

The Challenge of Digital Sources in the Web Age

**Common Tensions Across Three Web Histories,
1994-2015**

Ian Milligan
Assistant Professor



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 F...
--	---	--	--	---



199

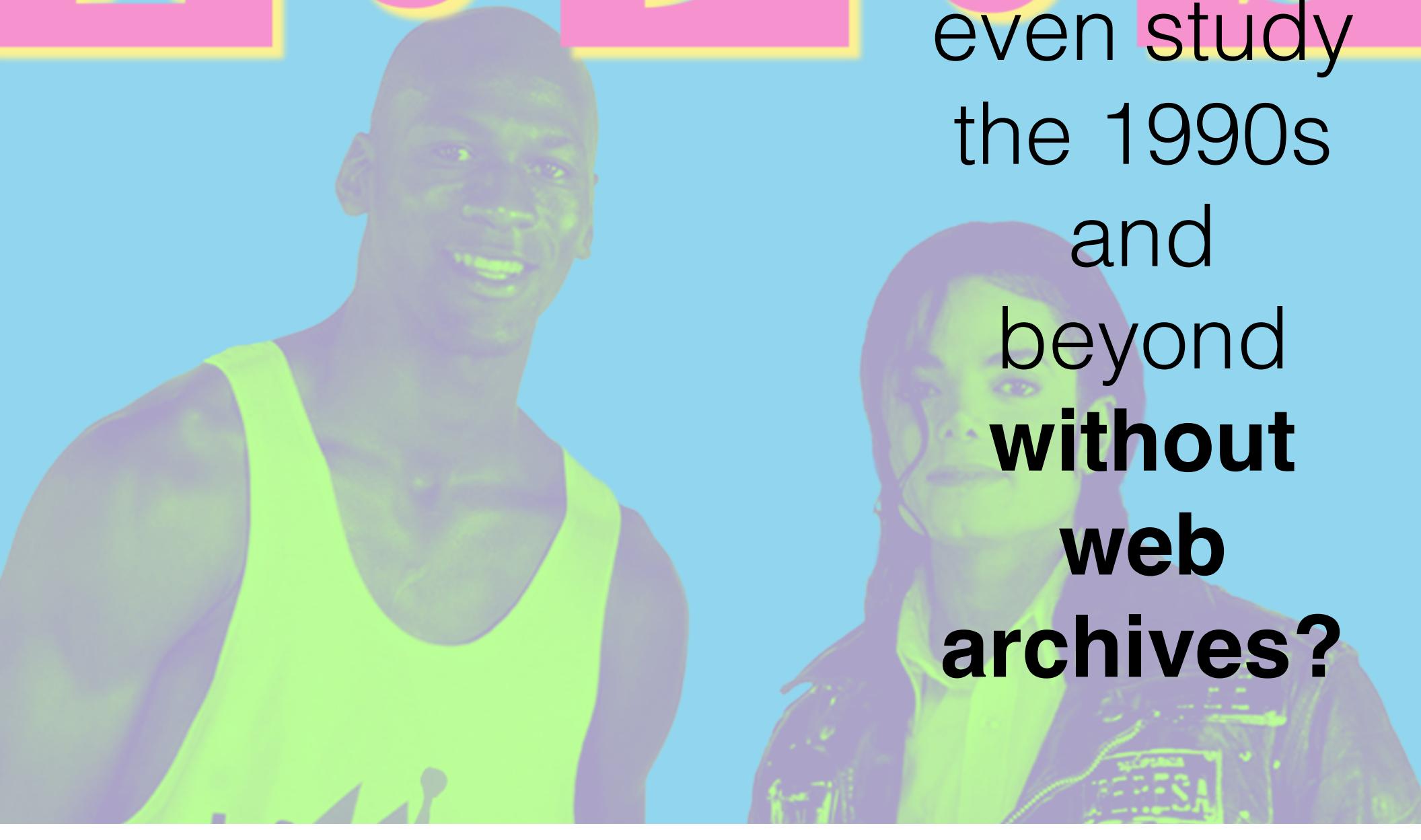
99

99

0

S

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 (similar to digitized newspapers);
- Historians are completely left out of post-1996 research, letting everybody else do the work (a la Culturomics project/*Nature* magazine article);
- Our profession gets left behind...

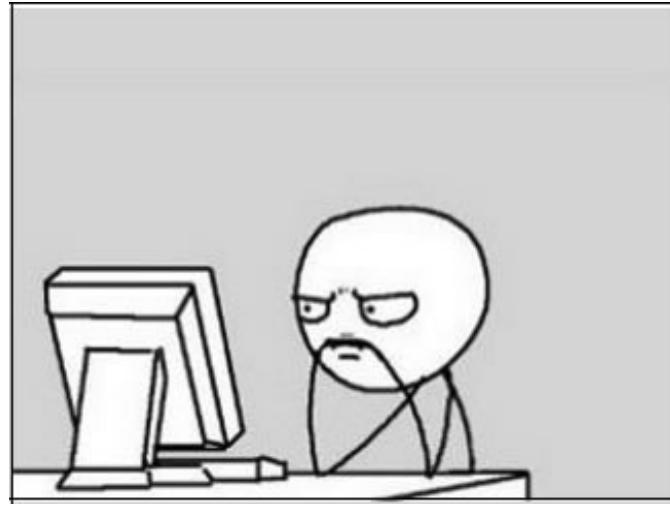
But what will web archives look like?

- 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**, 2005-2015 (Archive-It/University of Toronto)

Similarities -

Windows into the lives of
everyday people.





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

Case Study One

- The **Wide Web Scrape** (~ 80TB) - Snapshot of the Web
- **85,570** WARC files, CDX metadata
- Similar in some ways to traditional humanistic inquiry, just on a bigger scale.

The screenshot shows a Mac OS X desktop with a dark-themed browser window open. The title bar reads "Wide Crawl started March". The address bar shows the URL <https://archive.org/details/wide00002&tab=about>. The browser's menu bar includes "File", "Edit", "View", "Search", "Bookmarks", "History", "Help", and "Archive.org". Below the menu bar are various icons for file operations like Open, Save, Print, and Find. The main content area displays a page titled "Wide Crawl started March 2011". The page header includes a small thumbnail image of a classical building and the text "Wide Crawl started March 2011". Below the header, there is a brief 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." followed by a "MORE" link. At the top of the page content, there are three navigation links: "About" (which is underlined), "Collection", and "Forum". The "About" section contains a detailed description of the crawl, including its start date (09 March, 2011), end date (23 December, 2011), number of captures (2,713,676,341), number of unique URLs (2,273,840,159), and number of hosts (29,032,069). It also describes the seed list as a list of Alexa's top 1 million web 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 named "ARossi Archivist". Below this, there are sections for "ADDITIONAL CONTRIBUTOR" (with a profile picture of a man named "brewste Archivist") and "VIEWS" (with a profile picture of a person named "kngenie Archivist").

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%
TOTAL	41,542,172	1,687,664,620	2.46%

CDX Files (finding aids)



