

# **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<br>HOME<br>PAGES<br>AND<br>E-MAIL | ARTS<br>AUTOS<br>BUSINESS<br>COMPUTERS<br>CULTURE | EDUCATION<br>ENTERTAINMENT<br>ENVIRONMENT<br>FAMILY<br>FASHION | FOOD<br>GAMES<br>GAY & LESBIAN<br>GOVERNMENT<br>HEALTH | KIDS<br>MUSIC<br>PEOPLE<br>RECREATION<br>SCIENCE F... |
|--|---|--|--|---|



199

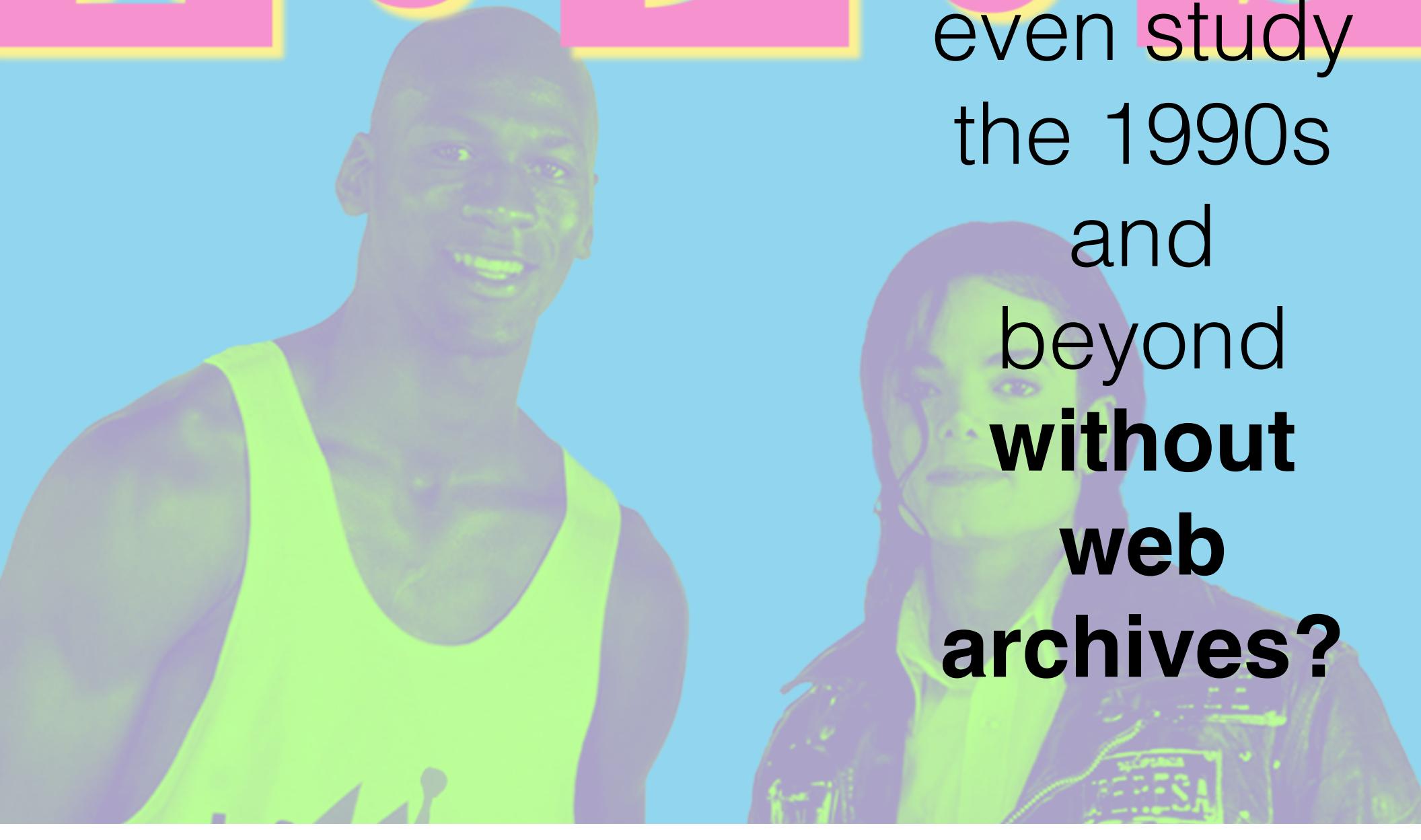
99

99

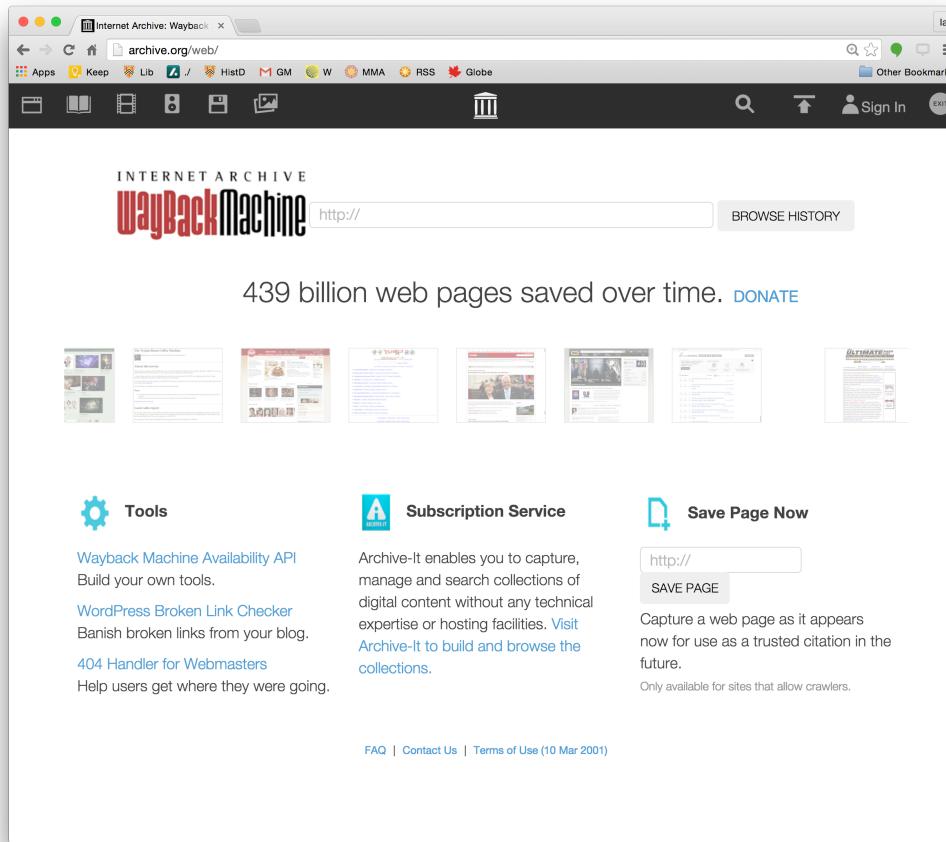
0

S

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?

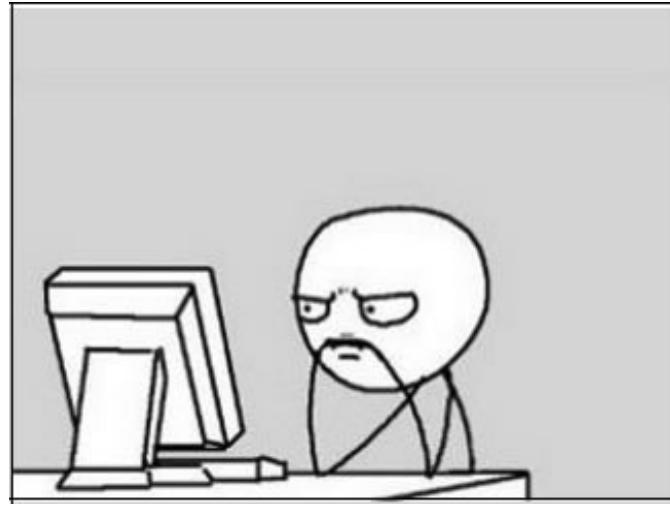
# But what will web archives look like?

- 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 a dark theme. The address bar displays the URL <https://archive.org/details/wide00002&tab=about>. The main content area is titled "Wide Crawl started March 2011". Below the title, a sub-headline reads: "Web wide crawl with initial seedlist and crawler configuration from March 2011. This uses the new HQ software for distributed crawling by Kenji Nagahashi." There is a "MORE" link. At the bottom of this section, there are three navigation links: "About" (which is underlined), "Collection", and "Forum".  
  
The "DESCRIPTION" section contains the same text as the sub-headline. It also includes a "What's in the data set:" section which lists the following statistics:

- 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

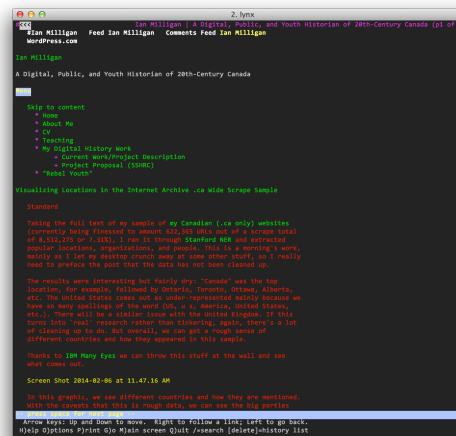
The "The seed list for this crawl was a list of Alexa's top 1 million web sites, retrieved close to the crawl start date. We used Heritrix (3.1.1-SNAPSHOT) crawler software and respected robots.txt directives. The scope of the crawl was not limited except for a few manually excluded sites."

On the right side of the page, there is a sidebar with a yellow header "Created on October 5 2010" and a profile picture of a woman. Below this, it says "ADDITIONAL CONTRIBUTOR" and lists two more profiles: "brewste" and "kngenie", both labeled "Archivist". At the very bottom right, there are "VIEWS" and other small icons.

