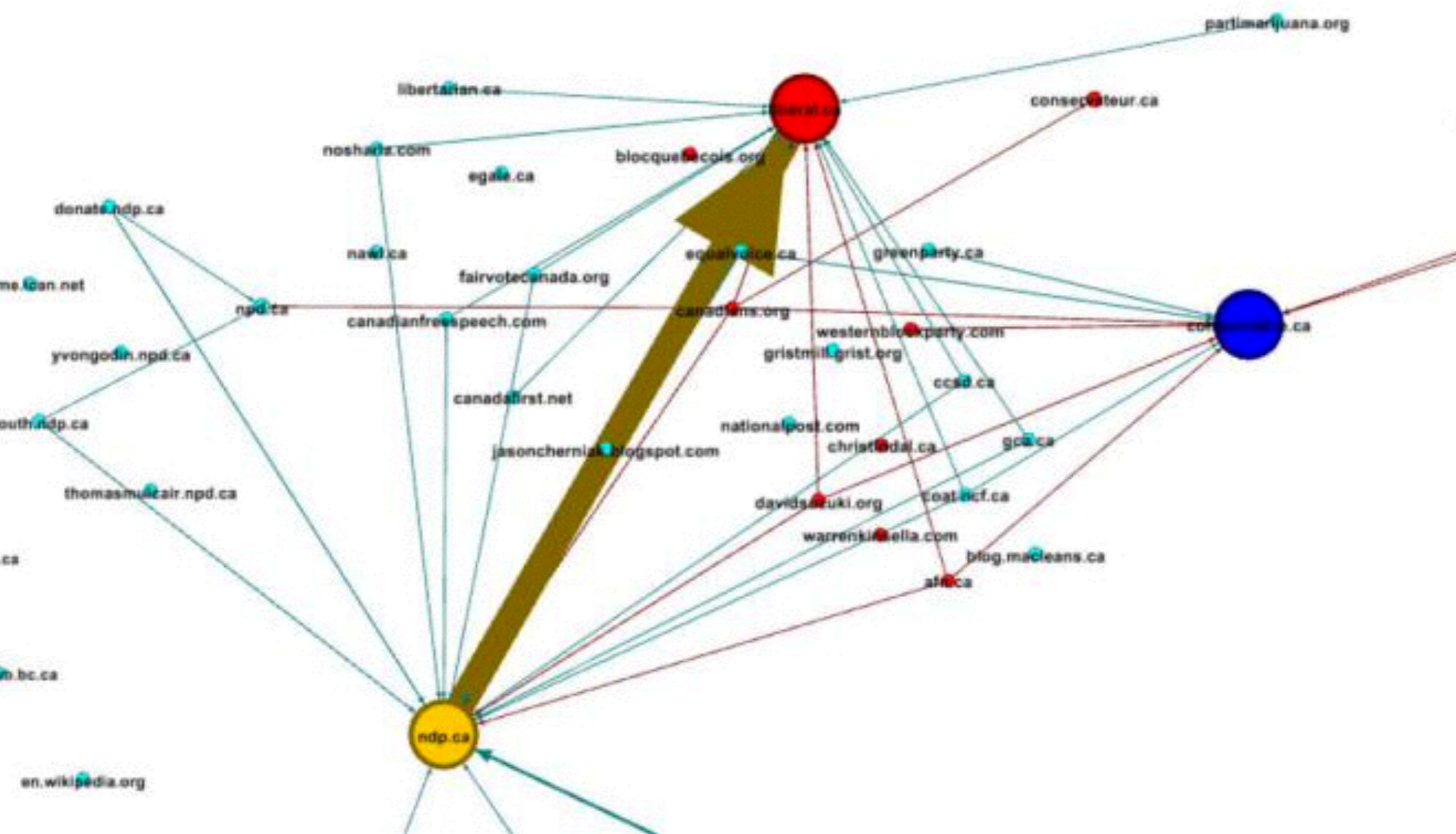
**Do we want metadata  
or content analysis?**

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

# Metadata Extraction



# 2005 Canadian Federal Election



# Case Study Three

- **GeoCities:** Archive Team End-of-Life Torrent
- 2009, content dating back to 1996; can find sites *created* pre-1999 using neighbourhood structure

The screenshot shows a web browser window with the URL <https://archive.org/details/2009-archiveteam-geocities-part1>. The page title is "The Archive Team Geocities". The main content area displays a "The Archive Team Geocities Snapshot (Part 1 of 8) (October 2009)". It includes a brief description of the collection, a media player showing two video files, and a sidebar with links like "Download item", "The Archive Team Geocities Snapshot (Part 1 of 8) (October 2009)", and "GeoCities Yellowpages".