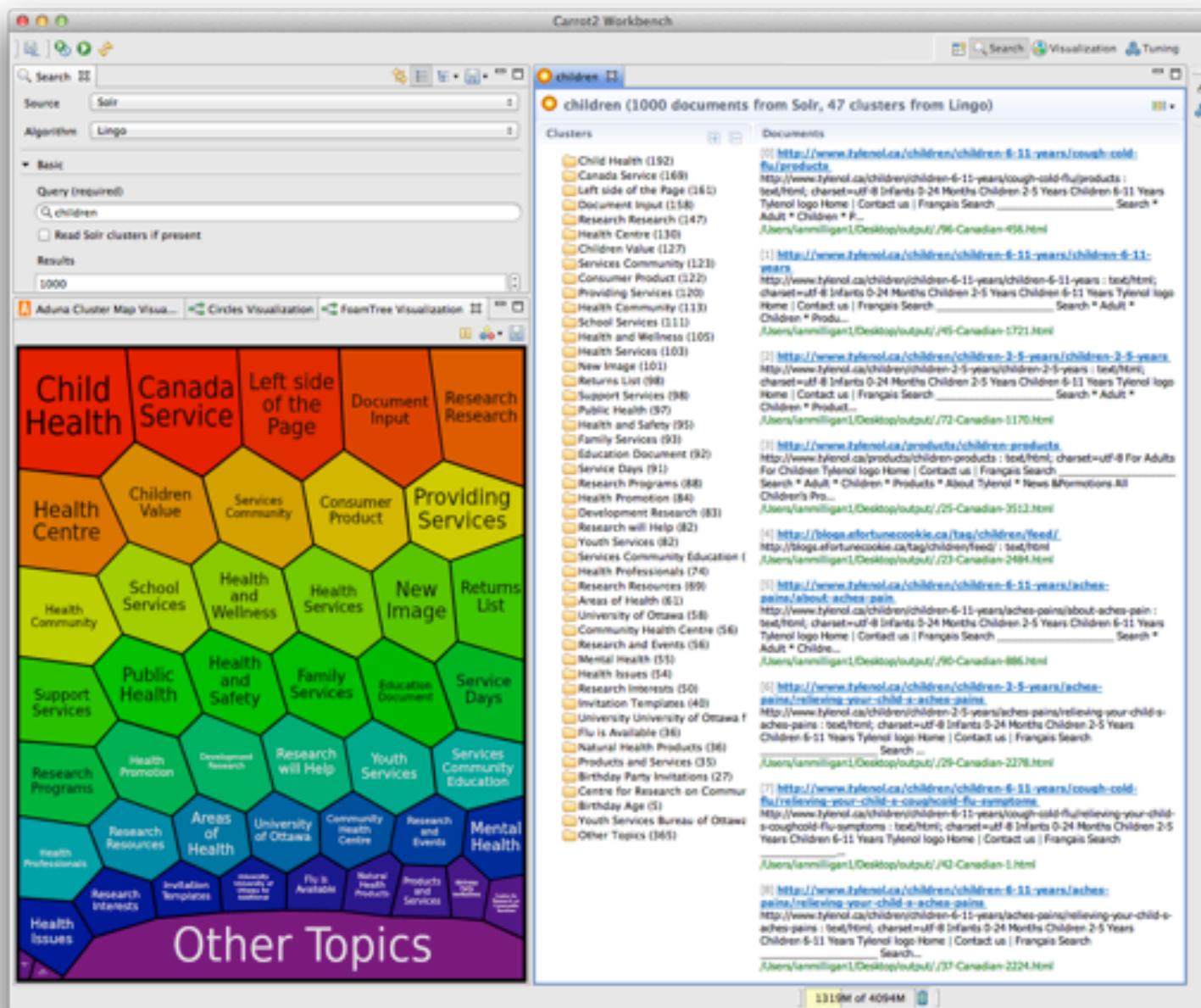
WARC File

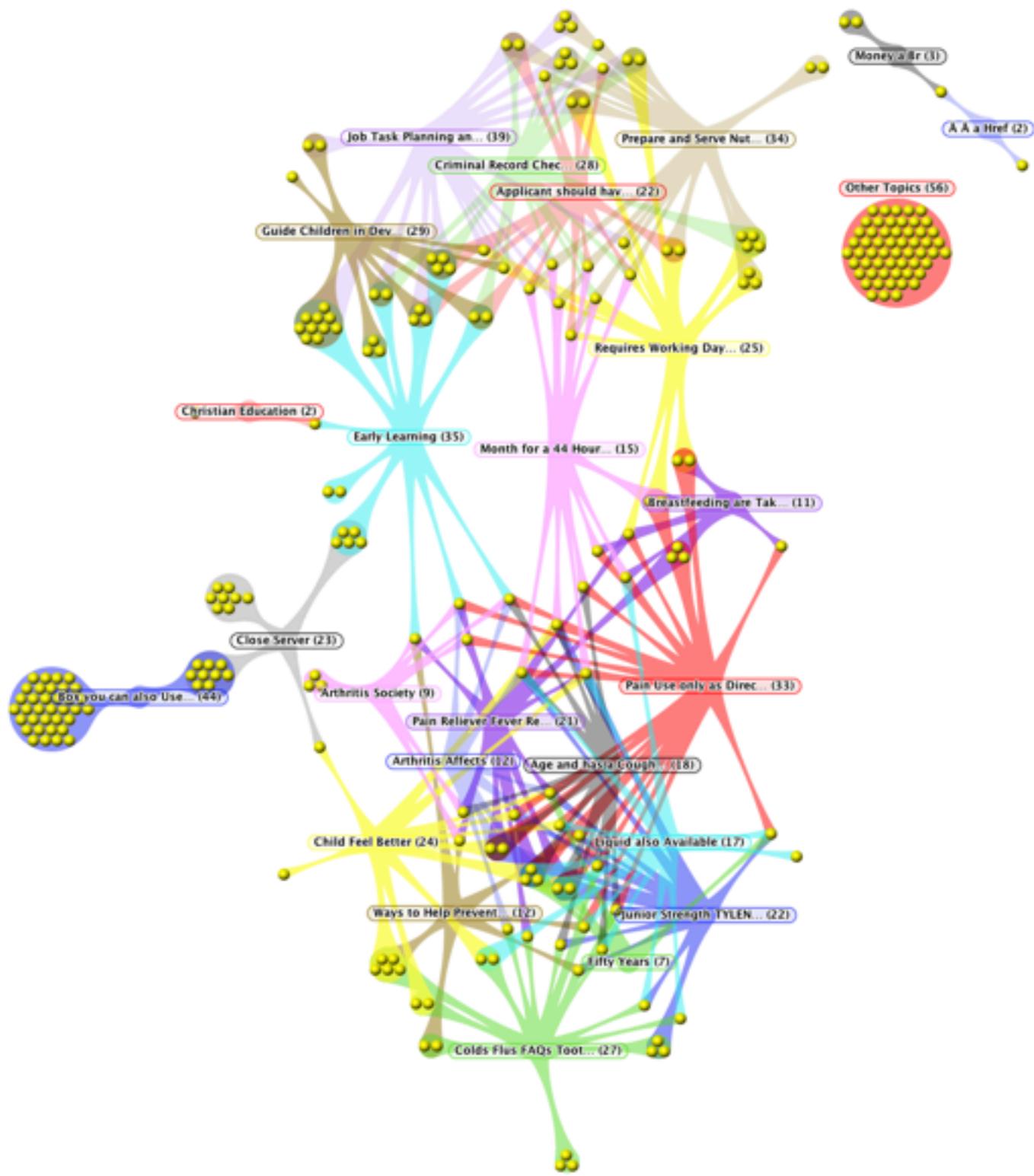
WARC-Tools/Lynx
(`warcfilter.py`,
`warchtmlindex.py`
and `filesdump.py`)

A screenshot of a terminal window titled "2. hmv". The window displays the output of the command "warcfilter.py" on a WARC file. The output lists numerous URLs, primarily from the Internet Archive, such as "http://www.archive.org/details/ianmilligan_0001of0001" and "http://www.archive.org/details/ianmilligan_0002of0001". The terminal also shows the command "Screen Shot 2014-02-06 at 11:47:16 AM" and help text for the Lynx browser.



Indexing





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.