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

| <b>Top-Level Domain</b> | <b>Number of Distinct URLs Downloaded in Sample</b> | <b>Number of Overall URLs in Wide Web Scrape (selected domains)</b> | <b>Percentage of URLs Captured</b> |
|-------------------------|---|---|------------------------------------|
| .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)



WARC File

Plain Text

Indexing

Carrot2 Workbench

Source: Solr  
Algorithm: Lingo

Basic  
Query (Required): children  
Read Solr clusters if present

Results: 1000

Aduna Cluster Map Visualizat... Circles Visualization FoamTree Visualization

**Clusters**

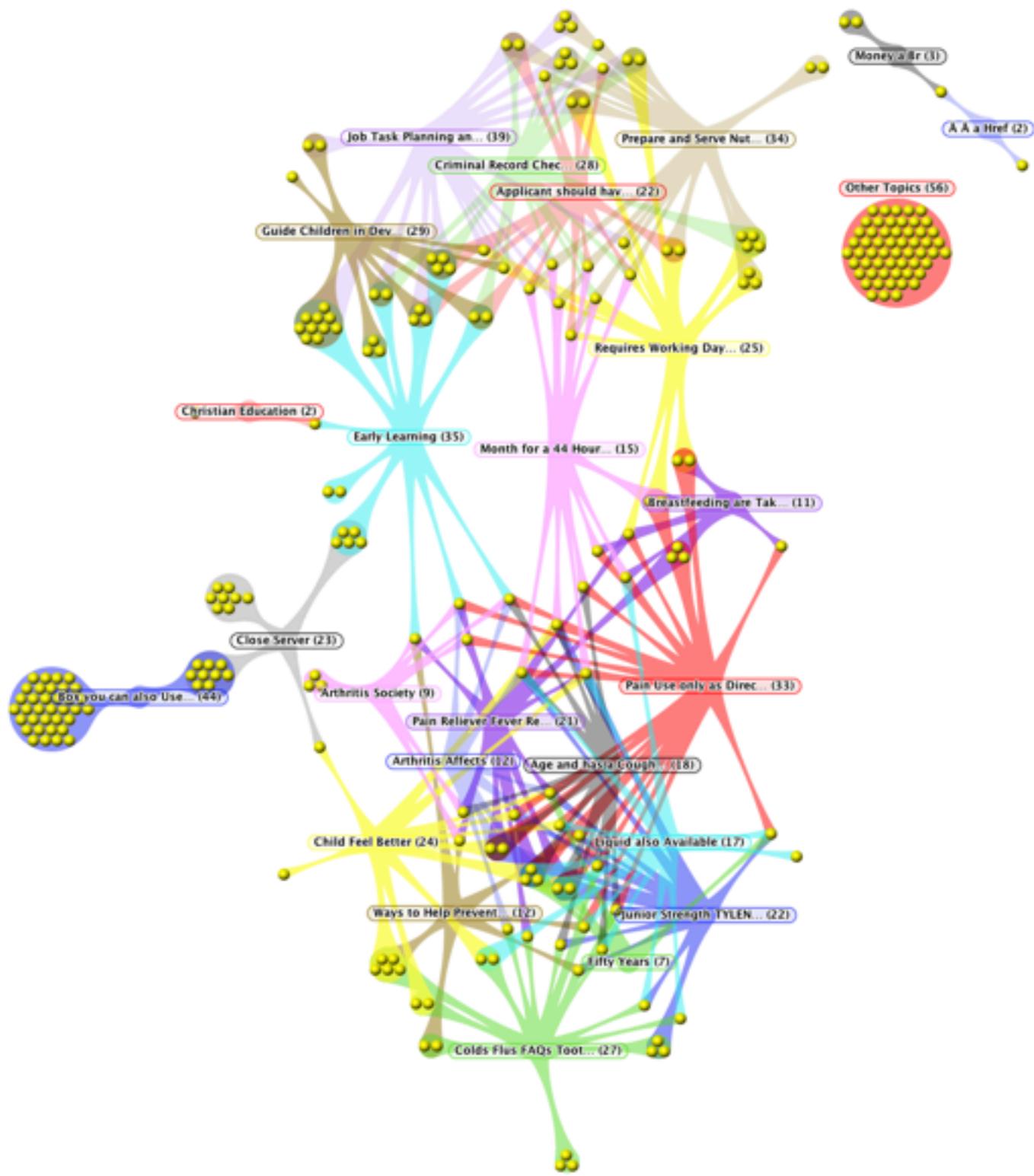
- Child Health (192)
- Canada Service (168)
- Left side of the Page (161)
- Document Input (158)
- Research Research (147)
- Health Centre (138)
- Children Value (127)
- Services Community (123)
- Consumer Product (122)
- Providing Services (120)
- Health Community (113)
- School Services (111)
- Health and Wellness (105)
- Health Services (103)
- New Image (101)
- Returns List (98)
- Support Services (98)
- Public Health (97)
- Health and Safety (95)
- Family Services (93)
- Education Document (92)
- Service Days (91)
- Research Programs (88)
- Health Promotion (84)
- Development Research (83)
- Research will Help (82)
- Youth Services (82)
- Services Community Education (74)
- Health Professionals (74)
- Research Resources (69)
- Areas of Health (63)
- University of Ottawa (58)
- Community Health Centre (54)
- Research and Events (56)
- Mental Health (53)
- Health Issues (54)
- Research Interests (50)
- Invitation Templates (48)
- University University of Ottawa f (46)
- Flu Is Available (38)
- Natural Health Products (38)
- Products and Services (35)
- Birthday Party Invitations (27)
- Centre for Research on Commun (24)
- Birthday Age (5)
- Youth Services Bureau of Ottawa (5)
- Other Topics (365)

**children (1000 documents from Solr, 47 clusters from Lingo)**

Documents

- [1] <http://www.tylenol.ca/children/children-6-11-years/cough-cold-products>: text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Français Search \_\_\_\_\_ Search \* Adult \* Children \* P...  
/Users/kanniligan1/Desktop/output/76-Canadian-456.html
- [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 | Français Search \_\_\_\_\_ Search \* Adult \* Children \* Prod...  
/Users/kanniligan1/Desktop/output/76-Canadian-1721.html
- [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 | Français Search \_\_\_\_\_ Search \* Adult \* Children \* Prod...  
/Users/kanniligan1/Desktop/output/72-Canadian-1170.html
- [1] <http://www.tylenol.ca/children/products>: text/html; charset=utf-8 For Adults For Children Tylenol logo Home | Contact us | Français Search Search \* Adult \* Children \* Products \* About Tylenol \* News & Information All Children's Pro...  
/Users/kanniligan1/Desktop/output/25-Canadian-3512.html
- [1] <http://blogs.afortunecookie.ca/tag/children/feed/>: text/html  
/Users/kanniligan1/Desktop/output/23-Canadian-2494.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 | Français Search \_\_\_\_\_ Search \* Adult \* Children...  
/Users/kanniligan1/Desktop/output/70-Canadian-886.html
- [1] <http://www.tylenol.ca/children/children-3-5-years/aches-pains/relieving-your-child-s-aches-pain>: text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Français Search \_\_\_\_\_ Search ...  
/Users/kanniligan1/Desktop/output/29-Canadian-2278.html
- [1] [http://www.tylenol.ca/children/children-6-11-years/cough-cold-flu/relieving-your-child-s-aches-pains](http://www.tylenol.ca/children/children-6-11-years/cough-cold-flu/relieving-your-child-s-aches-pains/relieving-your-child-s-aches-pain): text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Français Search \_\_\_\_\_ Search ...  
/Users/kanniligan1/Desktop/output/40-Canadian-1.html
- [1] <http://www.tylenol.ca/children/children-6-11-years/aches-pains/relieving-your-child-s-aches-pain>: text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Français Search \_\_\_\_\_ Search ...  
/Users/kanniligan1/Desktop/output/37-Canadian-2224.html

13198 of 4094M



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

**Clusters**

- Box you can also Use it Program
- Job Task Planning and Organizizi
- 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> : text/html; charset=latin1\_swedish\_ci

CLWR funds Nicaraguan medical and dental clinic, scholarships

2010 [Nicaraguan\_medic... /Users/ianmilligan1/Desktop]

**Services**

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

**Search With Google**

**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¢â€ša Href (2)
- Other Topics (56)

Lutheran Church-Canada

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

INTERNET ARCHIVE WaybackMachine 3 captures 5 Dec 10 - 14 Jul 11 DEC JUL 14 2010 2011

**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 attend 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.

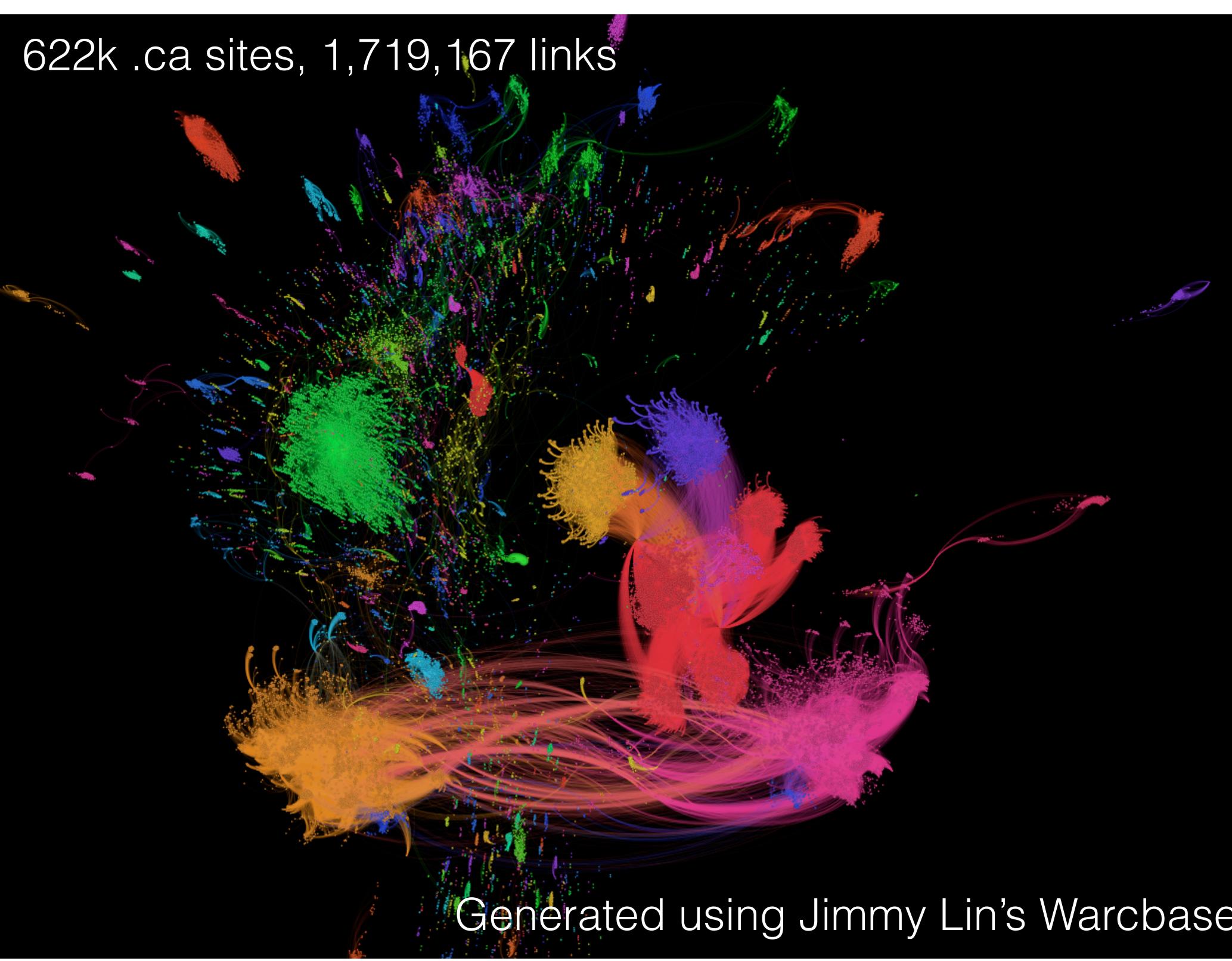
Robert Granke, executive director of CLWR, visited congregations of the ILSN in November. You can read more about his visit at [www.lccontheroad.ca](http://www.lccontheroad.ca), The Canadian Lutheran or in the forthcoming issue of CLWR's Partnership newsletter due out in early February.