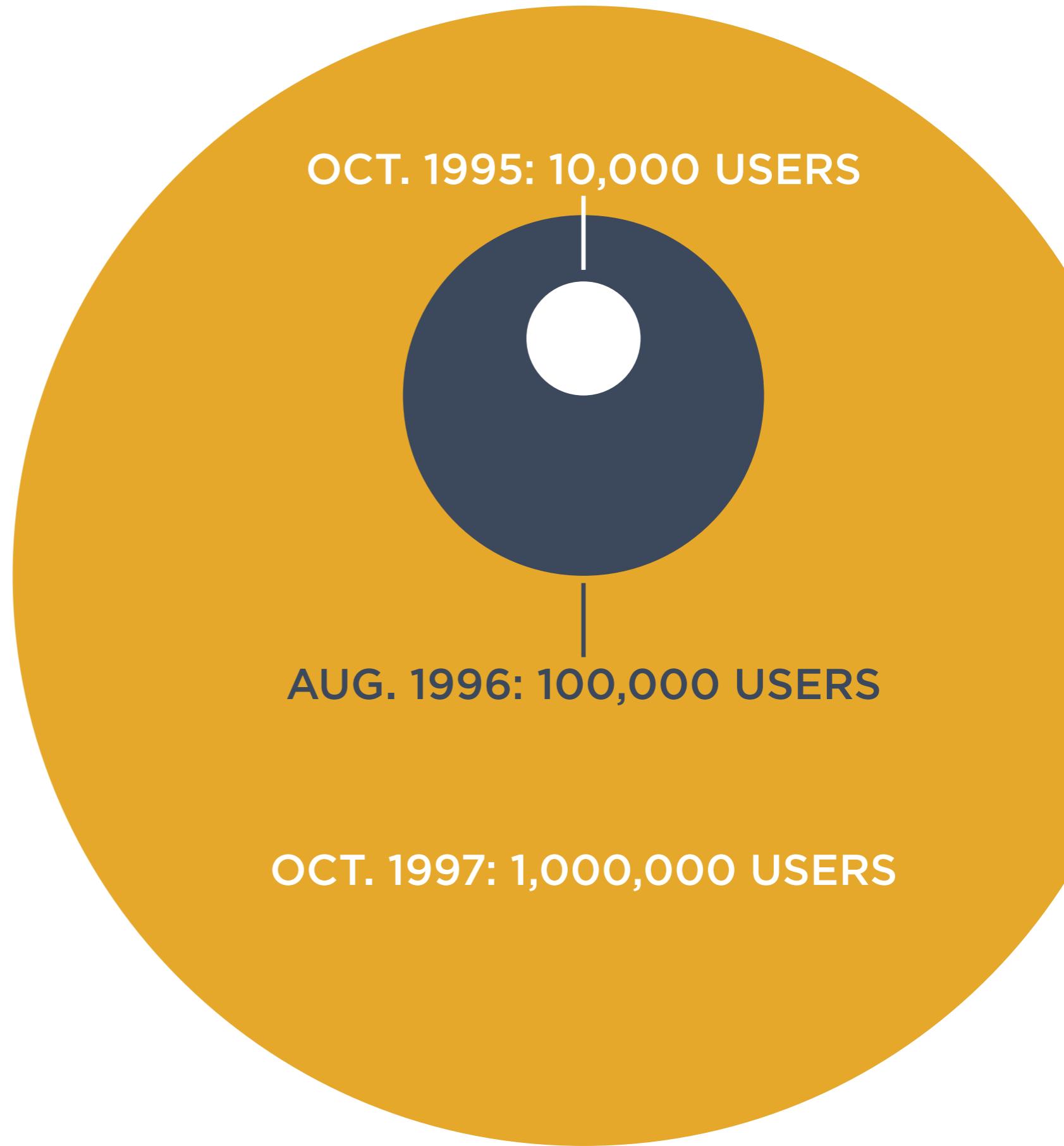
The screenshot shows a web browser window with the URL <https://web.archive.org/web/19961022173245/http://www.geocities.com/>. The page title is "Welcome to GeoCities Home". The main content area displays the "GEO CITIES" homepage from 1996, featuring a red banner with "1,669 captures" and a "TechWire" update. Below the banner, there's a "GEO CITIES" logo and a section about communities. To the right, there's a sidebar with "YOUR HOME ON THE WEB" and a list of links. At the bottom, there are sections for "Free Home Pages & Free Member Email" and "Advertiser Information".

A substantive  
research question?