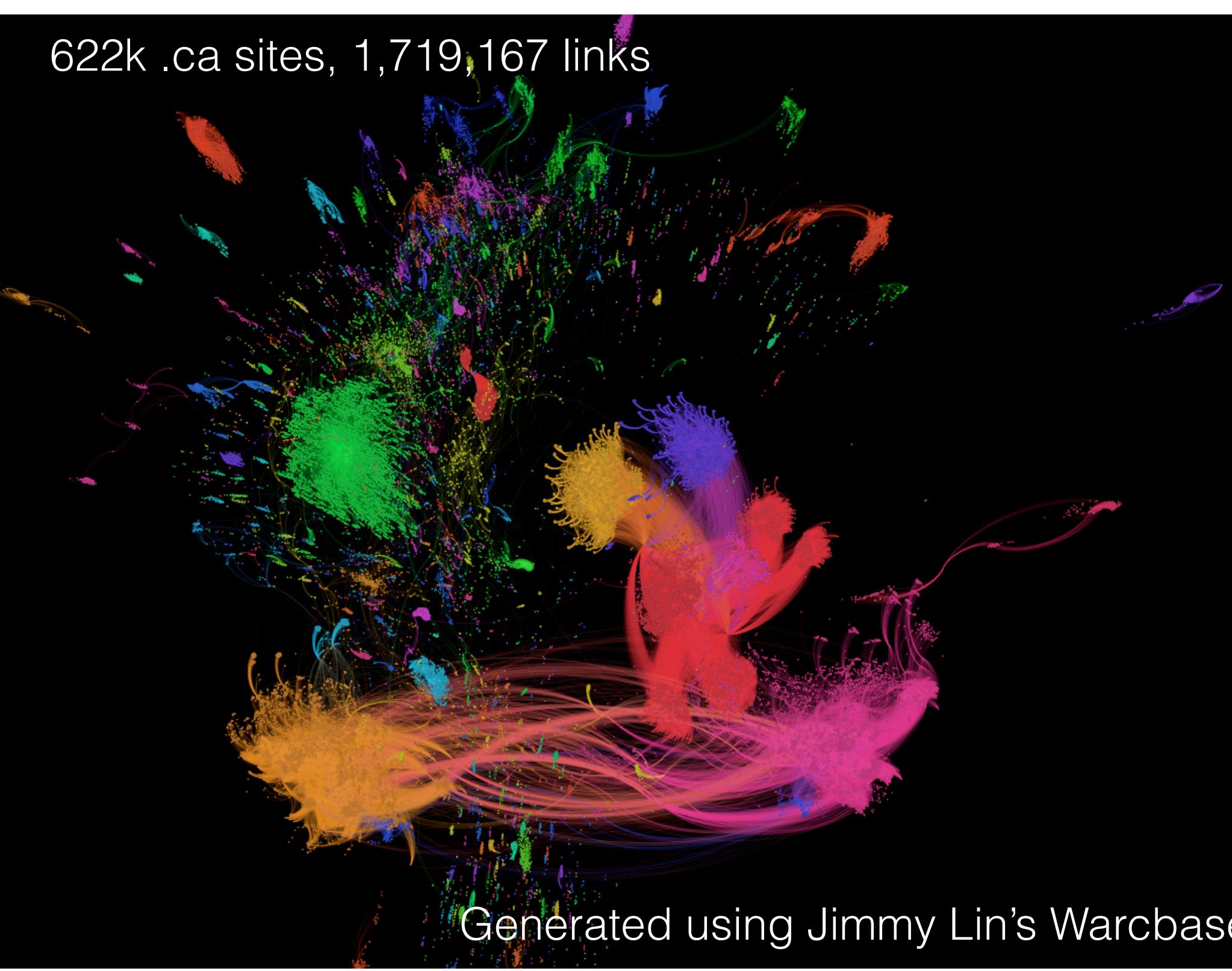
Robert Granke, executive director of CLWR, visited congregations of the ILSN in November. You can read more about his visit at www.lccontheroad.ca, The Canadian Lutheran or in the forthcoming issue of CLWR's Partnership newsletter due out in early February.



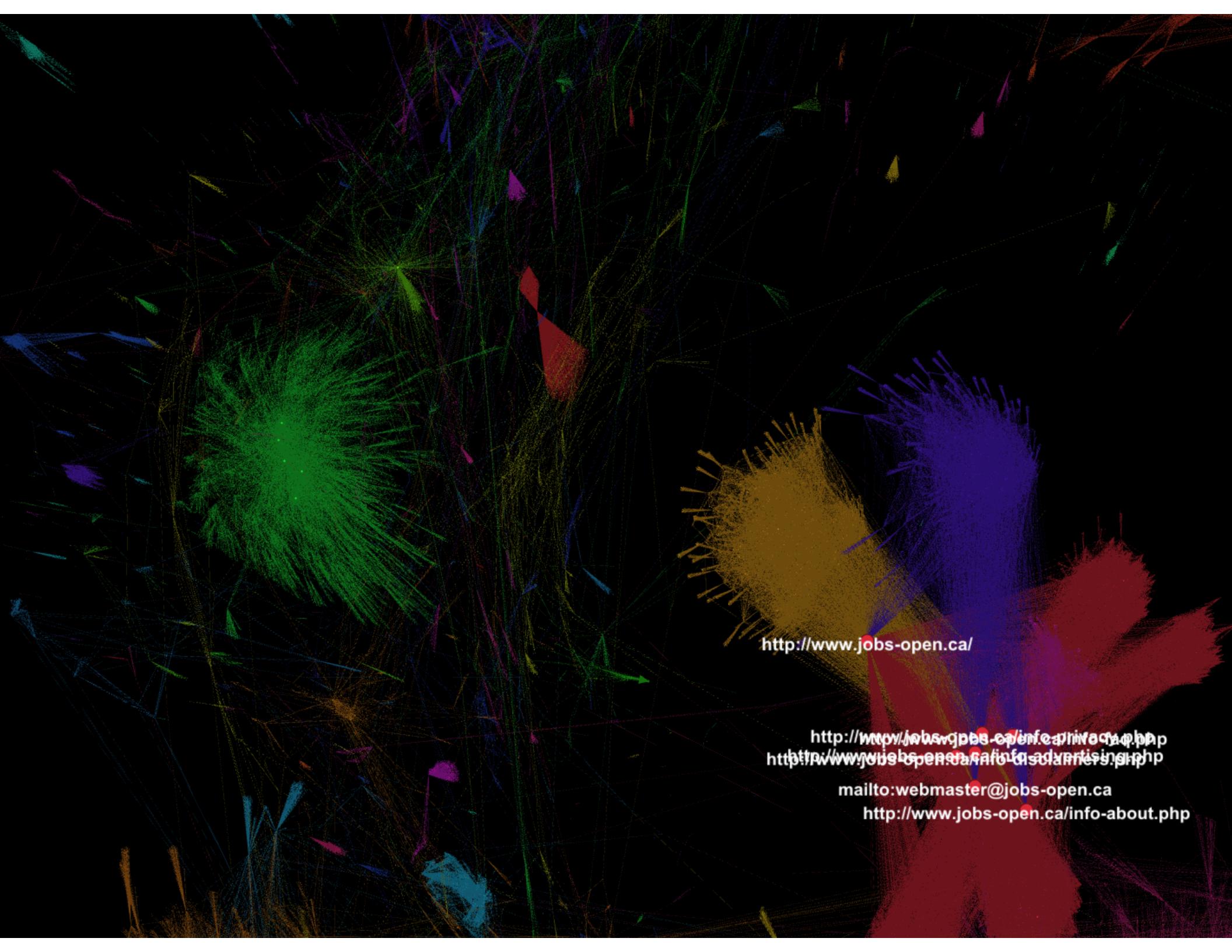
A medical clinic in Nicaragua.

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

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

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

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

<https://web3.uottawa.ca/infoweb/logon/fr.html>

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

[sourcesfinancieres.uottawa.ca/etudiant/payment-university-fees-fr.php](http://www.uottawa.ca/academics/info/reHS/CTS/programmes.htm?H=2565)
<http://www.uottawa.ca/academics/info/reHS/CTS/programmes.htm?H=2565>

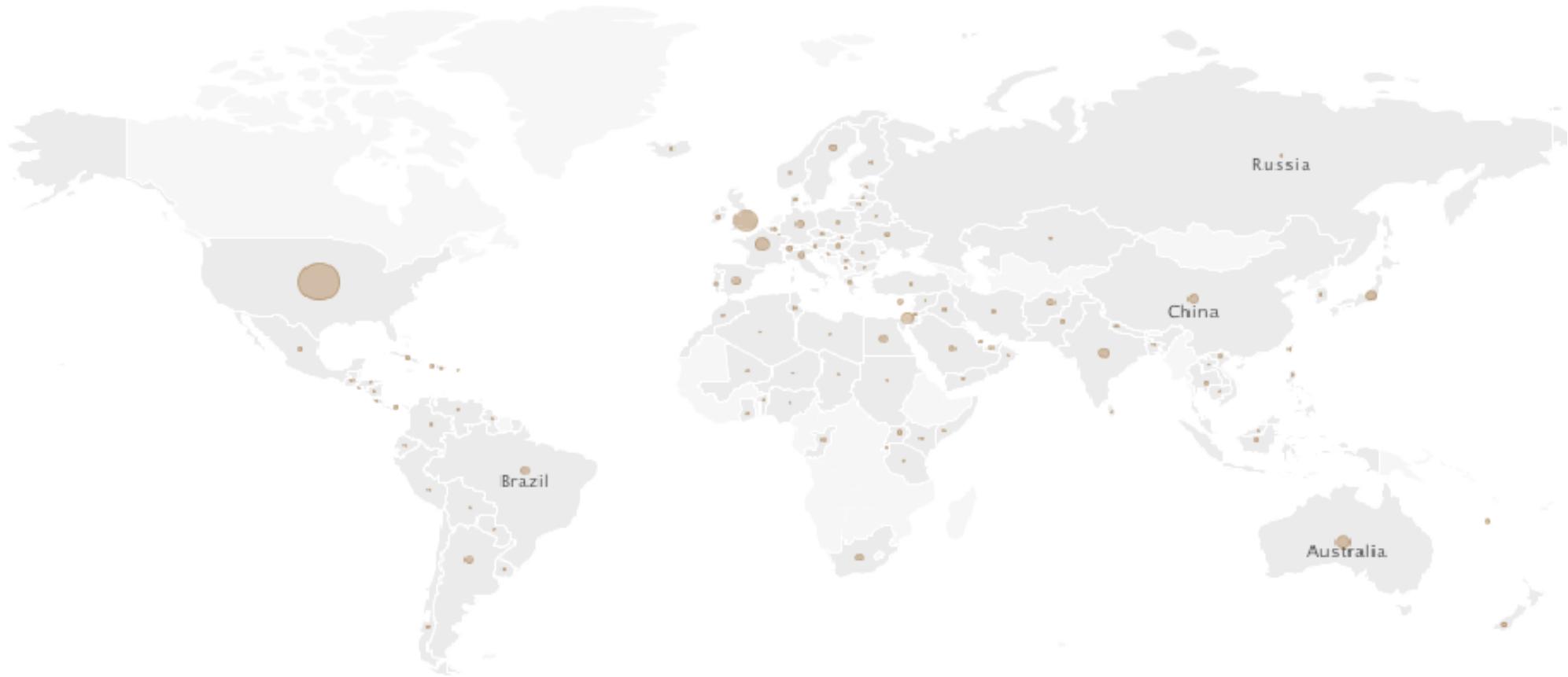
<http://www.admission.uottawa.ca/Default.aspx?tabid=2548&source=Eqp>