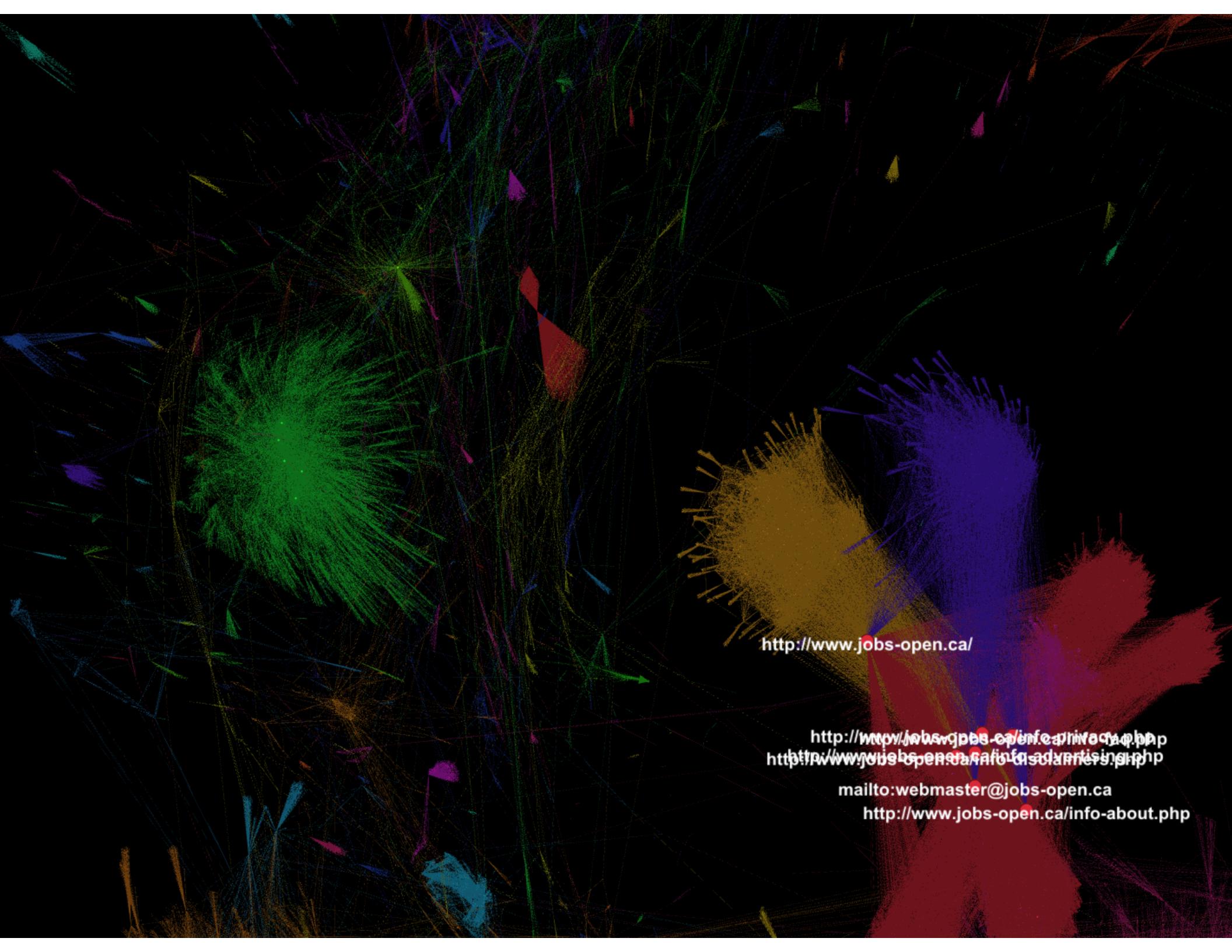
A medical clinic in Nicaragua.

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

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



Generated using Jimmy Lin's Warchbase



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

<http://www.jobs-open.ca/info-about.php>  
<http://www.jobs-open.ca/advertising.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/bienvenue.html>

<https://Web3.uottawa.ca/Htweb/logon/fr.html>

<http://www.ressourcesfinancieres.uottawa.ca/etudiant/payment-university-fees-fr.php>

<http://www.uottawa.ca/academicinfo/bourses/courseship?tabid=2565>

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

<http://www.uottawa.ca/contacteznous.htm>

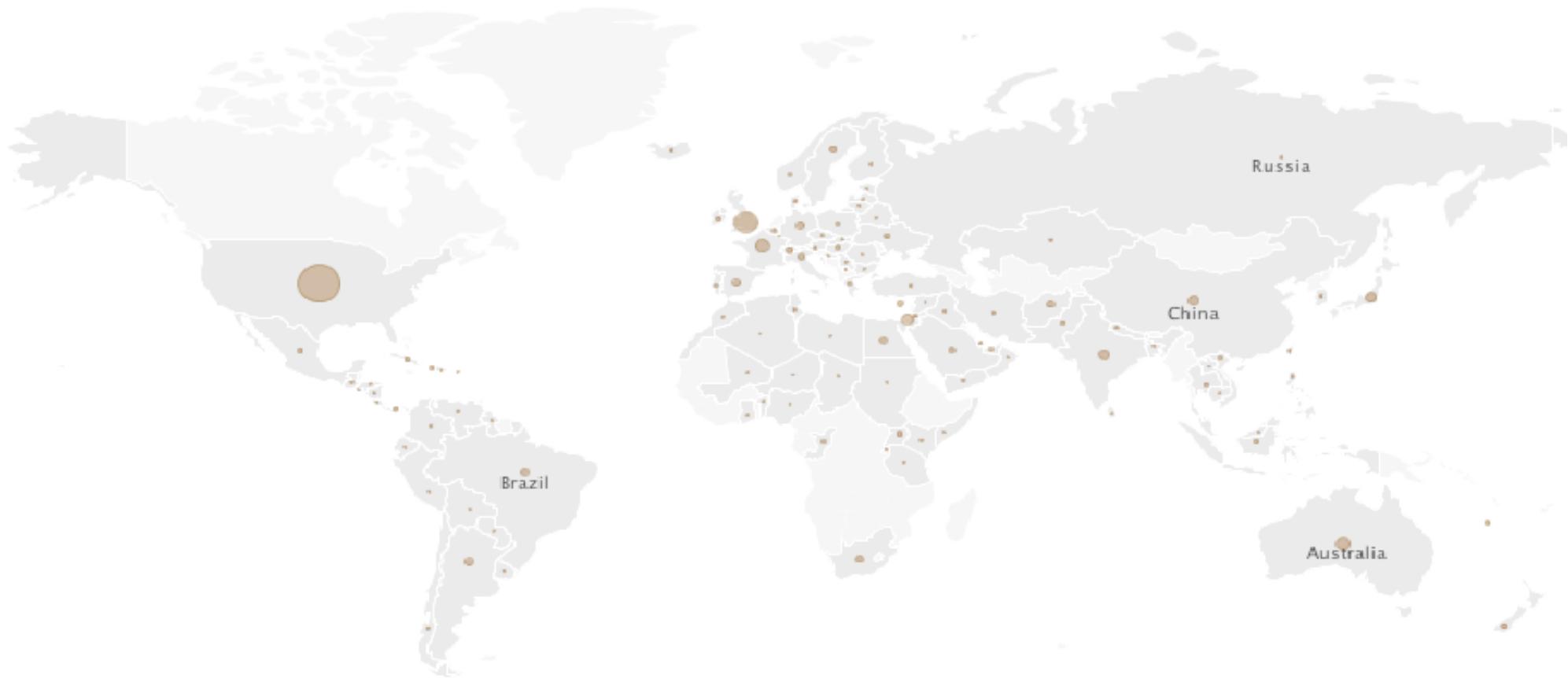
[http://www.uottawa.ca/etudesuniversitaires.asp?Lang=FR&U=University%C3%A9&utm\\_campaign=template&utm\\_source=uottawa.caaccueil.html?utm\\_medium=referral&utm\\_content=university&utm\\_term=university&utm\\_id=2548&utm\\_sourceFqp](http://www.uottawa.ca/etudesuniversitaires.asp?Lang=FR&U=University%C3%A9&utm_campaign=template&utm_source=uottawa.caaccueil.html?utm_medium=referral&utm_content=university&utm_term=university&utm_id=2548&utm_sourceFqp)

<http://www.uottawa.ca/Default.aspx?tabid=2672>

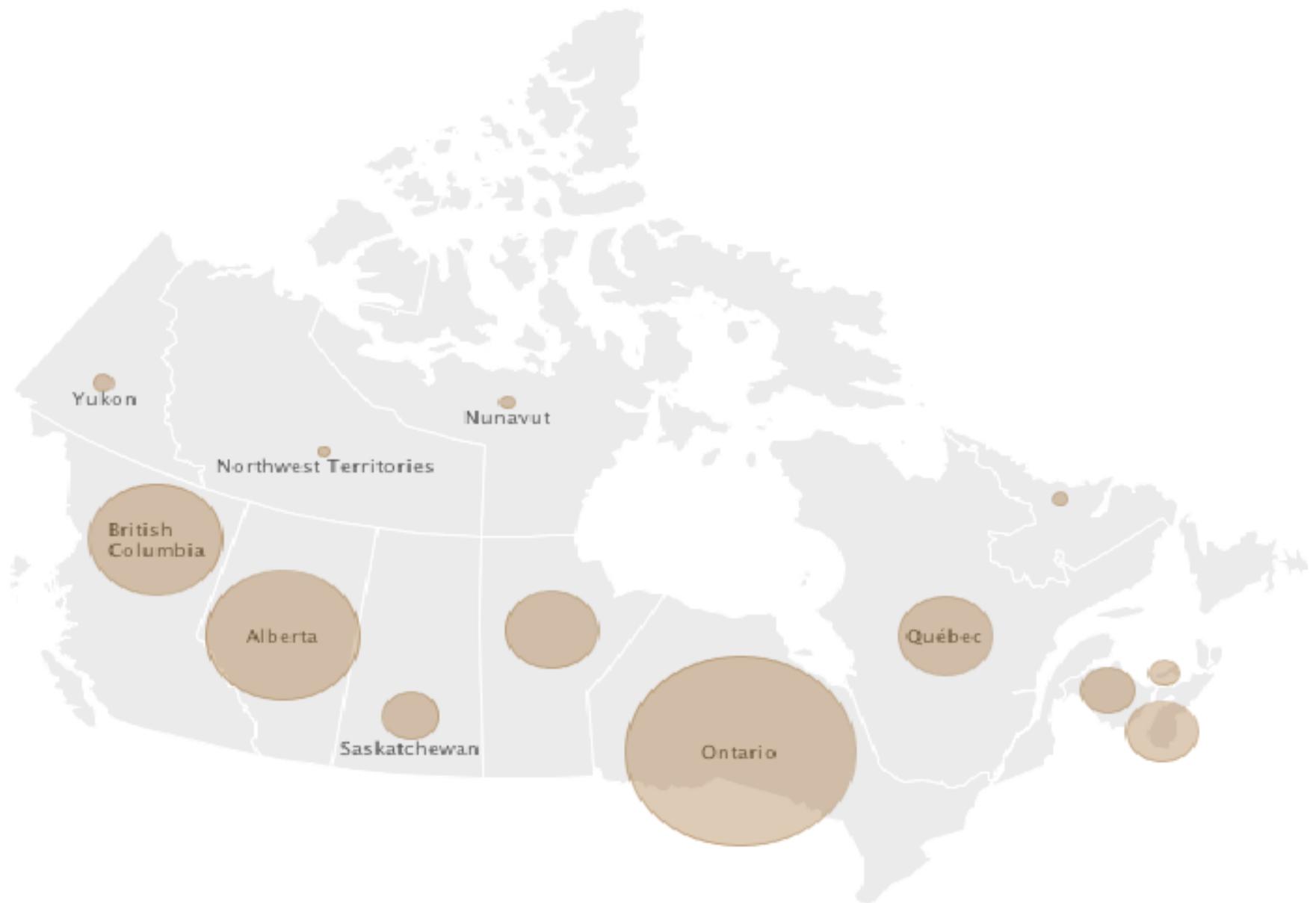
[https://web9.uottawa.ca/service/studesuniversitaires.asp?Lang=FR&U=University%C3%A9&utm\\_campaign=template&utm\\_source=uottawa.caaccueil.html?utm\\_medium=referral&utm\\_content=university&utm\\_term=university&utm\\_id=2548&utm\\_sourceFqp](https://web9.uottawa.ca/service/studesuniversitaires.asp?Lang=FR&U=University%C3%A9&utm_campaign=template&utm_source=uottawa.caaccueil.html?utm_medium=referral&utm_content=university&utm_term=university&utm_id=2548&utm_sourceFqp)