# GEOCITIES USERS:

**What was  
GeoCities?**

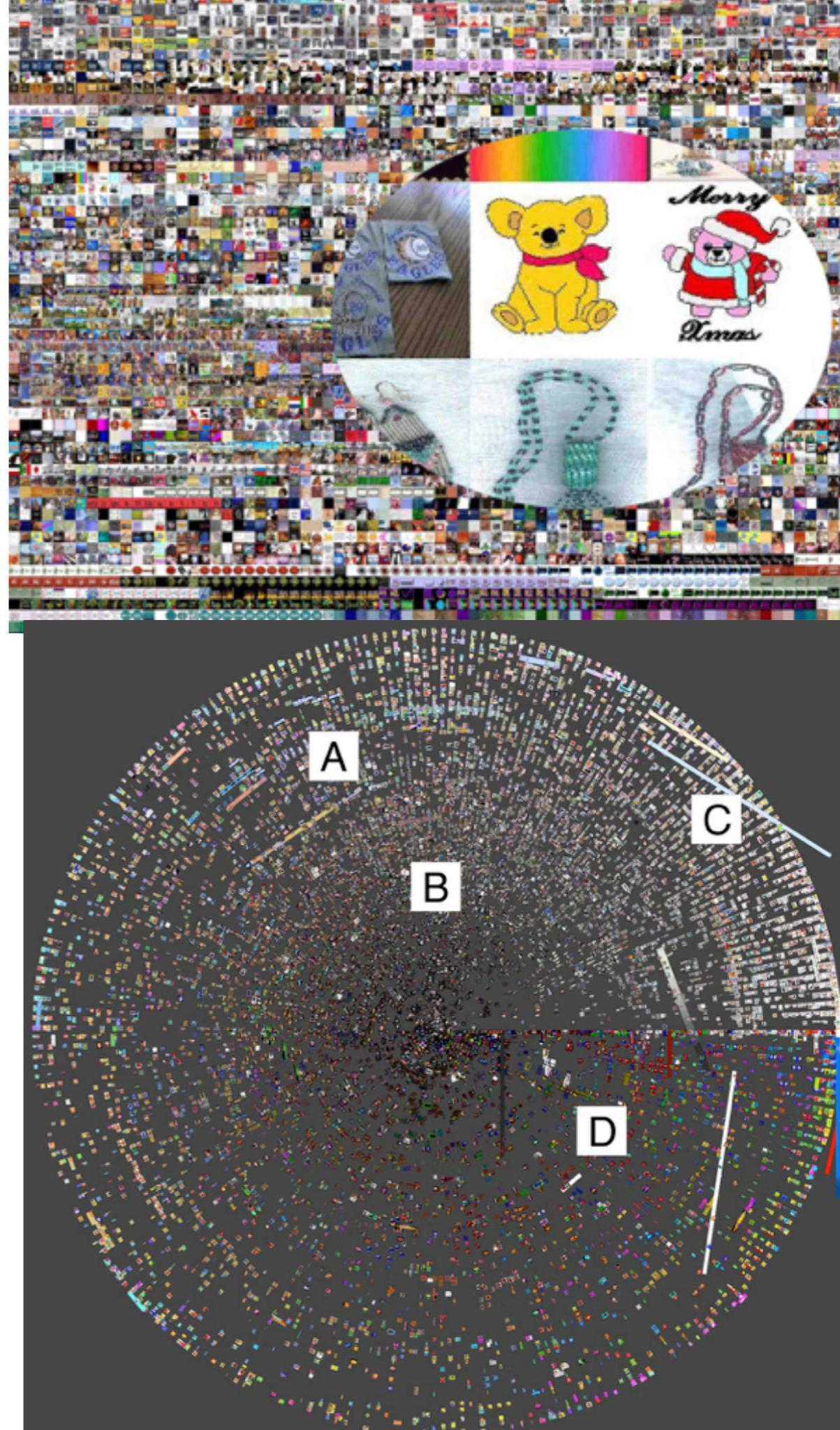
**Why does it  
matter?**



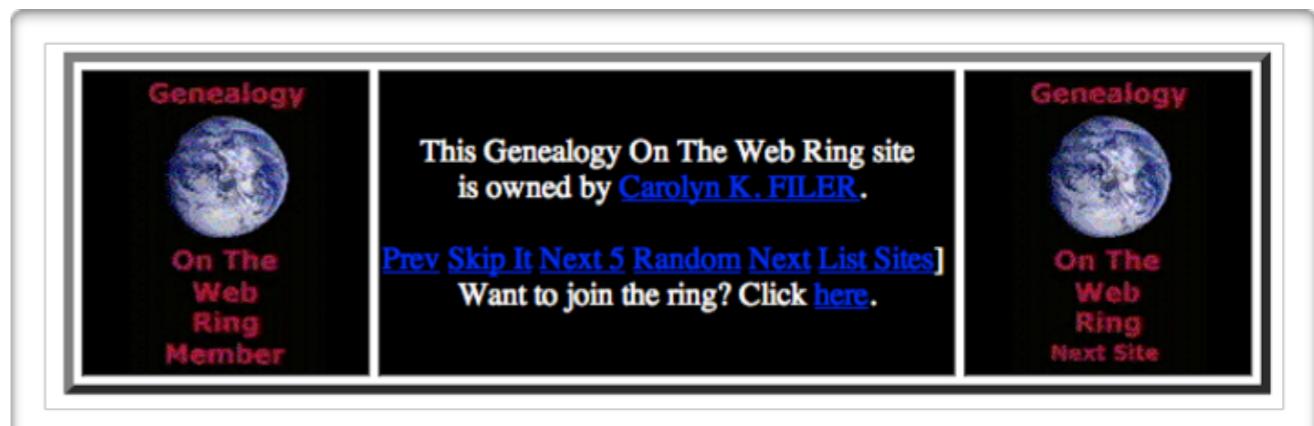
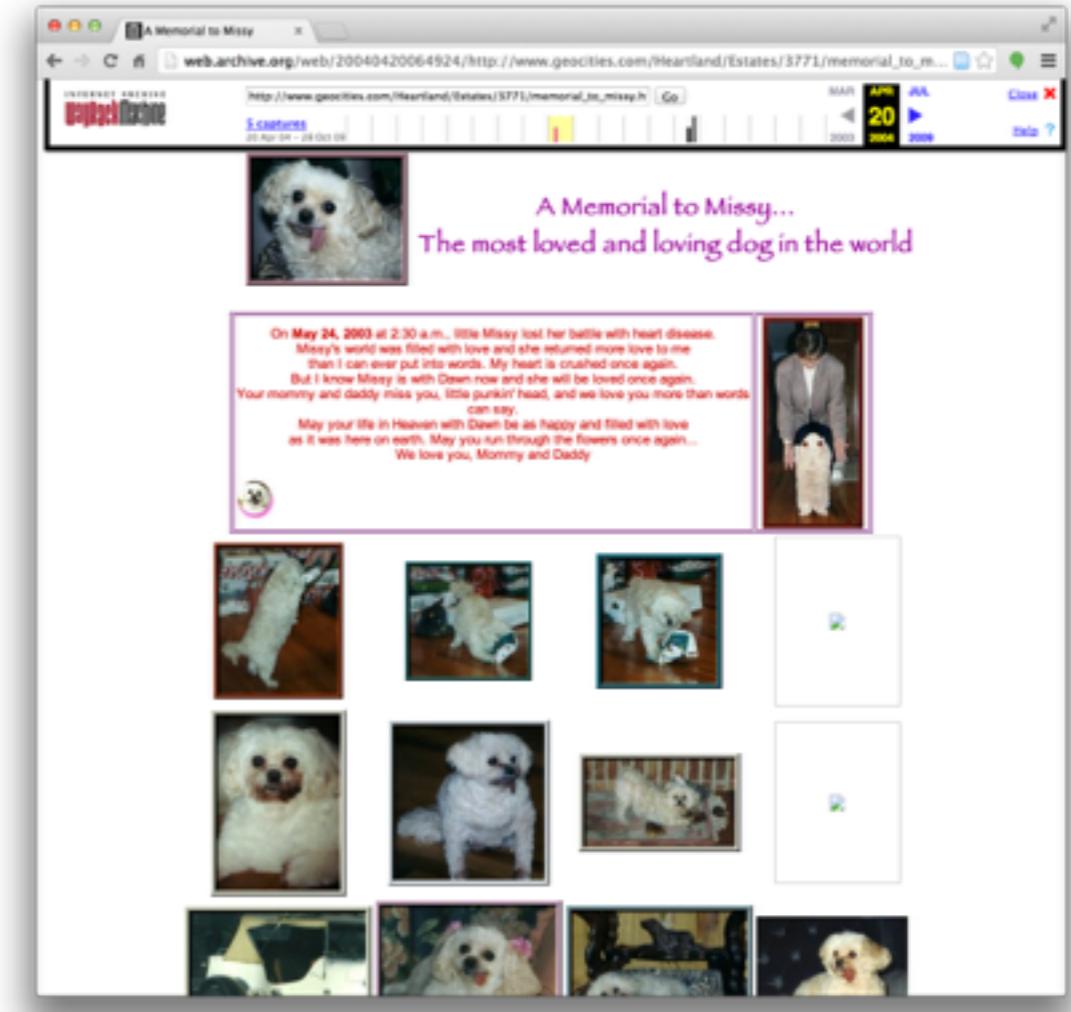
# Topic Modelling Community to Test Coherence

Selected Neighbourhoods	Top Two Topics
Athens <i>“... based on education, teaching, reading, writing and philosophy”.</i>	people things time person sense life man work world k soul make nature body case made point part parts goddess witch healing incense witchcraft lov shaman witches sun spirit protection light circle earth
EnchantedForest <i>“A place for and about kids. Games, stories, educational sites, and homepages created by kids themselves.”</i>	blue page school home day kids clues fun time year room birthday family mom jordan play great jq battalion show st jonny horse battery armored lt artillery camp sailor army field col pingu w
Heartland <i>“A family oriented neighborhood that represents Main Street in cyberspace. This is the place to find parenting, pets, and home town values.”</i>	people time children book years child information ye school person system state world books government g family county church home years information st city b school mrs history birth records great cemetery death
Hollywood <i>“Entertainment capital of the world. Movies, television, and our live video camera at the corner of Hollywood and Vine!”</i>	joey rachel ross monica chandler don yeah phoebe hey mike back gonna ll chris big uh g frasier niles martin daphne roz don back ll door room scene ve dad turns takes crane good
Pentagon <i>Military men and women.</i>	war people president government american world state united general military public soviet political clinton ar fort war civil island iran world adams army british hist german french american forts walther cap newport
WestHollywood <i>“A community with a culture based on gay and lesbian identity.”</i>	gender women sex male female people men person woman sexual crossdressing femin transgendered marriage man children transsexual

**Looking at  
millions of  
user-  
contributed &  
generated  
images**



And the stories of significant users and meaningful experiences.



# Shared Problems

- We actually have common questions – but accessing each of these different case studies required different tools.
- What if there was a platform that could do it all?

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

The screenshot shows the GitHub repository page for 'lintool/warcbase'. At the top, it displays the repository name 'lintool / warcbase'. Below the name, a brief description states: 'Warcbase is an open-source platform for managing web archives built on Hadoop and HBase'. A link to 'http://warcbase.org' is provided. Key statistics are shown: 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.