<http://www.adminmission.com> <http://www.pw5354.com/getitfree.htm>

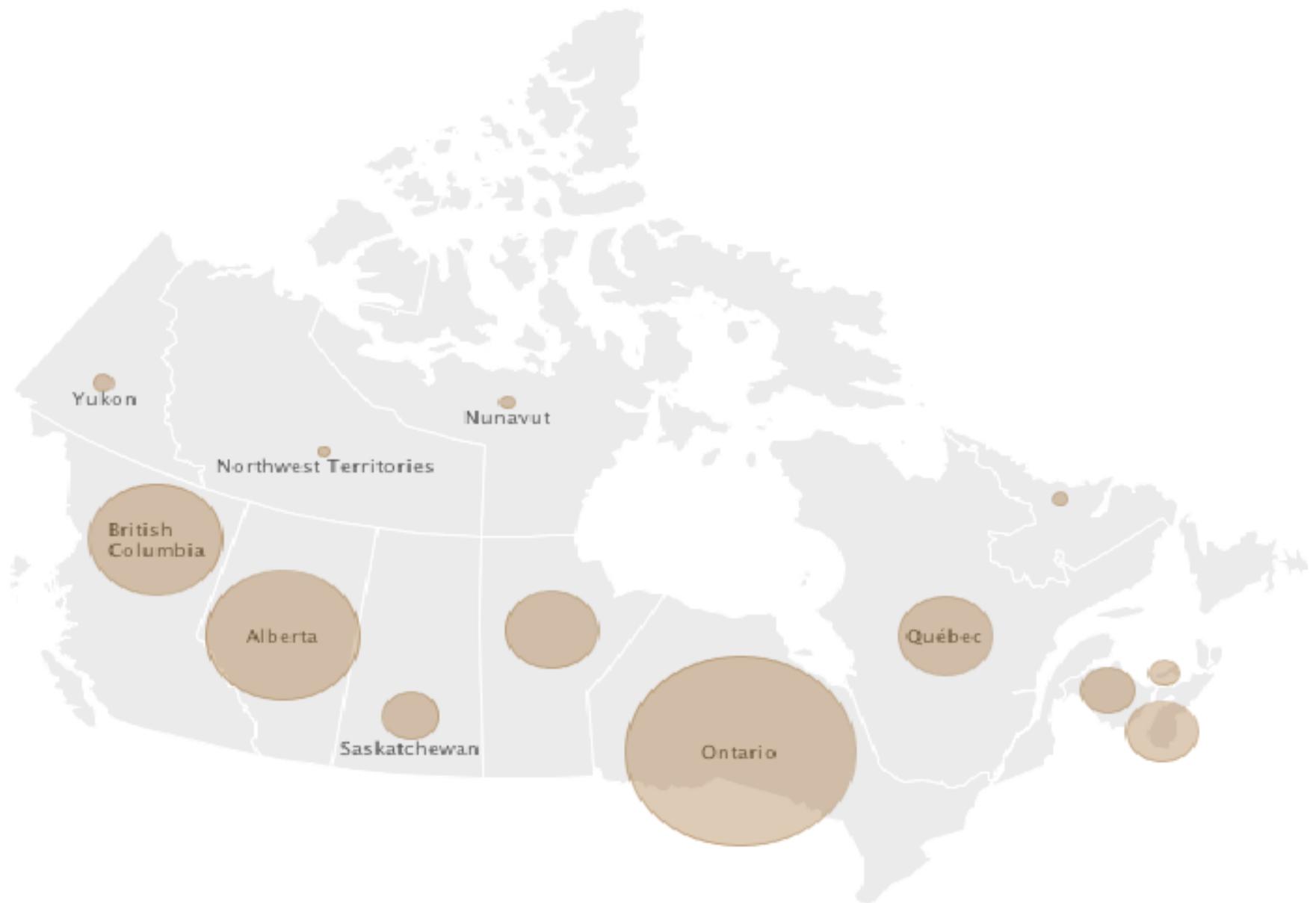
<https://web9.uottawa.ca/study/studies/university/arts/Default.aspx?tabid=1736>

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

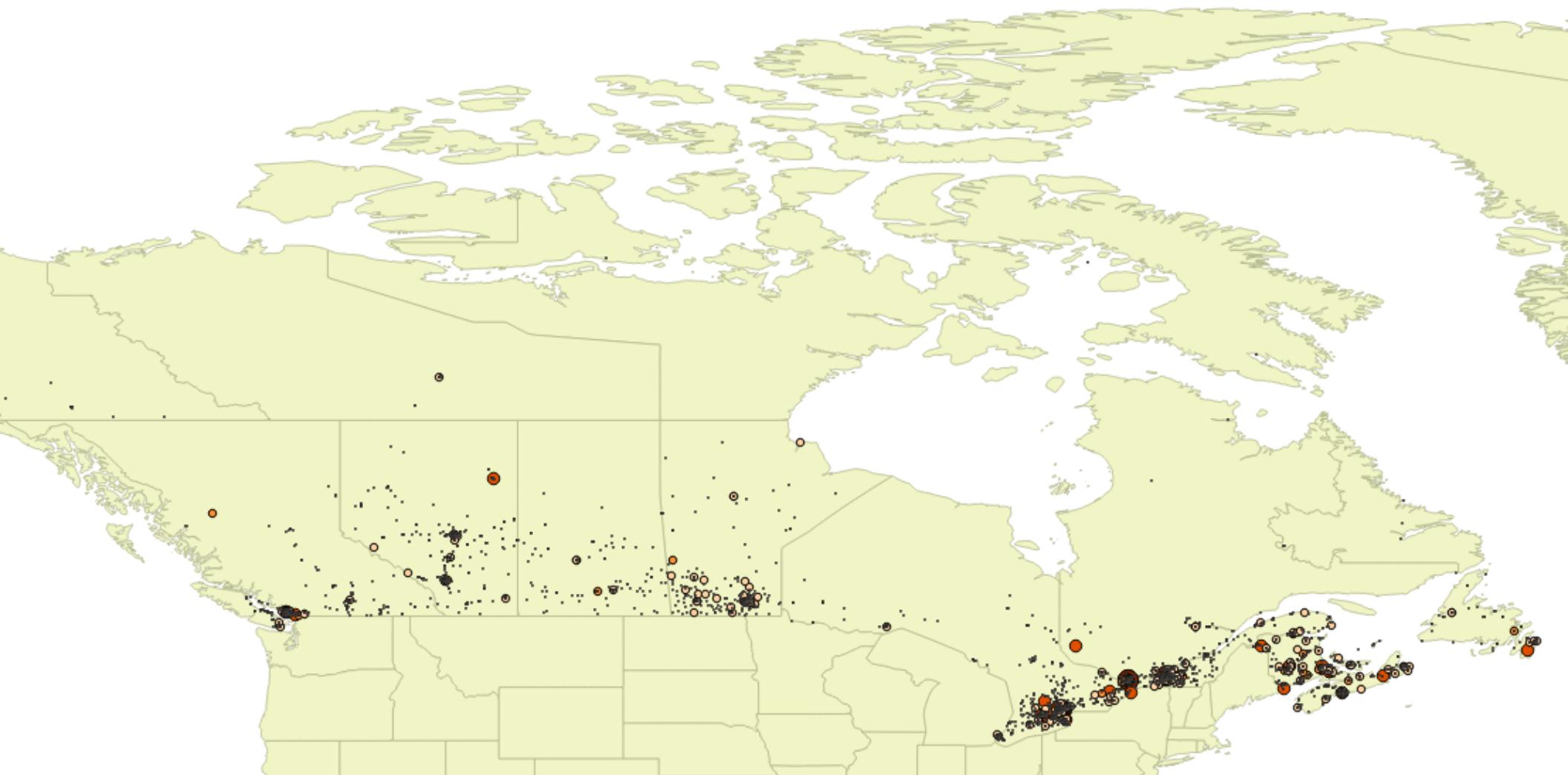
Countries Mentioned in .ca TLD (excluding Canada)



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” and “Canadian Labour Unions.”
- 2005 - 2015
- WAT & WARC 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 main content area displays the "Canadian Political Parties and Political Interest Groups" collection, which was collected by the University of Toronto and archived since October 2005. The collection includes a thumbnail image of the Archive-It logo, the title, collector information, and details about the scope and subject. Below this, there are sections for "Narrow Your Results" and a search bar. At the bottom, there are buttons for "Sites" and "Search Page Text", and a footer indicating "Page 1 of 1 (54 Total)".