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

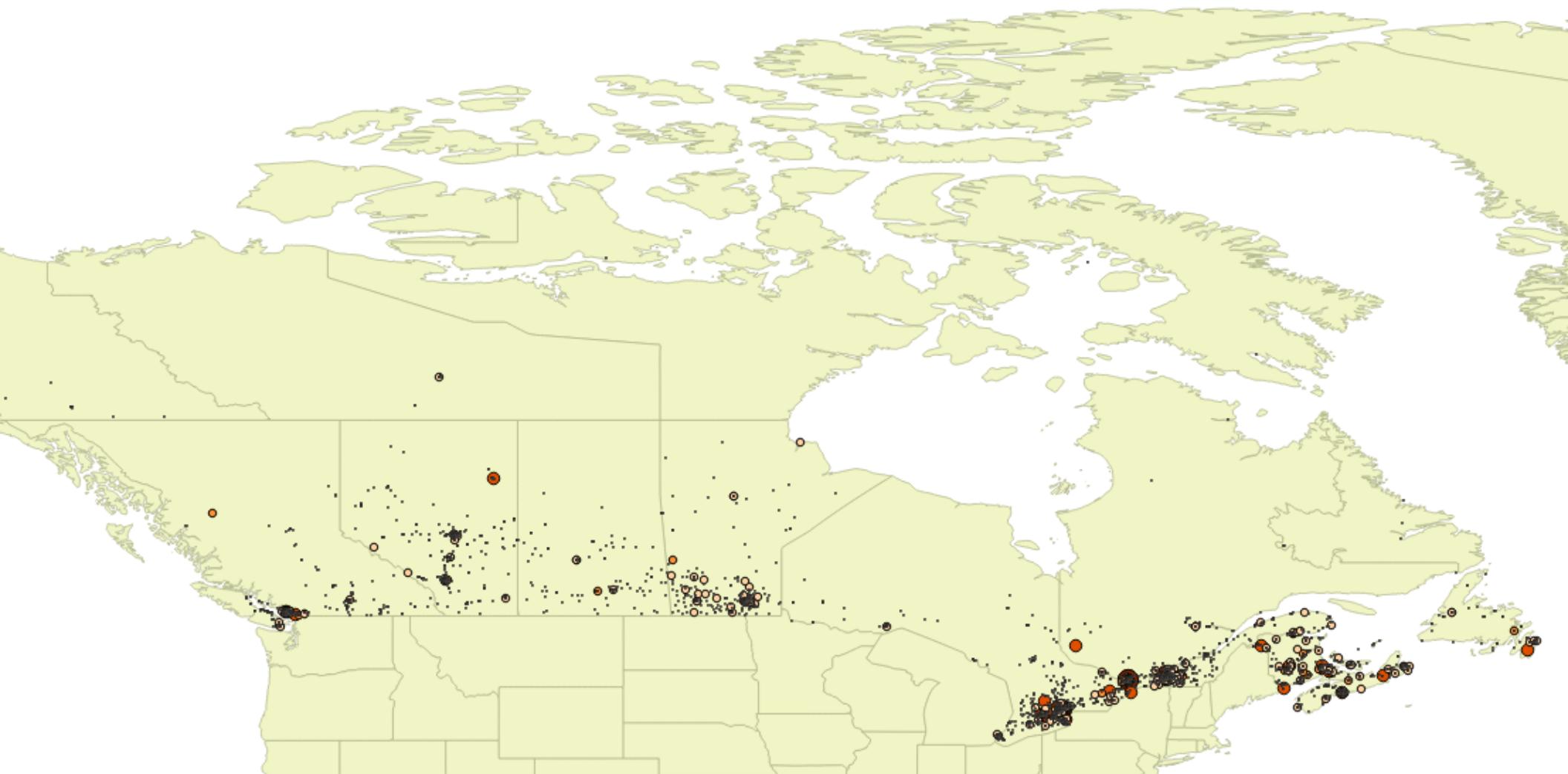
# Countries Mentioned in .ca TLD (excluding Canada)



# 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 "Archive-It - Canadian Political Parties and Political Interest Groups". The header includes links for HOME, EXPLORE, LEARN MORE, and CONTACT US. A banner at the top right says "The leading... for co... culture... Built...". Below the banner, the collection title "Canadian Political Parties and Political Interest Groups" is displayed, along with "Collected by: University of Toronto", "Archived since: Oct, 2005", and a brief description. A large "ARCHIVE-IT" logo is visible. The main content area features a section titled "Narrow Your Results" with a search bar and buttons for "Sites" and "Search Page Text". At the bottom, it shows "Page 1 of 1 (54 Total)" and sorting options.

Explore > University of Toronto > Canadian Political Parties and Political Interest Groups

**Canadian Political Parties and Political Interest Groups**  
Collected by: [University of Toronto](#)

Archived since: Oct, 2005  
Description: Canadian Political Parties and Political Interest Groups, national Canadian political parties, and a number of special interest groups.  
Subject: [Politics & Elections](#)  
Collector: [University of Toronto](#)

Narrow Your Results

Enter search terms here

Sites Search Page Text

Page 1 of 1 (54 Total)

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?