A large section titled 'Warcbase' provides a detailed description of the project: 'Warcbase is an open-source platform for managing web archives built on Hadoop and HBase. This 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.'

A 'Getting Started' section is present, along with a 'Clone the repo:' button.

A platform for all kinds  
of questions?

# 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 repository page for 'lintool/warcbase'. The title bar indicates it's the 'Home - lintool/warcbase' page. The URL in the address bar is 'GitHub, Inc. [US] https://github.com/lintool/warcbase/wiki'. The main content area is titled 'Home' and features a sub-header 'Ian Milligan edited this page on Jul 23 · 9 revisions'. Below this, there's a welcome message: 'Welcome to the warcbase wiki! Here, you'll find various pig scripts and instructions to unlock your rich web archive collections.' A note states: 'These pages are under active development, as of June 2015.' It encourages users to share their experiences: 'If you are using warcbase, we would love to hear from you. Please let us know!' At the bottom, a note cautions: '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 website: <http://programminghistorian.org/lessons/intro-to-bash>.'

## Getting Started?

This is still actively under development, with several features in the pipeline (no promises).

# 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

Elizabeth May  
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

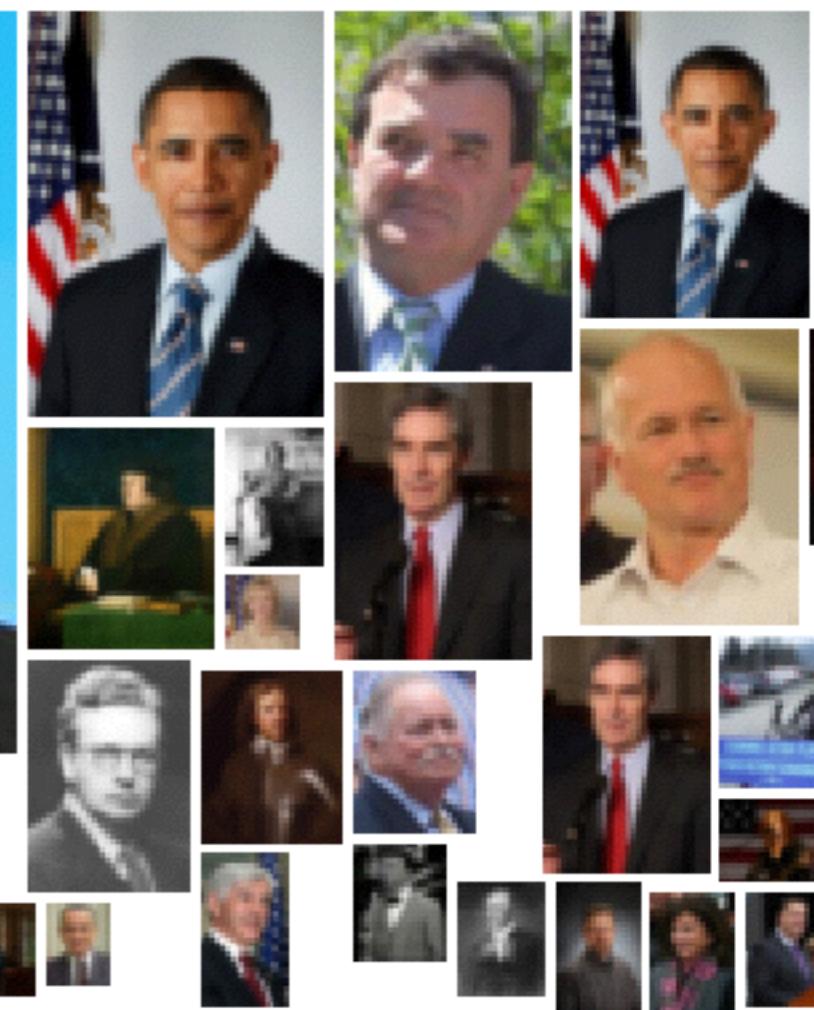
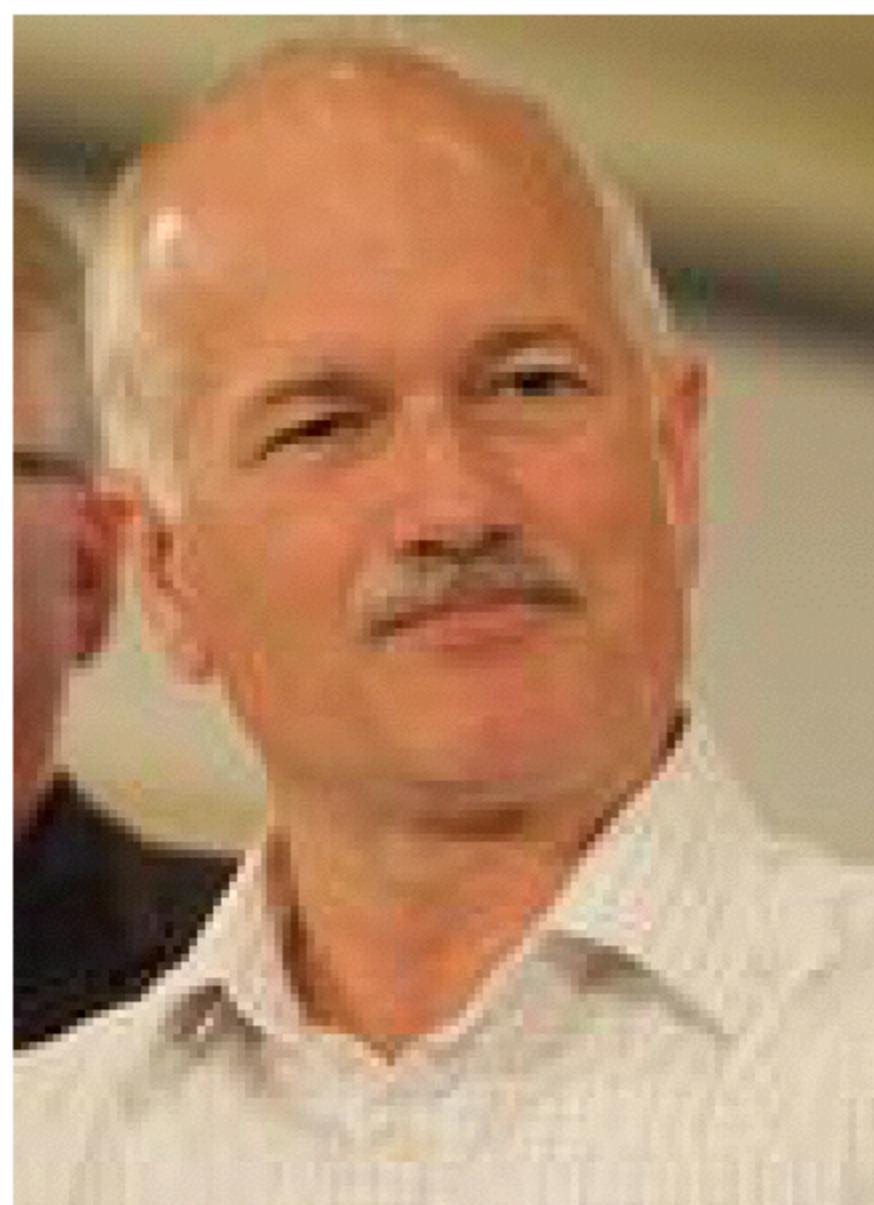
Paul