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.

How to provide
access?

WAT Files?

Potential sweet spot between
the lightweight CDX and the
heavy-duty WARC?

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

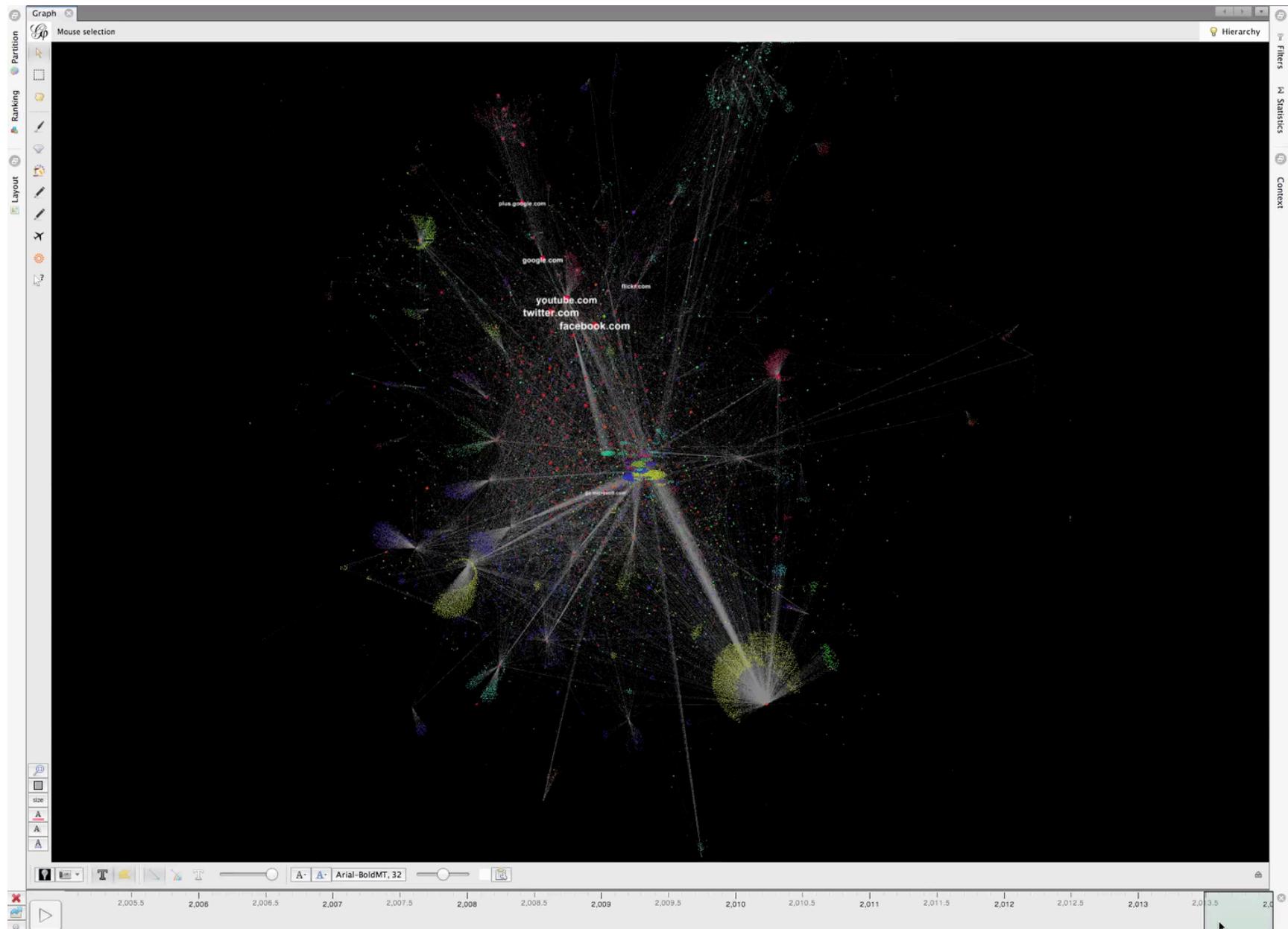
Two problems

Problem One:

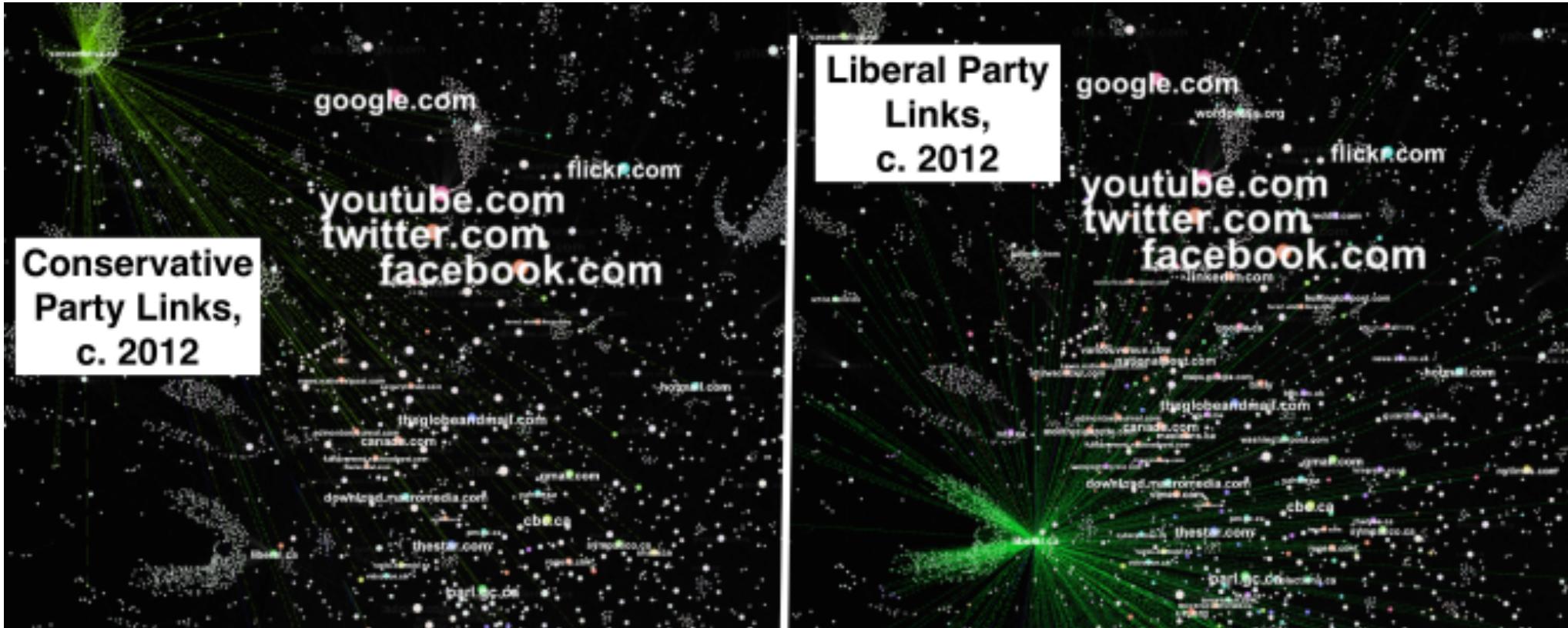
Historians want to work with content, but we can only use metadata on most computer systems.