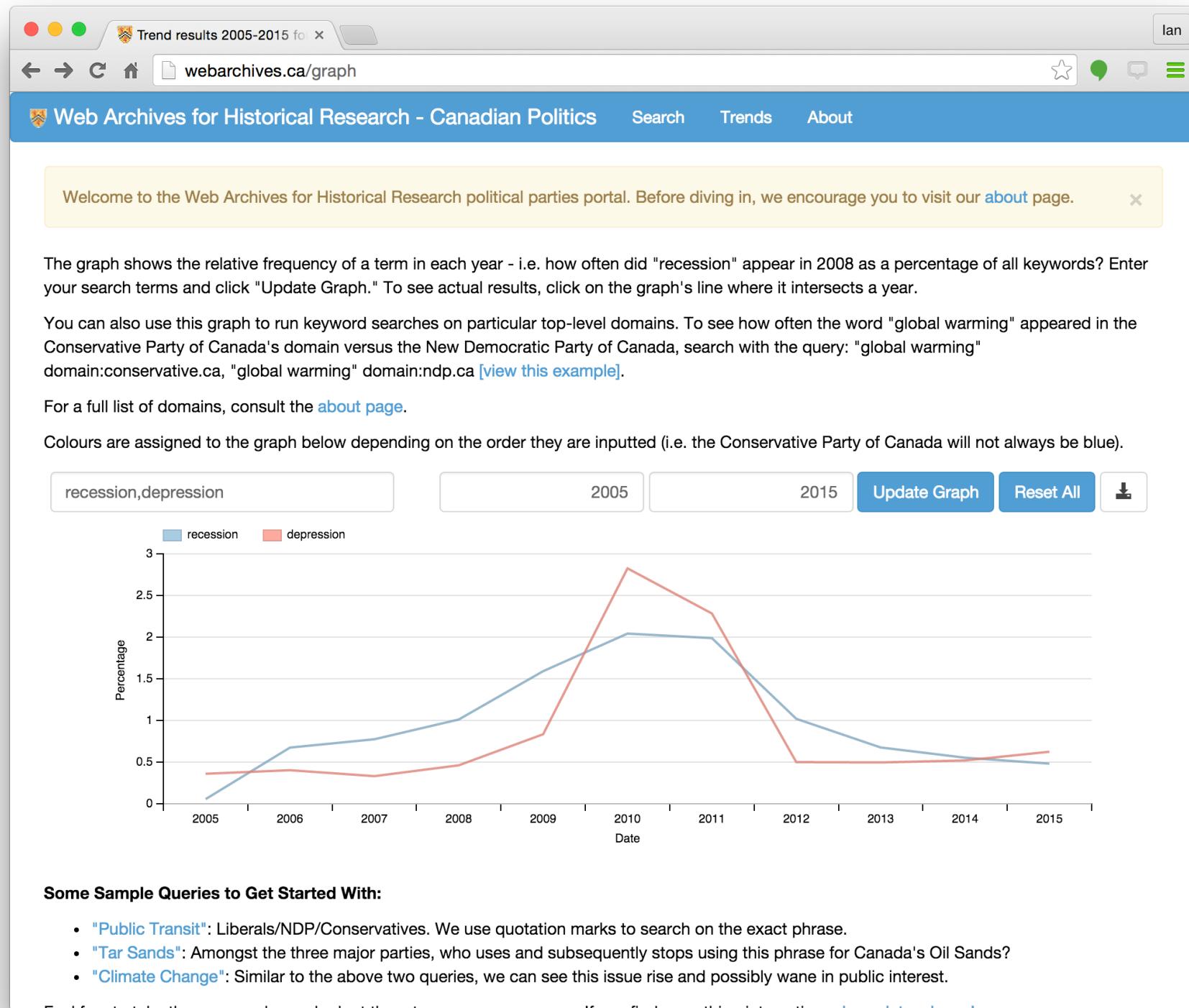


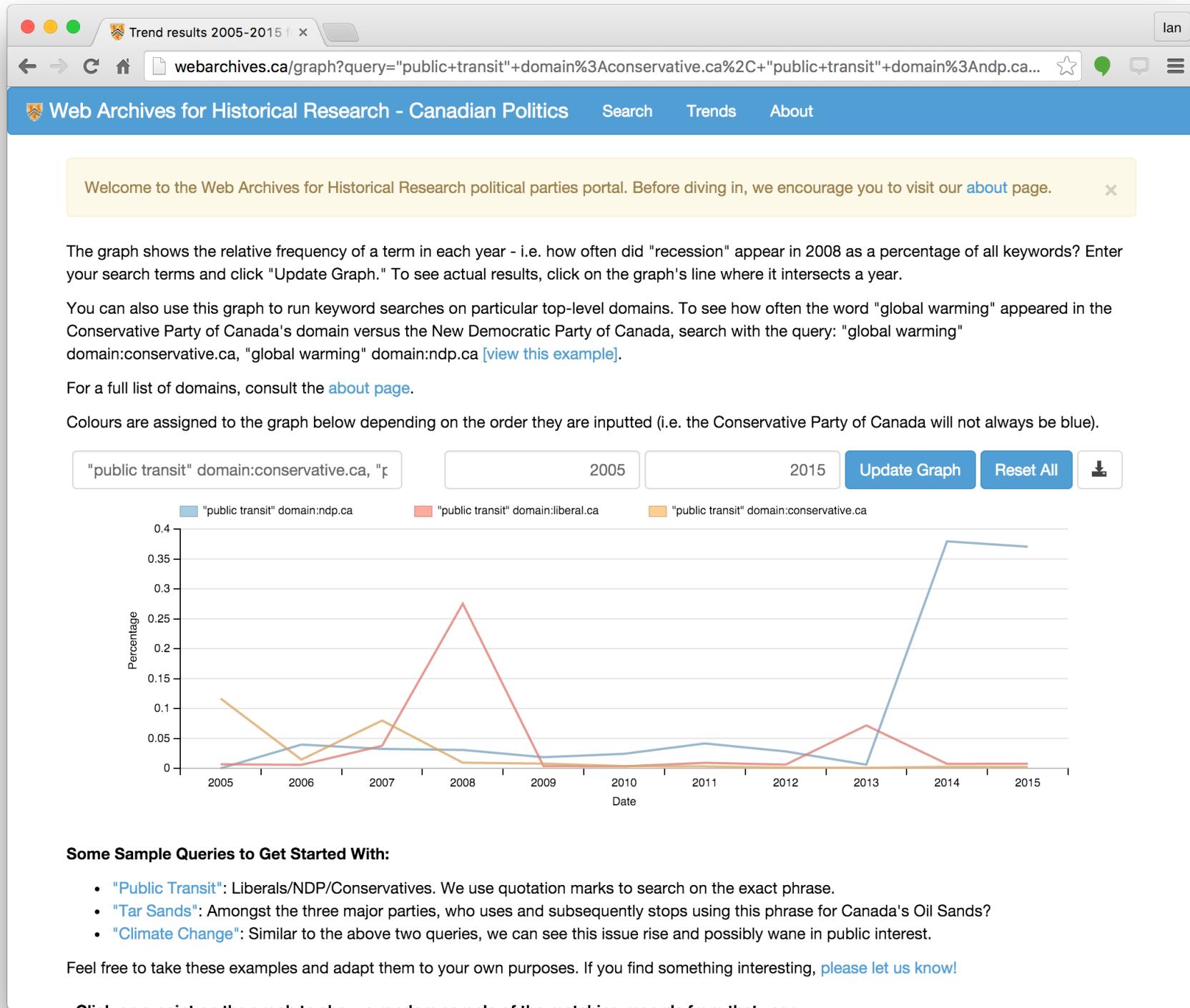
# Current Interface

- Very limited - simple search engine, some advanced options; no facets
- Great collections.. but nobody uses them!

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

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....

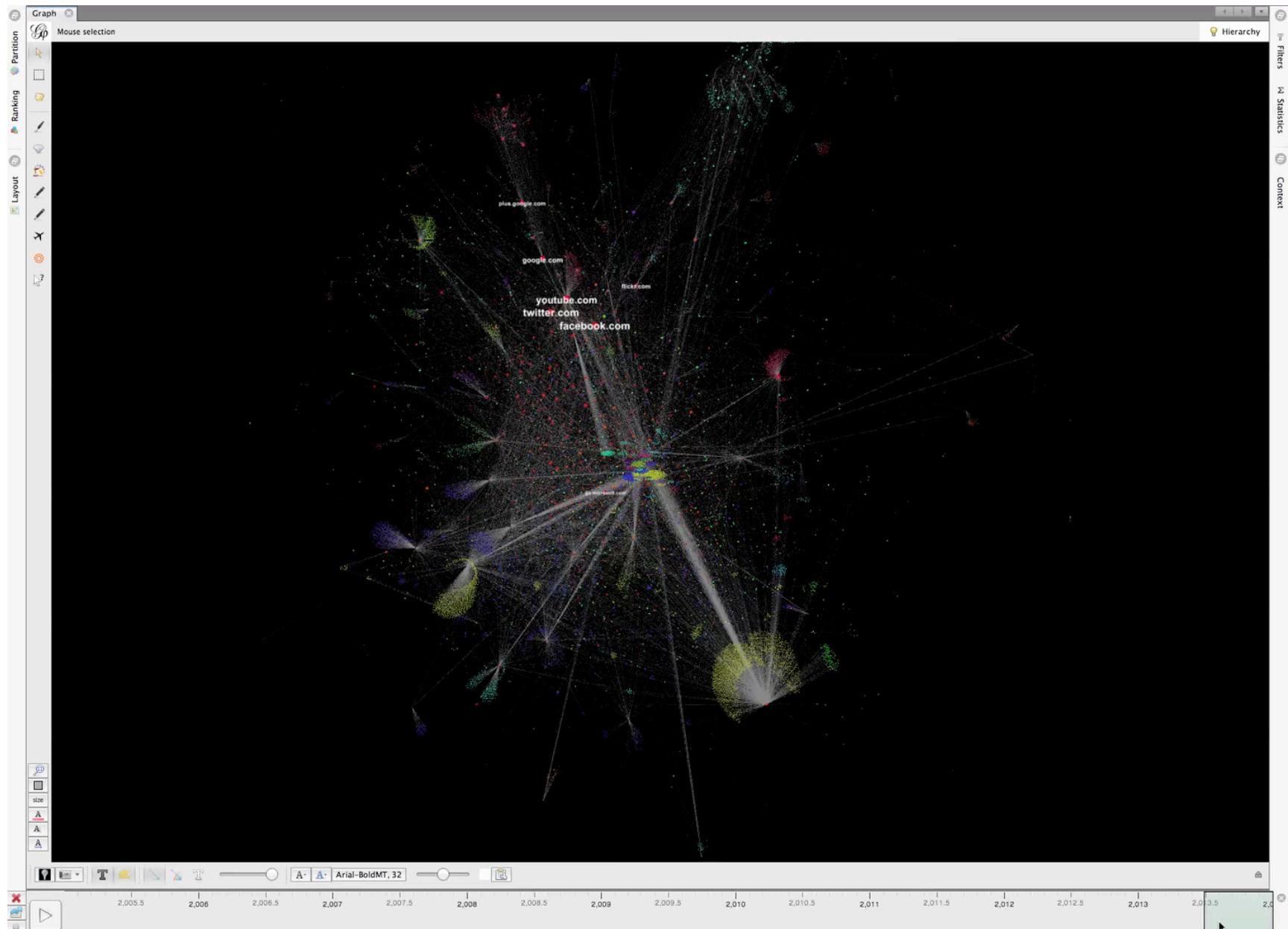
The screenshot shows a web browser window with the following details:

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

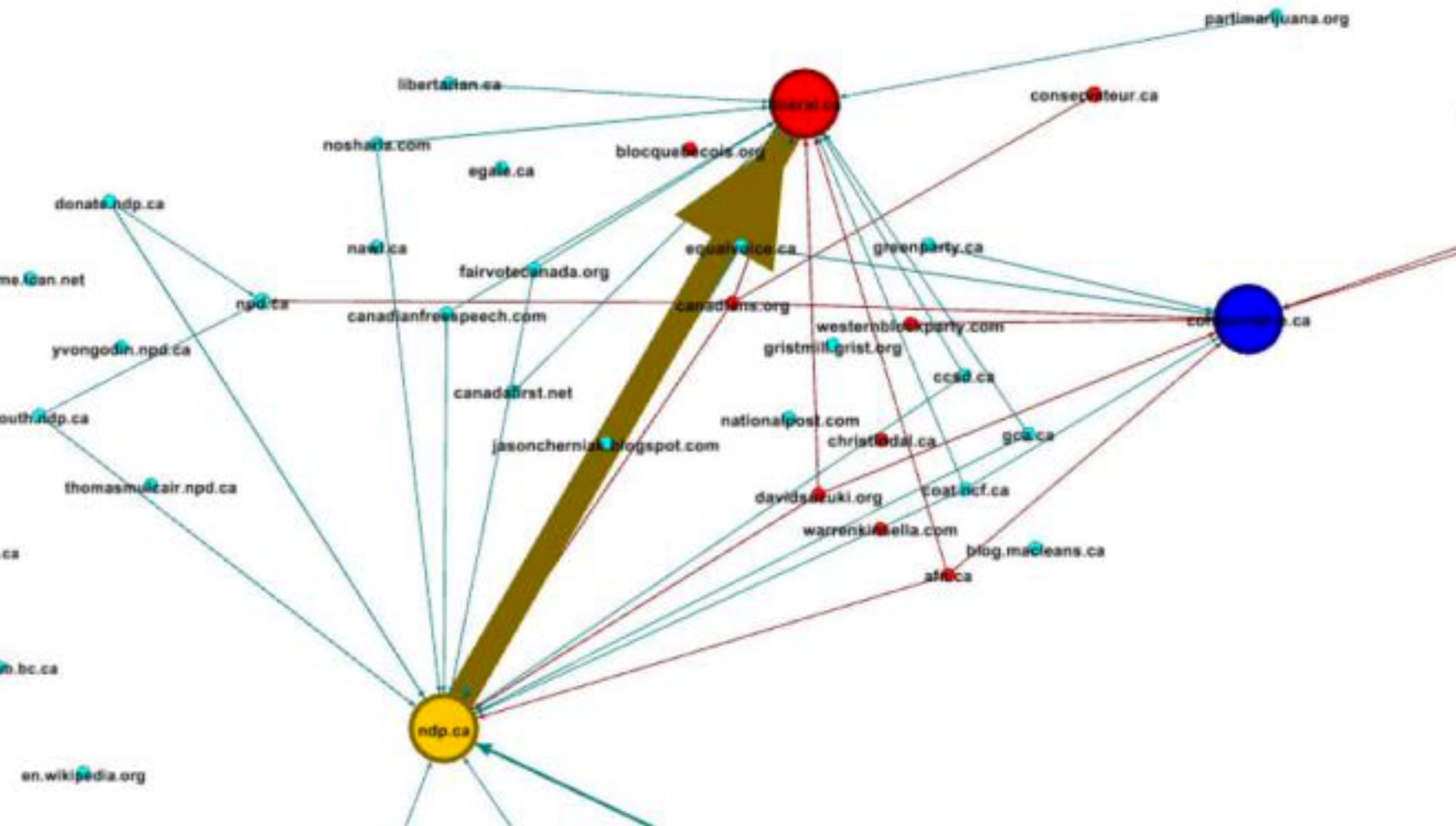
**Do we want metadata  
or content analysis?**

**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 collection of Geocities data from October 2009, including a thumbnail of a website for "Events & Adventures" and a media player showing two video clips: "www.geocities.com.7z" and "Interview with Jason Scott regarding Geocities". The bottom of the page includes a "Download item" section and a "Web Crawls > Archive Team > The Archive Team Geocities Valhalla > The Archive Team Geocities Snapshot (Part 1 of 8)" link.