Paul Martin

Ross

Sharon Labchuk

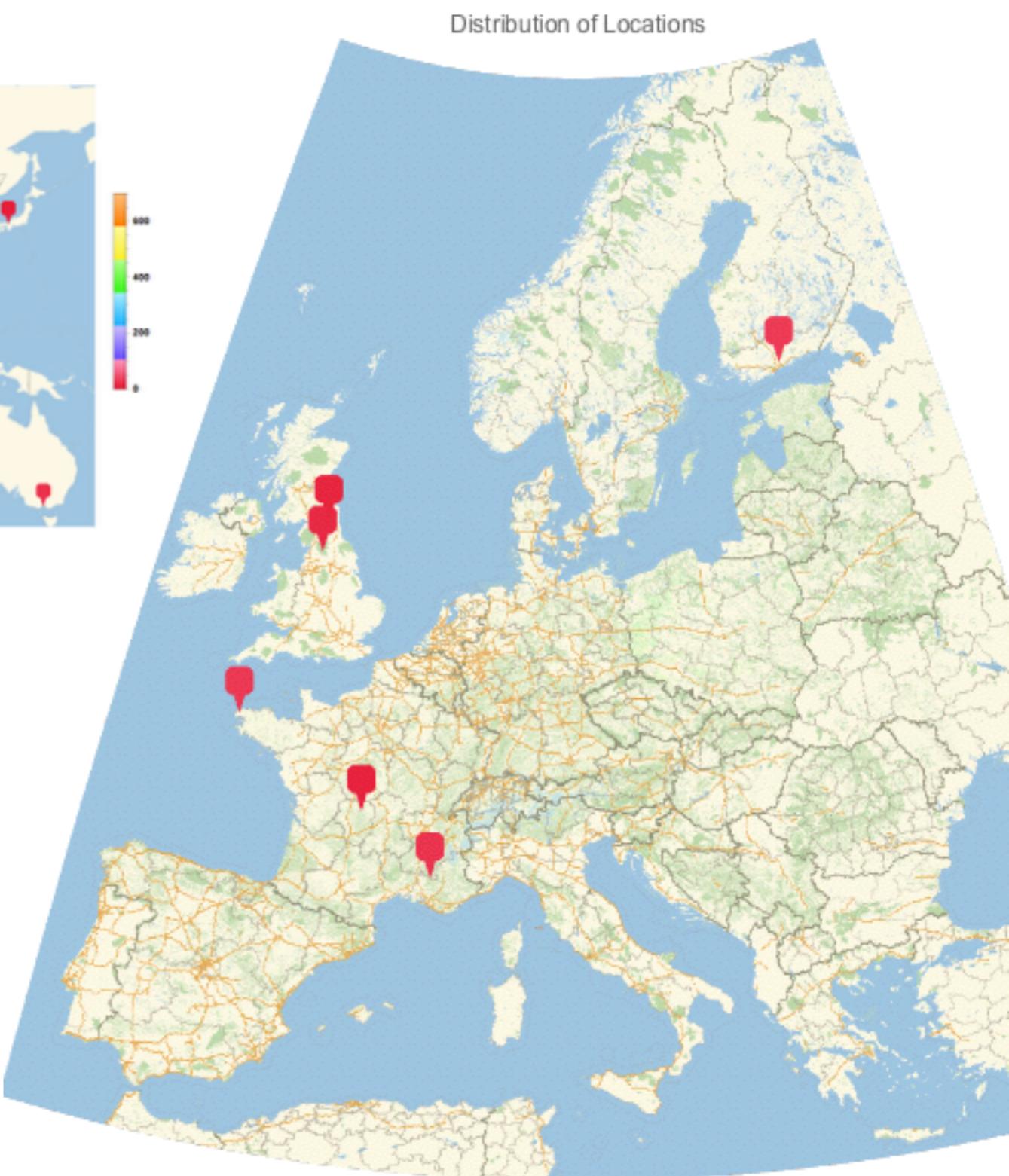
# Extract Entities



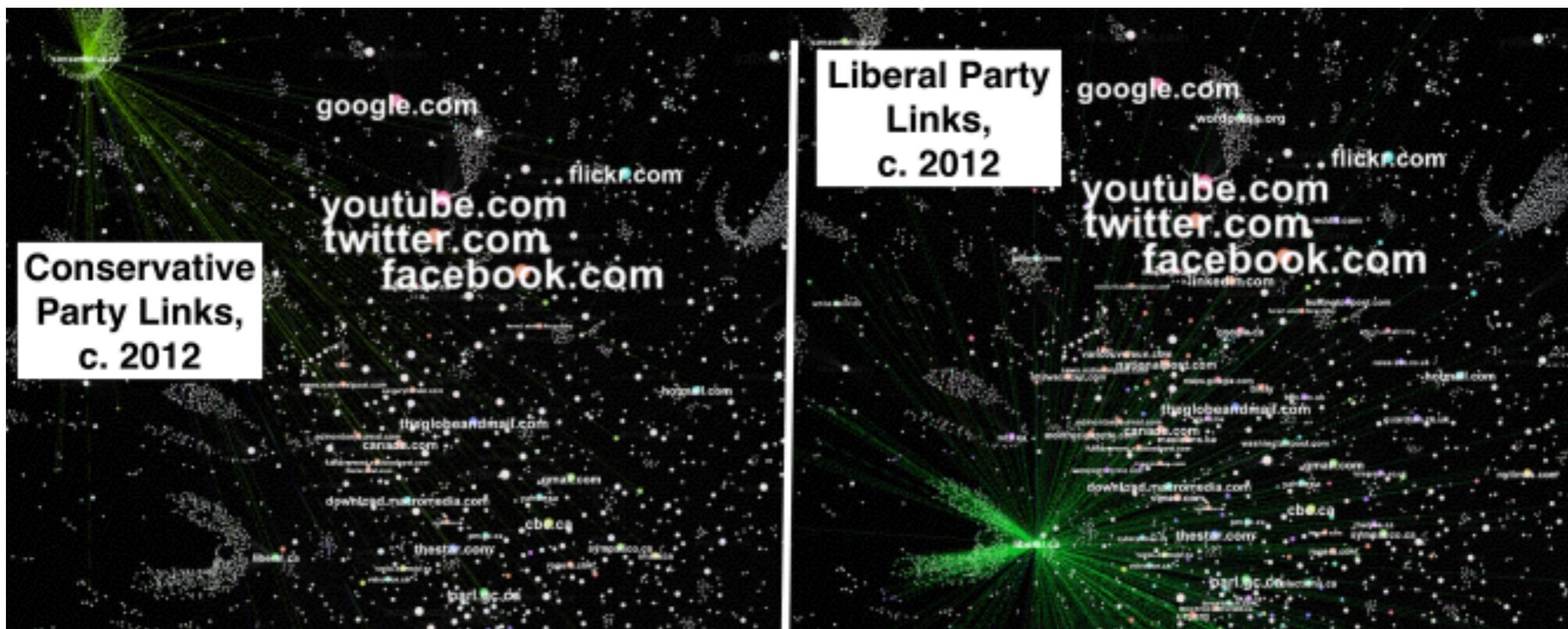
# Extract Entities



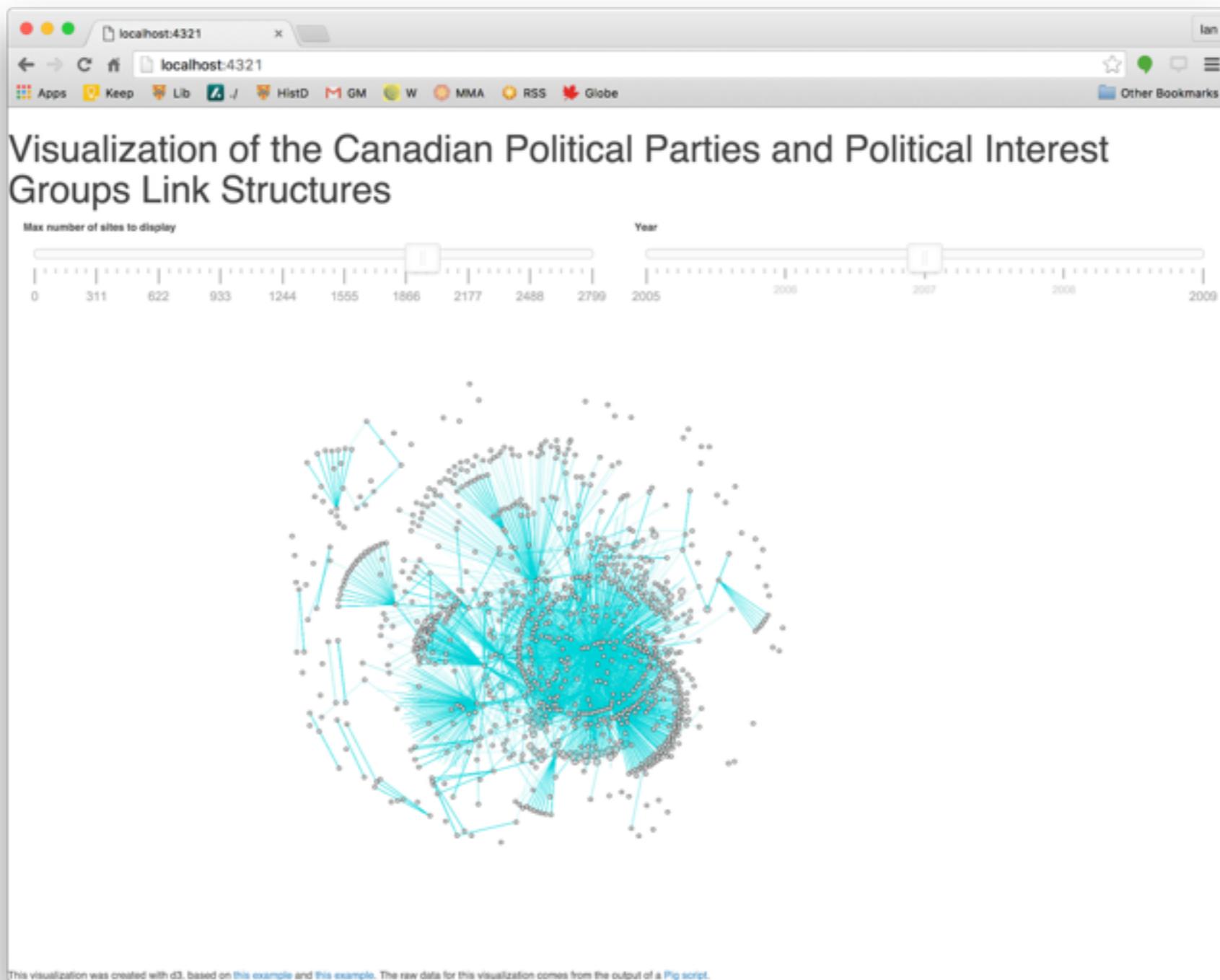
```
In[26]:= int = SemanticInterpretation[#] & /@ processedfreq[[All, 1]]  
Out[26]= {Canada, Calgary, Colombia, Montreal, $Failed, Ontario, Canada, Afghanistan,  
Ottawa, Manitoba, Canada, British Columbia, Canada, Toronto, Nova Scotia, Canada,  
Saskatchewan, Canada, Quebec, Canada, Alberta, Canada,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



# Or D3.js link networks in browser



Spark Notebook    TTOW    lan

localhost:9000/notebooks/TTOW.snb#tab1461784360-0

SPARK NOTEBOOK TTOW (autosaved)