(but that's ok - we can
use metadata to do great
work)

Metadata Extraction



Metadata Extraction



Metadata Extraction

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

colinbarriemp.ca	12
colinbarriemp.ca&lang=fr	12
colinmayes.ca	12
colinmayes.ca&lang=fr	12
congrespcc.ca	12
conservateur.ca	12
conservateur.us5.list-manage.com	12
conservative.ca	12
conservative.us5.list-manage.com	12
consumersfirst.ca	12
corneliuchisu.ca	12
corneliuchisu.ca&lang=fr	12
costasmenegakis.ca	12
costasmenegakis.ca&lang=fr	12
cpcconvention.ca	12

Metadata Extraction

- 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

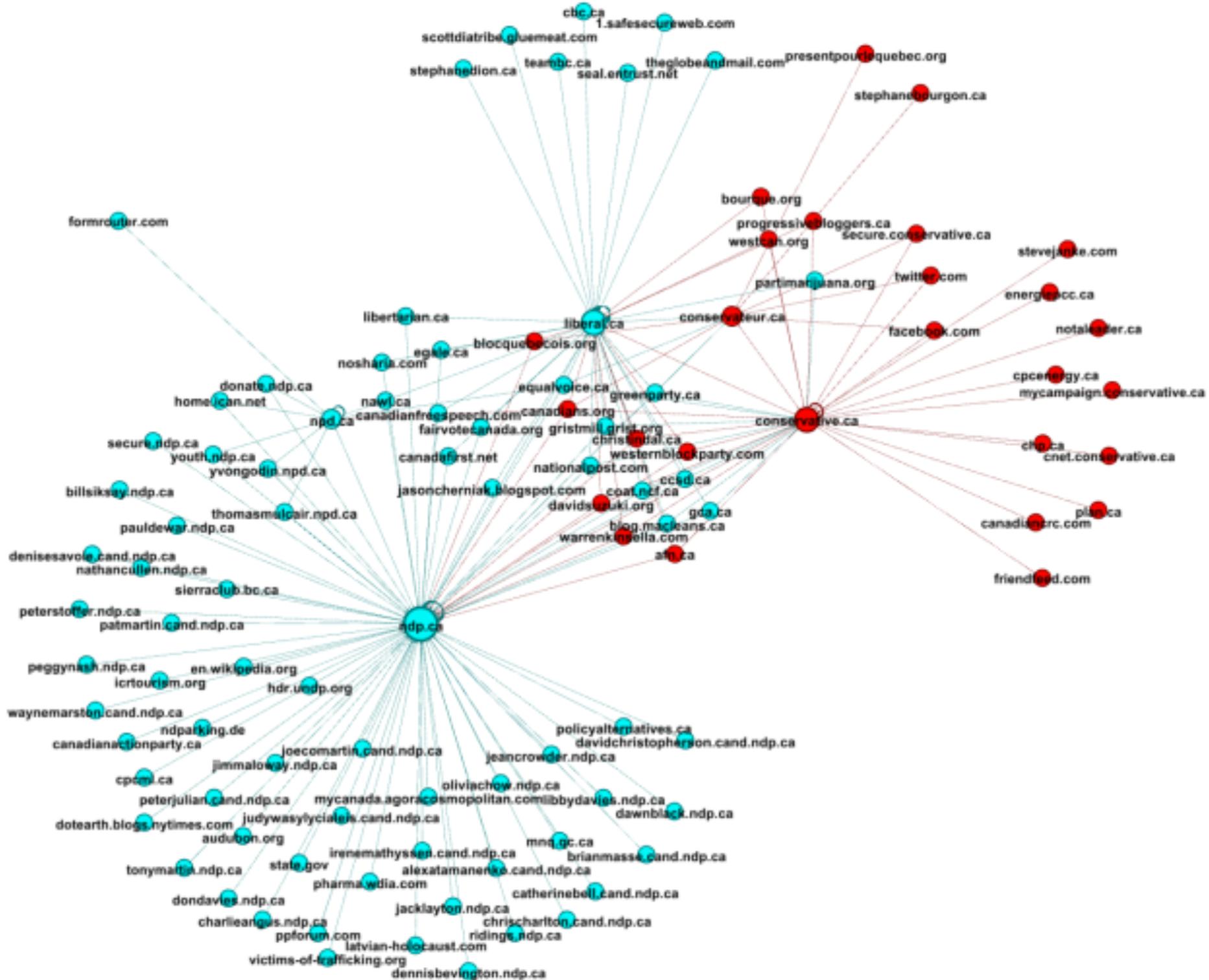
Election Campaign - Advertisement Sites



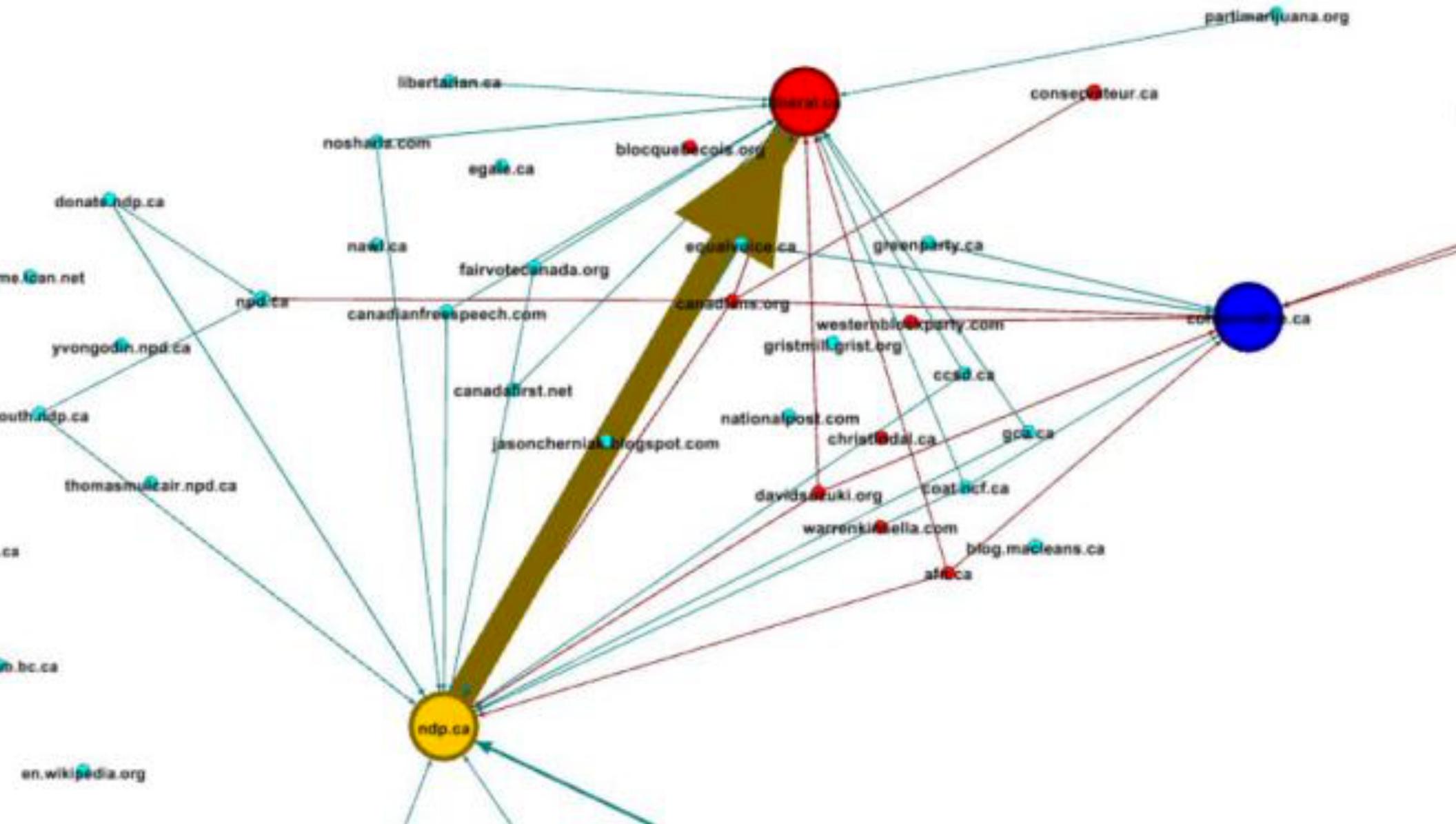
December 2008

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





2005 Canadian Federal Election



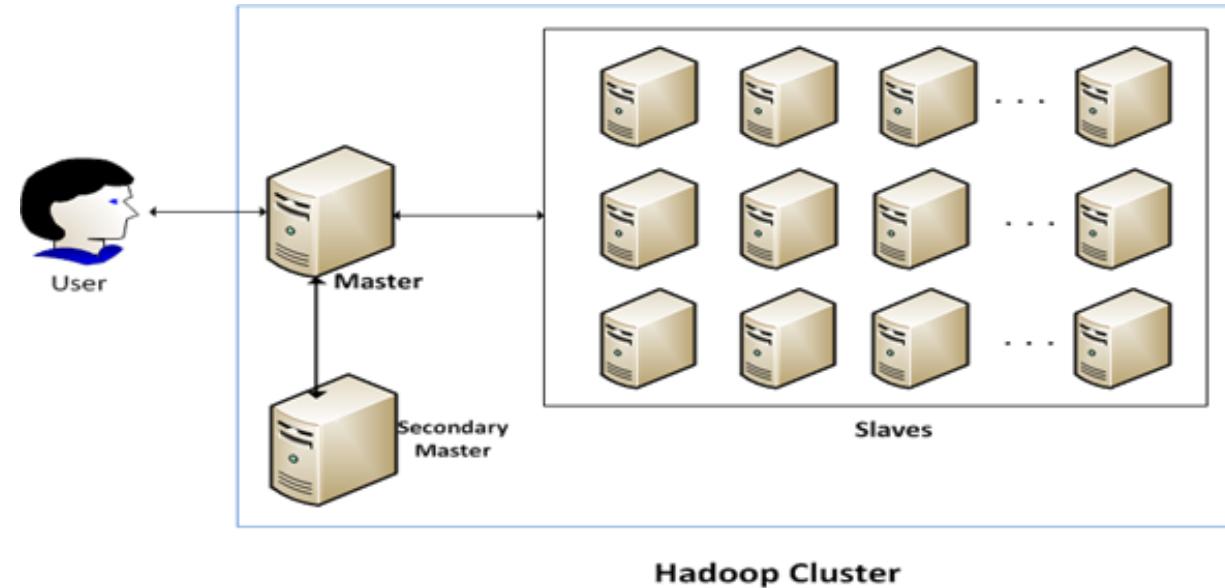


WATs help us find the files
we need to use - and to
contextualize them



Problem Two:

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



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



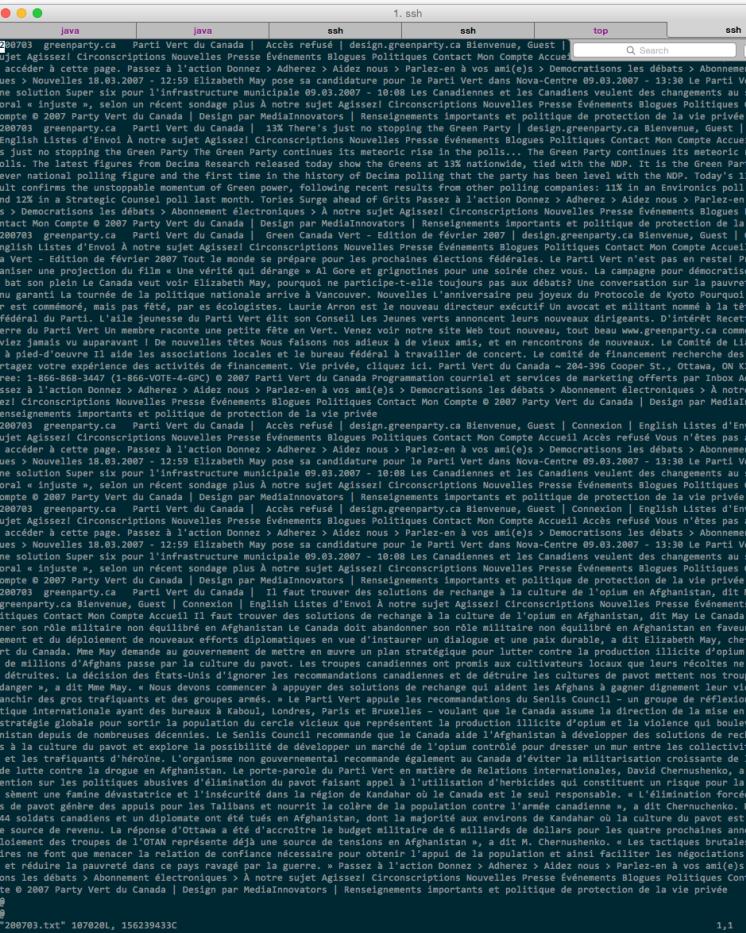
Large-Scale Text Analysis

- 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

```
java          java          ssh          ssh
200703 greenparty.ca Parti Vert du Canada | Accès refusé | design.greenparty.ca Bienvenue
ujet Agissez! Circonscriptions Nouvelles Presse Événements Blogues Politiques Contact Mon Com
accéder à cette page. Passez à l'action Donnez > Adherez > Aidez nous > Parlez-en à vos ami
ues > Nouvelles 18.03.2007 - 12:59 Elizabeth May pose sa candidature pour le Parti Vert dans
ne solution Super six pour l'infrastructure municipale 09.03.2007 - 10:08 Les Canadiens et
oral « injuste », selon un récent sondage plus À notre sujet Agissez! Circonscriptions Nouvel
ompte © 2007 Party Vert du Canada | Design par MediaInnovators | Renseignements importants et
200703 greenparty.ca Parti Vert du Canada | 13% There's just no stopping the Green Party
English Listes d'Envoi À notre sujet Agissez! Circonscriptions Nouvelles Presse Événements Bl
s just no stopping the Green Party The Green Party continues its meteoric rise in the polls..
olls. The latest figures from Decima Research released today show the Greens at 13% nationwide
ever national polling figure and the first time in the history of Decima polling that the part
ult confirms the unstoppable momentum of Green power, following recent results from other poll
nd 12% in a Strategic Counsel poll last month. Tories Surge ahead of Grits Passez à l'action
s > Democratisons les débats > Abonnement électroniques > À notre sujet Agissez! Circonscrip
ntact Mon Compte © 2007 Party Vert du Canada | Design par MediaInnovators | Renseignements im
200703 greenparty.ca Parti Vert du Canada | Green Canada Vert - Edition de février 2007 |
nglish Listes d'Envoi À notre sujet Agissez! Circonscriptions Nouvelles Presse Événements Bl
a Vert - Edition de février 2007 Tout le monde se prépare pour les prochaines élections fédé
aniser une projection du film « Une vérité qui dérange » Al Gore et grignotines pour une soin
bat son plein Le Canada veut voir Elizabeth May, pourquoi ne participe-t-elle toujours pas au
nu garanti La tournée de la politique nationale arrive à Vancouver. Nouvelles L'anniversaire
r est commémoré, mais pas fêté, par les écologistes. Laurie Arron est le nouveau directeur exé
fédéral du Parti. L'aile jeunesse du Parti Vert élit son Conseil Les Jeunes verts annoncent l
erre du Parti Vert Un membre raconte une petite fête en Vert. Venez voir notre site Web tout
viez jamais vu auparavant ! De nouvelles têtes Nous faisons nos adieux à de vieux amis, et en
à pied-d'œuvre Il aide les associations locales et le bureau fédéral à travailler de concert
rtagez votre expérience des activités de financement. Vie privée, cliquez ici. Parti Vert du
rée: 1-866-868-3447 (1-866-VOTE-4-GPC) © 2007 Parti Vert du Canada Programmation courriel et
ssez à l'action Donnez > Adherez > Aidez nous > Parlez-en à vos ami(e)s > Democratisons les