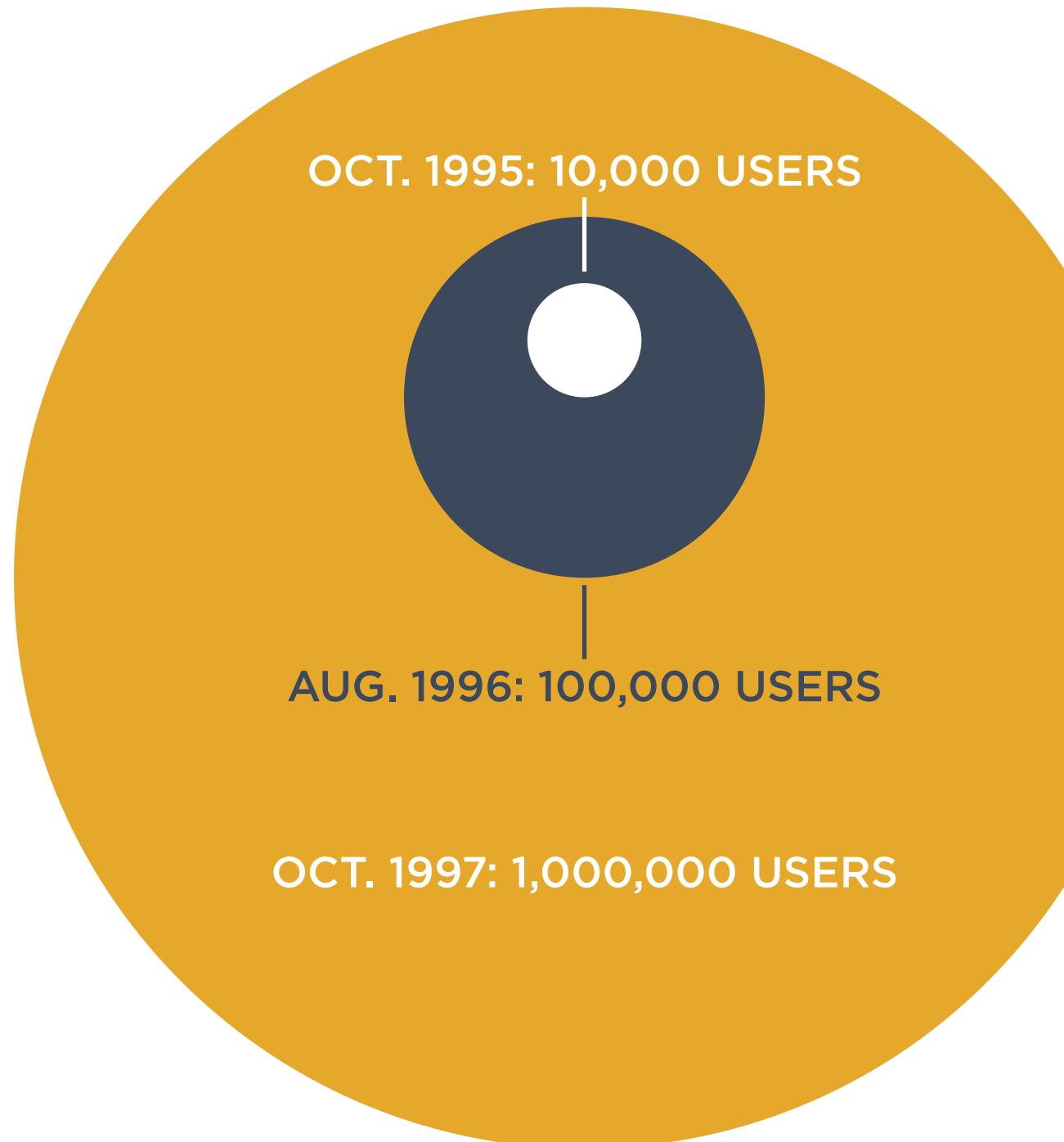
The screenshot shows a web browser window with the URL <https://web.archive.org/web/19961022173245/http://www.geocities.com/>. The page title is "Welcome to GeoCities Home". The main content area features the "Wayback Machine" logo and a banner stating "1,669 captures" from "22 Oct 96 – 13 Oct 14". Below the banner, there's a green text message: "TechWire just got more reporters...more news, and of course, it just got a whole lot better. You should come see what all the talk is about." To the right, there's a "GEOCLITIES" banner with neighborhood names like AREA 51, PARIS, HEARTLAND, ATHENS, and TIMES SQUARE. On the far right, there's a sidebar with links like "ENTER HERE", "INFORMATION", "NEIGHBORHOOD", "WHAT'S NEW", "WHAT'S COOL", and "WHAT IS GEOCITIES". 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?**

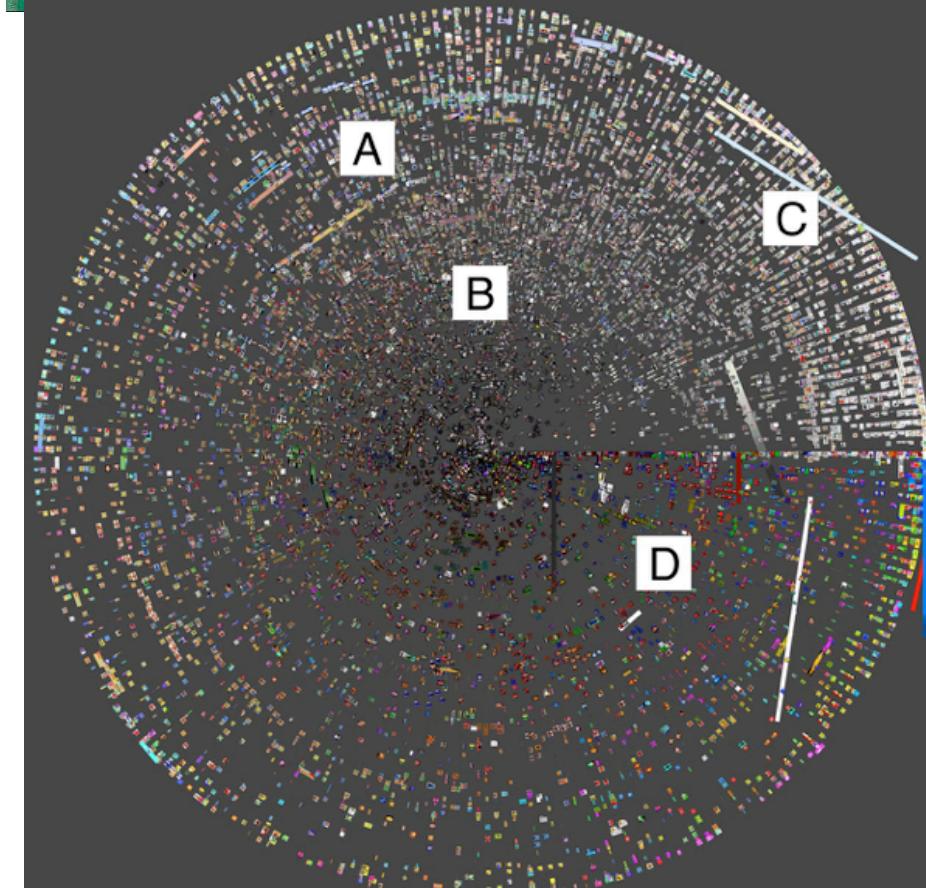
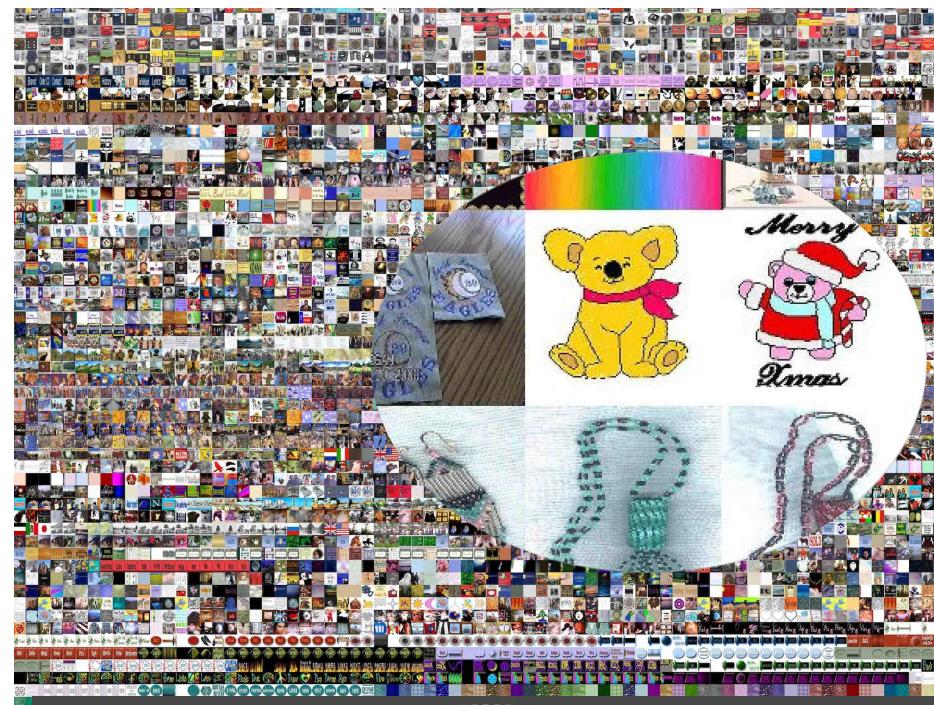
**Why does it  
matter?**



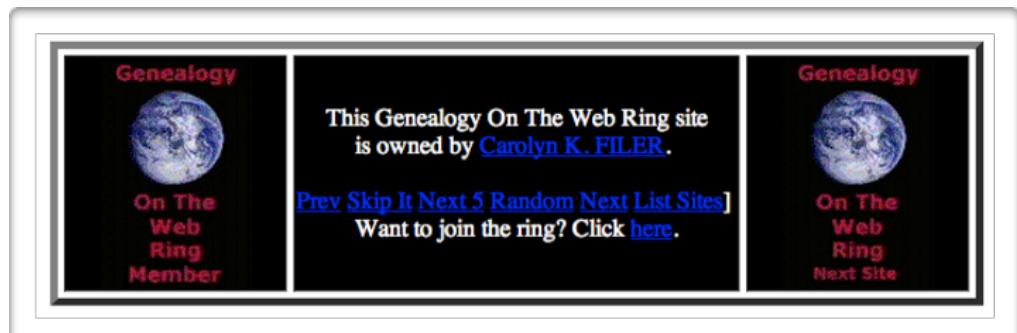
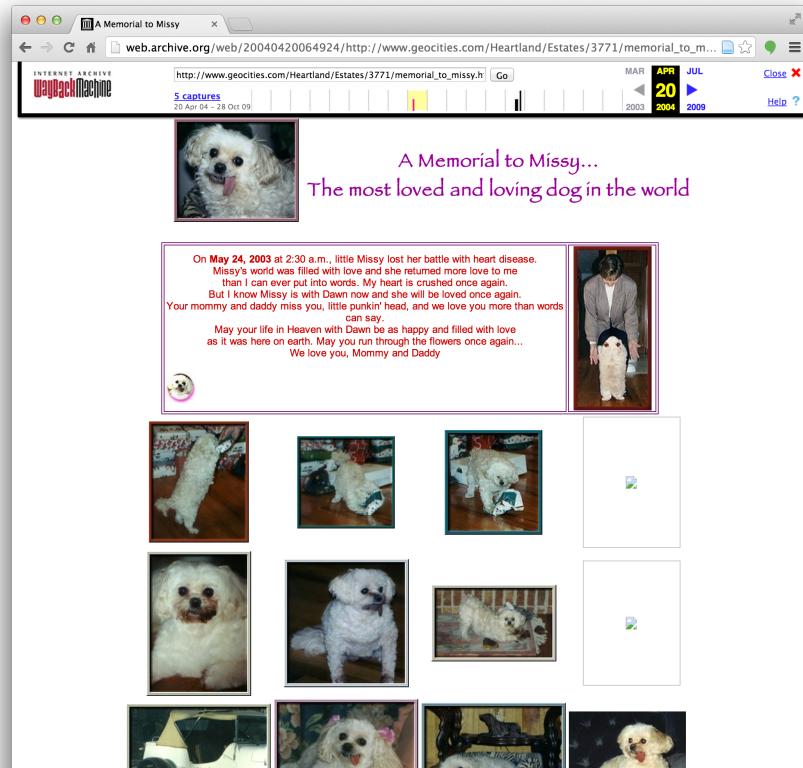
# Topic Modelling Community to Test Coherence

| Selected Neighbourhoods  | Top Two Topics  |
|--|---|
| Athens<br><i>“... based on education, teaching, reading, writing and philosophy”.</i>  | people things time person sense life man work world h soul make nature body case made point<br><br>part parts goddess witch healing incense witchcraft lov shaman witches sun spirit protection light circle earth        |
| EnchantedForest<br><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<br><br>jq battalion show st jonny horse battery armored lt artillery camp sailor army field col pingu w                     |
| Heartland<br><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 year 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<br><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<br><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<br><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 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.