File Edit View Insert Cell Kernel Help Scala [2.10.4] Spark [1.3.0] Hadoop [2.6.0]

Cell Toolbar: None

## TTOW Demo, December 2015

This is a notebook to demo how we're forseeing the rapid prototyping of work with web archives.

Note that we can begin to intersperse text with the code that we're writing, to enable the sharing of notebooks and research ideas.

```
In [1]: :cp /Users/ianmilligan1/dropbox/warcbase/target/warcbase-0.1.0-SNAPSHOT-fat
```

```
In [2]: import org.warcbase.spark.matchbox.  
import org.warcbase.spark.rdd.RecordRDD.  
  
import org.warcbase.spark.matchbox.  
import org.warcbase.spark.rdd.RecordRDD.
```

Out[2]: 161 milliseconds

```
In [3]: var arc="/Users/ianmilligan1/Dropbox/warcs-workshop/227-20051004191331-0000  
var warc="/Users/ianmilligan1/dropbox/wahr/sample-data/arc-warc/ARCHIVEIT-2  
var arcdir="/Users/ianmilligan1/dropbox/warcs-workshop";
```

```
In [4]: val r =  
RecordLoader.loadArc(arc,  
sc)
```

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

# Let's figure it out together!

“Archives Unleashed”  
Hackathon

March 3 - 5 2016, University  
of Toronto Library

[archivesunleashed.ca](http://archivesunleashed.ca)

Travel funds for grad  
students/contingent faculty &  
researchers



But the shared  
promise...



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

# Thank you!

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

---

**Ian Milligan**  
**Assistant Professor**  
**@ianmilligan1**



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