ez! Circonscriptions Nouvelles Presse Événements Blogues Politiques Contact Mon Compte © 2007
enseignements importants et politique de protection de la vie privée
200703 greenparty.ca Parti Vert du Canada | Accès refusé | design.greenparty.ca Bienvenu
ujet Agissez! Circonscriptions Nouvelles Presse Événements Blogues Politiques Contact Mon Com
accéder à cette page. Passez à l'action Donnez > Adherez > Aidez nous > Parlez-en à vos ami
ues > Nouvelles 18.03.2007 - 12:59 Elizabeth May pose sa candidature pour le Parti Vert dans
ne solution Super six pour l'infrastructure municipale 09.03.2007 - 10:08 Les Canadiens et
oral « injuste », selon un récent sondage plus À notre sujet Agissez! Circonscriptions Nouvel
ompte © 2007 Party Vert du Canada | Design par MediaInnovators | Renseignements importants et
200703 greenparty.ca Parti Vert du Canada | Accès refusé | design.greenparty.ca Bienvenu
ujet Agissez! Circonscriptions Nouvelles Presse Événements Blogues Politiques Contact Mon Com
accéder à cette page. Passez à l'action Donnez > Adherez > Aidez nous > Parlez-en à vos ami
ues > Nouvelles 18.03.2007 - 12:59 Elizabeth May pose sa candidature pour le Parti Vert dans
ne solution Super six pour l'infrastructure municipale 09.03.2007 - 10:08 Les Canadiens et
oral « injuste », selon un récent sondage plus À notre sujet Agissez! Circonscriptions Nouvel
ompte © 2007 Party Vert du Canada | Design par MediaInnovators | Renseignements importants et
200703 greenparty.ca Parti Vert du Canada | Il faut trouver des solutions de rechange à l'Afghanistan
greenparty.ca Bienvenue, Guest | Connexion | English Listes d'Envoi À notre sujet Agissez! Cir
cuitiques Contact Mon Compte Accueil Il faut trouver des solutions de rechange à la culture de
nner son rôle militaire non équilibré en Afghanistan Le Canada doit abandonner son rôle militai
rement et du déploiement de nouveaux efforts diplomatiques en vue d'instaurer un dialogue et u
rt du Canada. Mme May demande au gouvernement de mettre en œuvre un plan stratégique pour lut
de millions d'Afghans passe par la culture du pavot. Les troupes canadiennes ont promis aux
détruites. La décision des États-Unis d'ignorer les recommandations canadiennes et de détruire
danger », a dit Mme May. « Nous devons commencer à appuyer des solutions de rechange qui aider
anchir des gros trafiquants et des groupes armés. » Le Parti Vert appuie les recommandations de
tique internationale ayant des bureaux à Kaboul, Londres, Paris et Bruxelles - voulant que la
stratégie globale pour sortir la population du cercle vicieux que représentent la production d'Afgh
nistan depuis de nombreuses décennies. Le Senlis Council recommande que le Canada aide l'Afghanist
s à la culture du pavot et explore la possibilité de développer un marché de l'opium contrôlé par
et les trafiquants d'héroïne. L'organisme non gouvernemental recommande également au Canada de
de lutte contre la drogue en Afghanistan. Le porte-parole du Parti Vert en matière de Relati
on sur les politiques abusives d'élimination du pavot faisant appel à l'utilisation d'heroin
sément une famine dévastatrice et l'insécurité dans la région de Kandahar où le Canada est l
s de pavot génère des appuis pour les Talibans et nourrit la colère de la population contre l'Afghan
44 soldats canadiens et un diplomate ont été tués en Afghanistan, dont la majorité aux environs
e source de revenu. La réponse d'Ottawa a été d'accroître le budget militaire de 6 milliards
lolement des troupes de l'OTAN représente déjà une source de tensions en Afghanistan », a dit
ires ne font que menacer la relation de confiance nécessaire pour obtenir l'appui de la popula
et réduire la pauvreté dans ce pays ravagé par la guerre. » Passez à l'action Donnez > Adh
ons les débats > Abonnement électroniques > À notre sujet Agissez! Circonscriptions Nouvelles
ez! Circonscriptions Nouvelles Presse Événements Blogues Politiques Contact Mon Compte © 2007
Party Vert du Canada | Design par MediaInnovators | Renseignements importants et politiques
@
@
"200703.txt" 107020L, 156239433C
```