**Warcbase**

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.

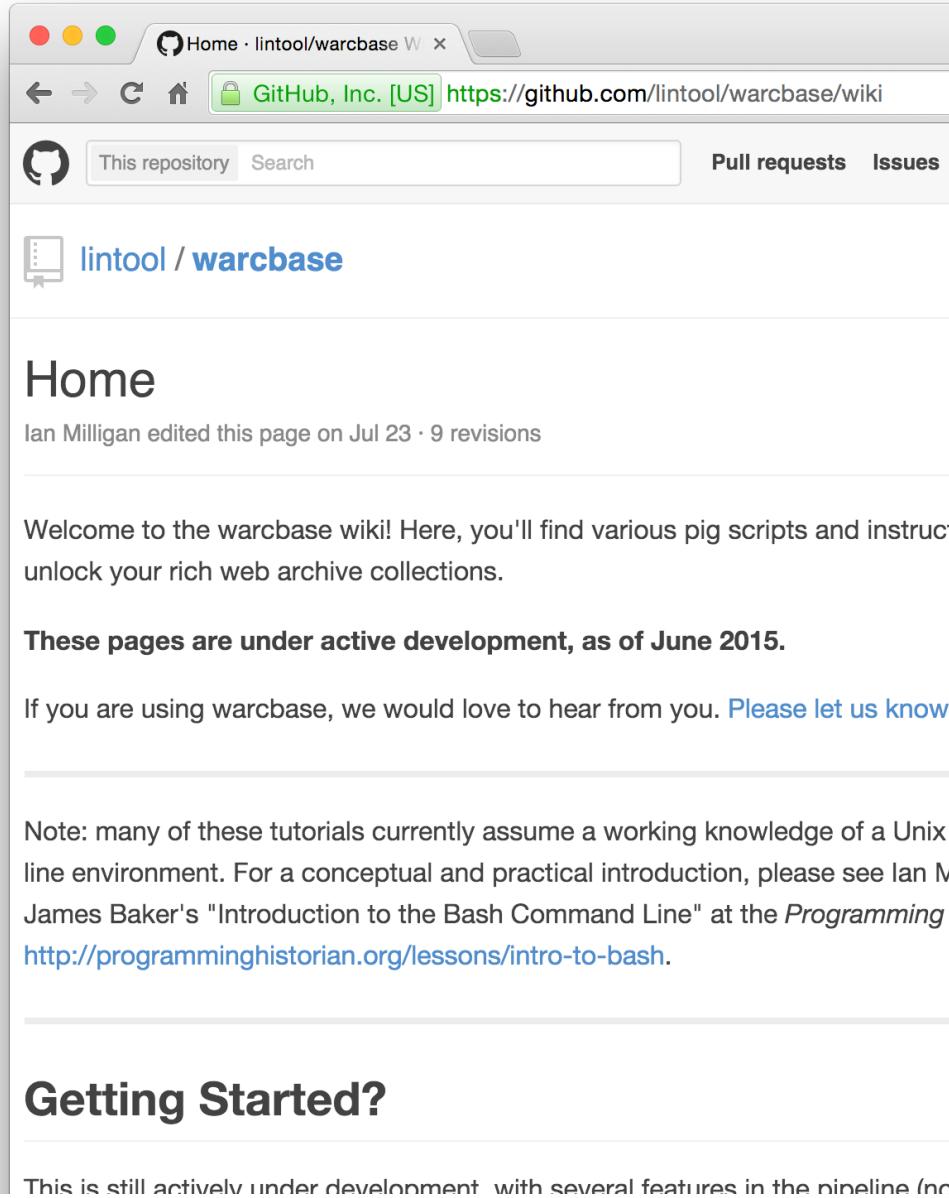
**Getting Started**

Clone the repo:

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 Mac OS X window displaying a GitHub wiki page. The title bar says "Home · lintool/warcbase". The address bar shows "GitHub, Inc. [US] https://github.com/lintool/warcbase/wiki". The main content area is titled "warcbase" and has a "Home" section. It includes a note from Ian Milligan about editing the page on Jul 23. Below that is a welcome message about the warcbase wiki and its purpose. A note states that the pages are under active development as of June 2015. There's also a note about Unix line environments and a link to a Bash Command Line introduction. At the bottom, there's a "Getting Started?" section and a footer note about the page being actively developed.

Home · lintool/warcbase

GitHub, Inc. [US] <https://github.com/lintool/warcbase/wiki>

This repository Search Pull requests Issues

## warcbase

### Home

Ian Milligan edited this page on Jul 23 · 9 revisions

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

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

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

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

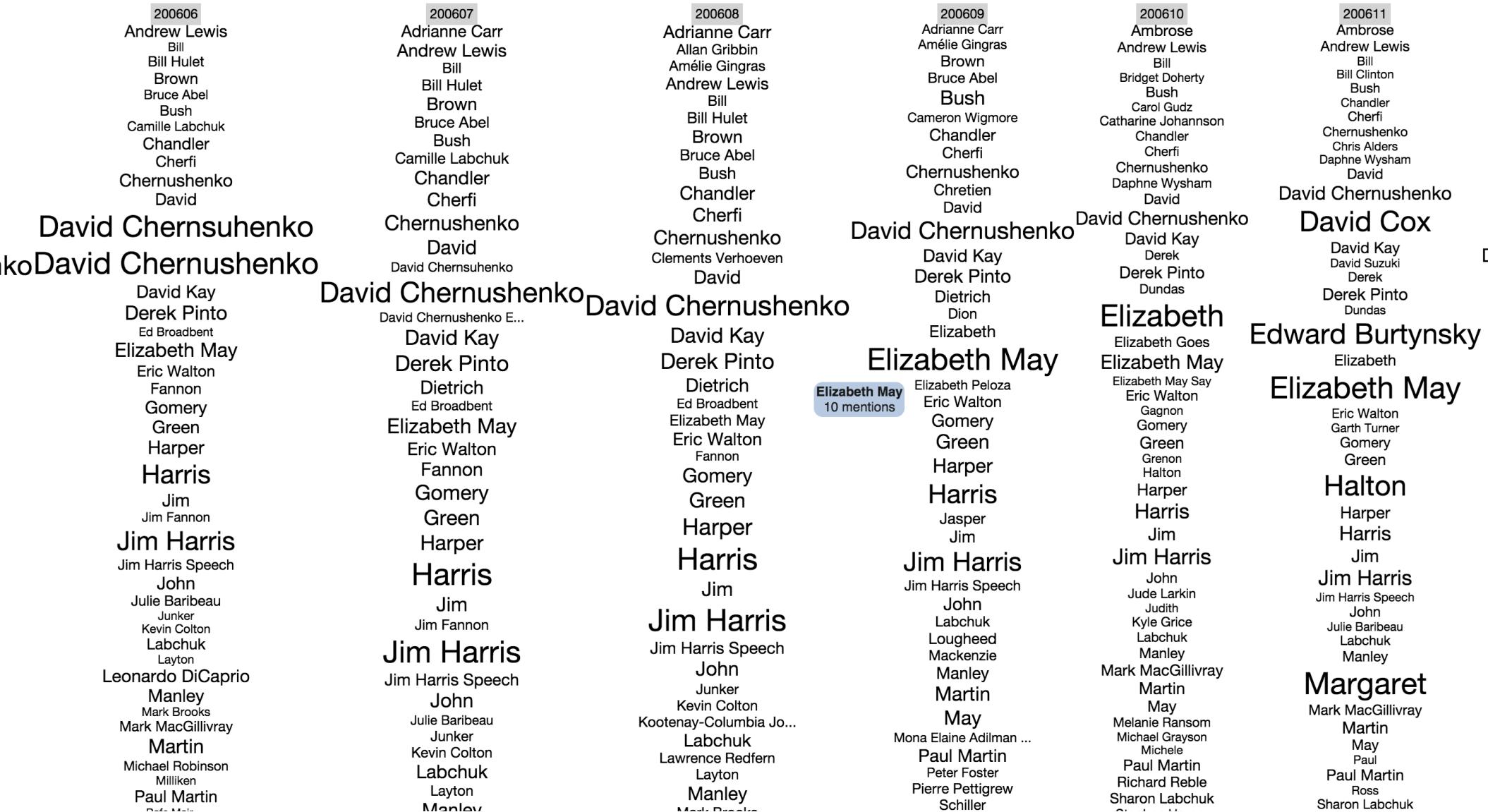
### Getting Started?

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

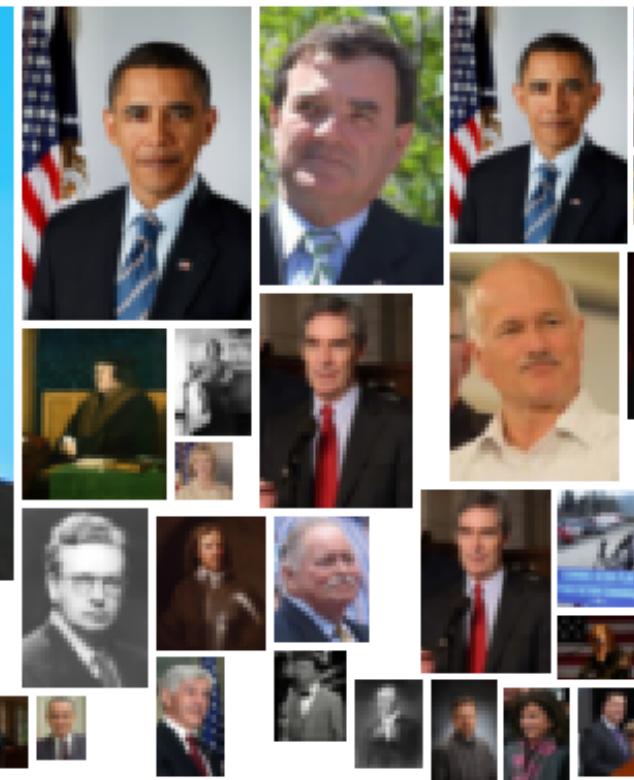
# Extract all Plain Text

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

# Extract Entities



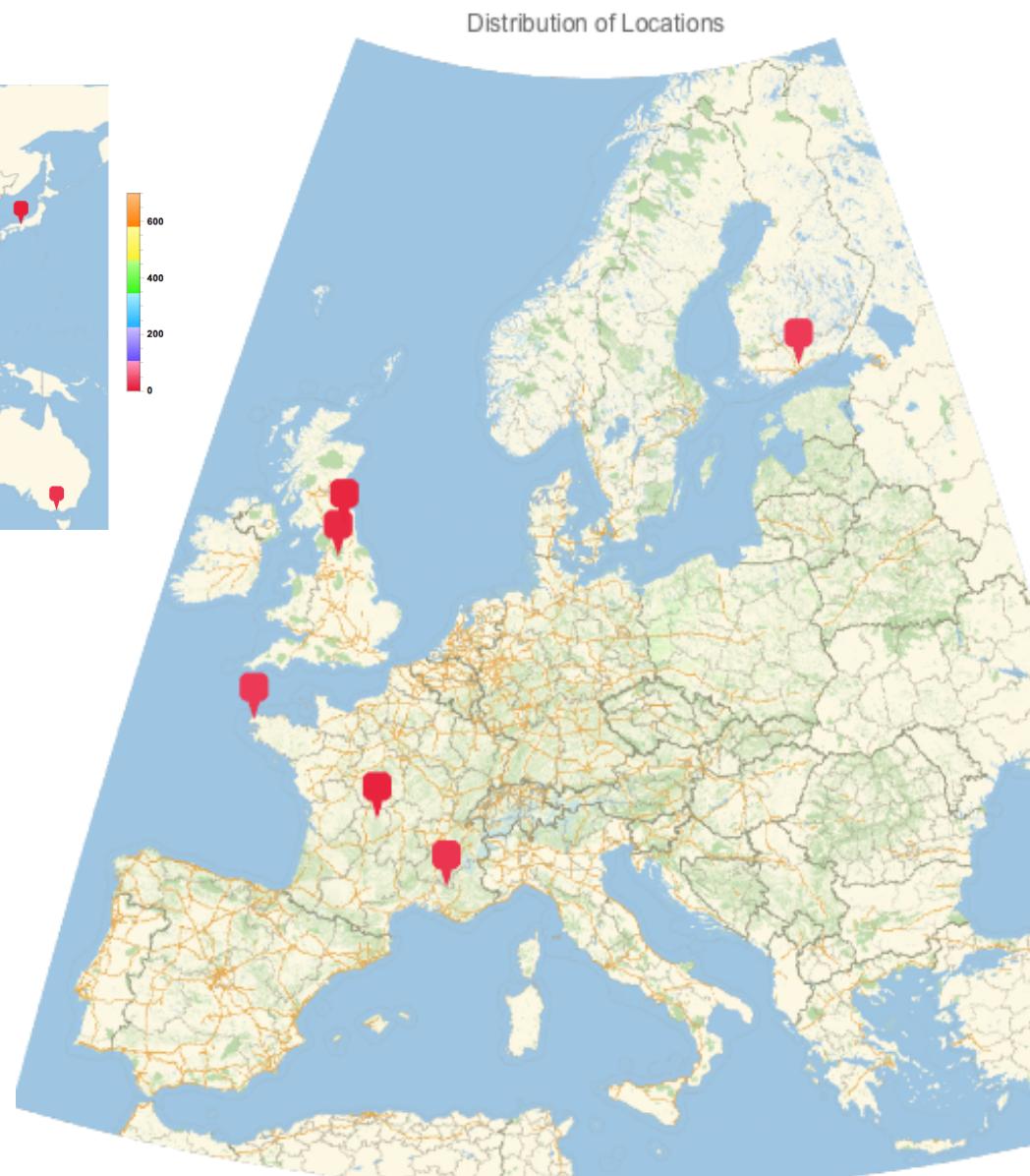
# Extract Entities



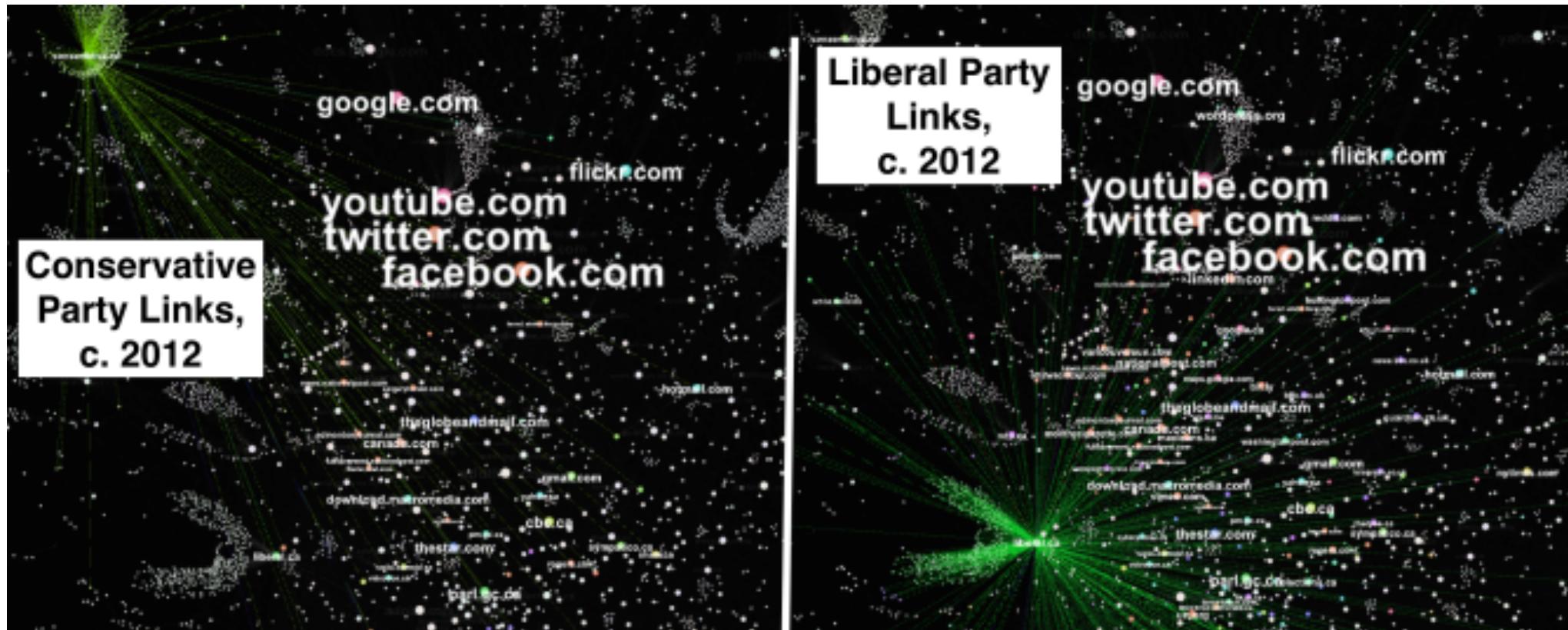
# Extract Entities



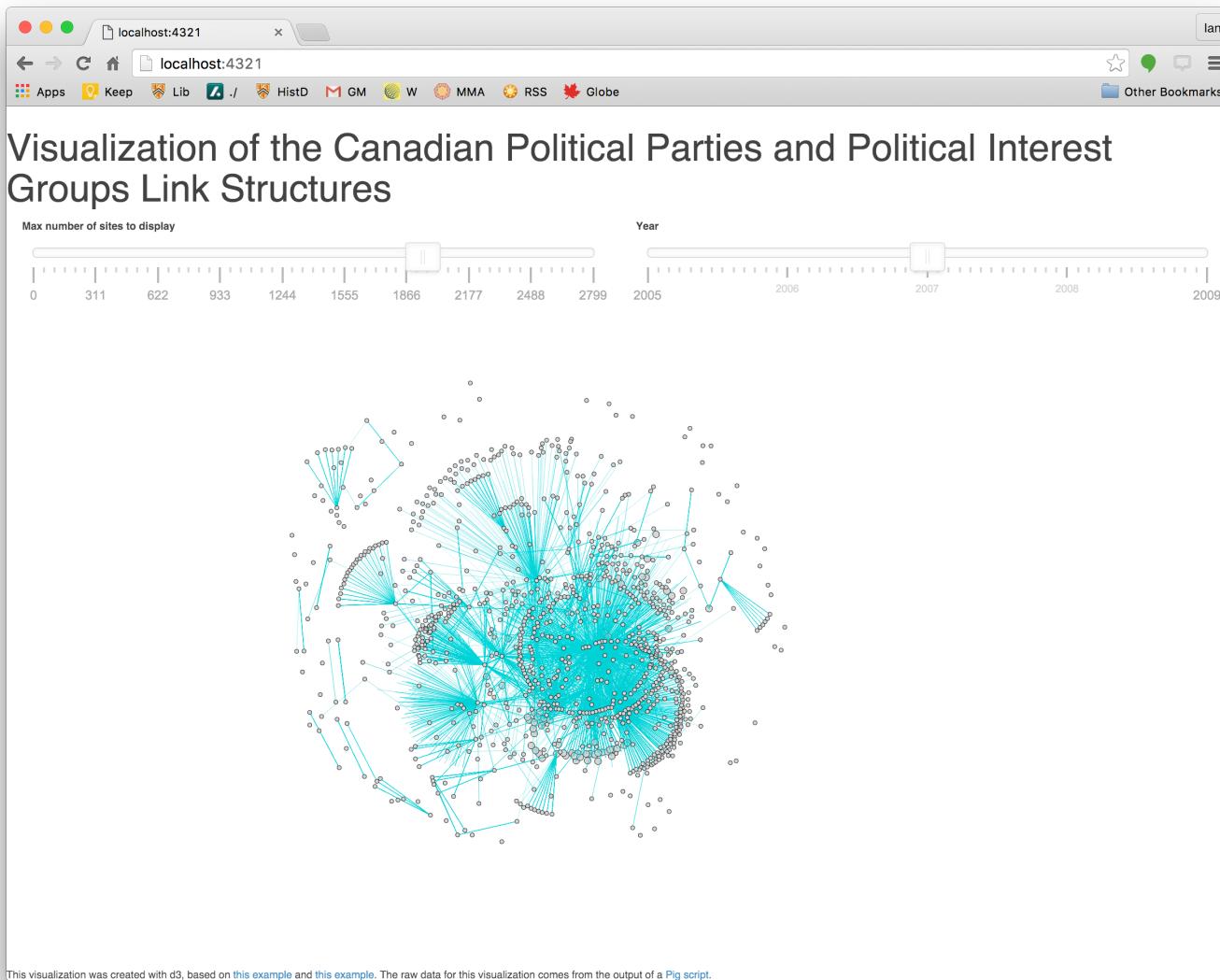
```
In[95]:= int = SemanticInterpretation[#] & /@ processedfreq[[All, 1]]  
Out[95]= {Canada, Calgary, Colombia, Montreal, $Failed, Ontario, Canada, Afghanistan,  
Ottawa, Manitoba, Canada, British Columbia, Canada, Toronto, Nova Scotia, Canada,  
Saskatchewan, Canada, Quebec, Canada, Alberta, Canada, Winnipeg, Washington,  
Canada, Nunavut, Canada, White Horse, United States, Petawawa, Mumbai,  
Lima, The Americas, Saskatoon, Peru, Edmonton, Mexico, Chicoutimi-Jonquiere,  
New Brunswick, Canada, United States, Newfoundland and Labrador, Canada, Qandahar,  
$Failed, Asia-Pacific, Newfoundland and Labrador, Canada, Victoria, United States,  
Quebec City, India, $Failed, Atlantic Ocean, St. Germain, Durham, Battle of Waterloo,
```



# Extract 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