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

A Return to Cooking

25 Top Recipes From
Top Atlanta Chefs

MELISSA LIBBY

atlanta cooks

A Return to Cooking

25 Top Recipes From
Top Atlanta Chefs

MELISSA LIBBY

Recipe Book Idea

michel richard happy in the kitchen

FRANK STITT'S BOTTEGA FAVORITA

My Beverly Hills Kitchen

COOKBOOK

BOULEVARD

COOKBOOK

MAKES AND MAZZOLA THE COOKBOOK

RED SAGE

THE BLACKBERRY FARM COOKBOOK

COOKBOOK

SAM BEALL THE JOYCE LaFRAY

Hearst Books

'Cuba Cocina! Joyce LaFray

COOKBOOK

Charlie Trotter's Desserts

HARPER

One&Only Palmilla SPACIUSINE by Charlie Trotter

HARPER

EFFORTLESS ELEGANCE WITH COLIN COWIE A W

TEN SPEED PRESS

LEMONGRASS AND LIME

ARTIZAN

RIPERT / RUHLMAN

MELISSA LIBBY

ARTIZAN

A Return to Cooking

25 Top Recipes From
Top Atlanta Chefs

MELISSA LIBBY

Using Warcbase to
analyze links and full-text

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

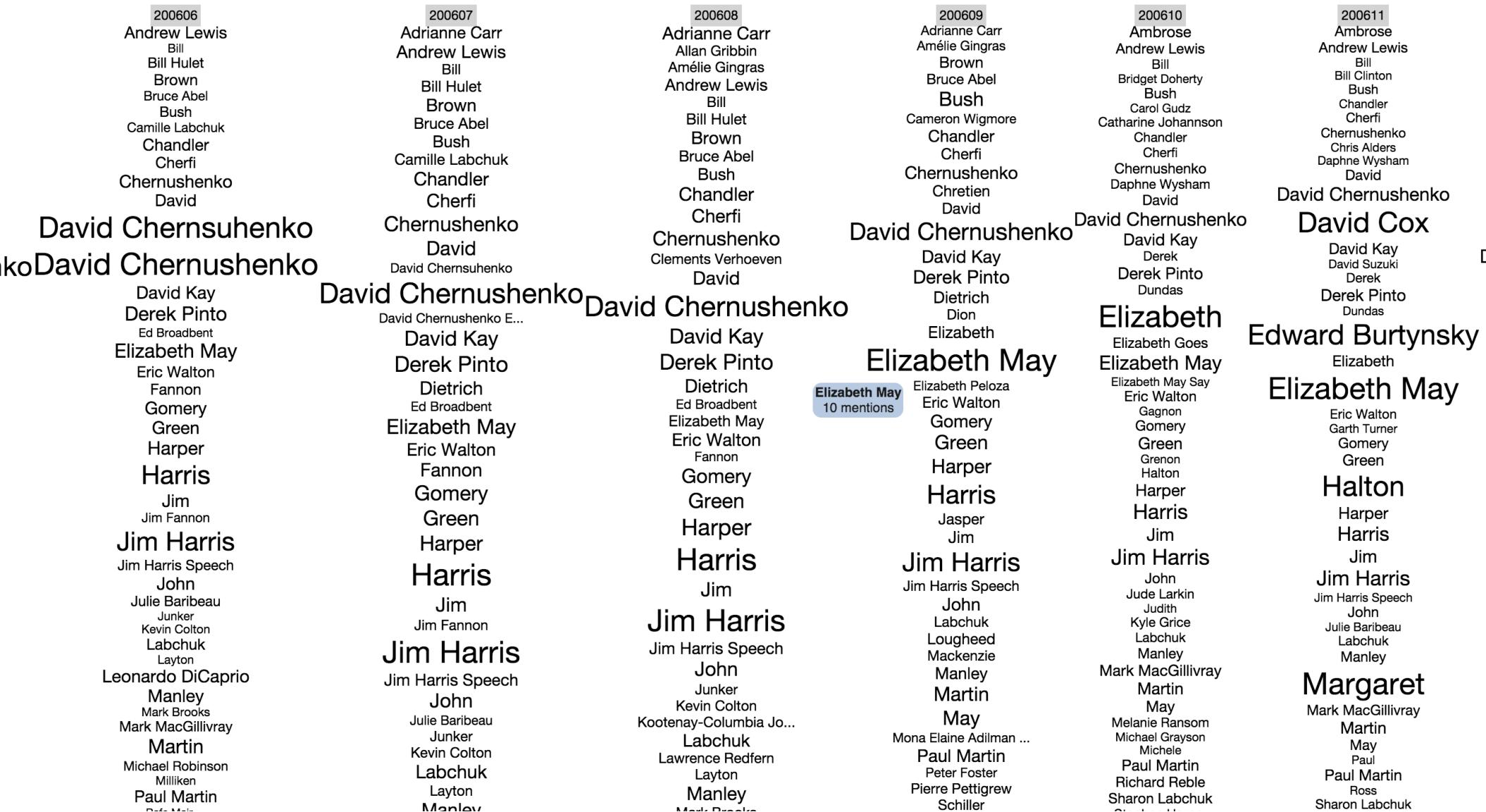
2557 Stephen Harper

2471 Stephen HarperLaureen

2410 Dion

2356 Harper

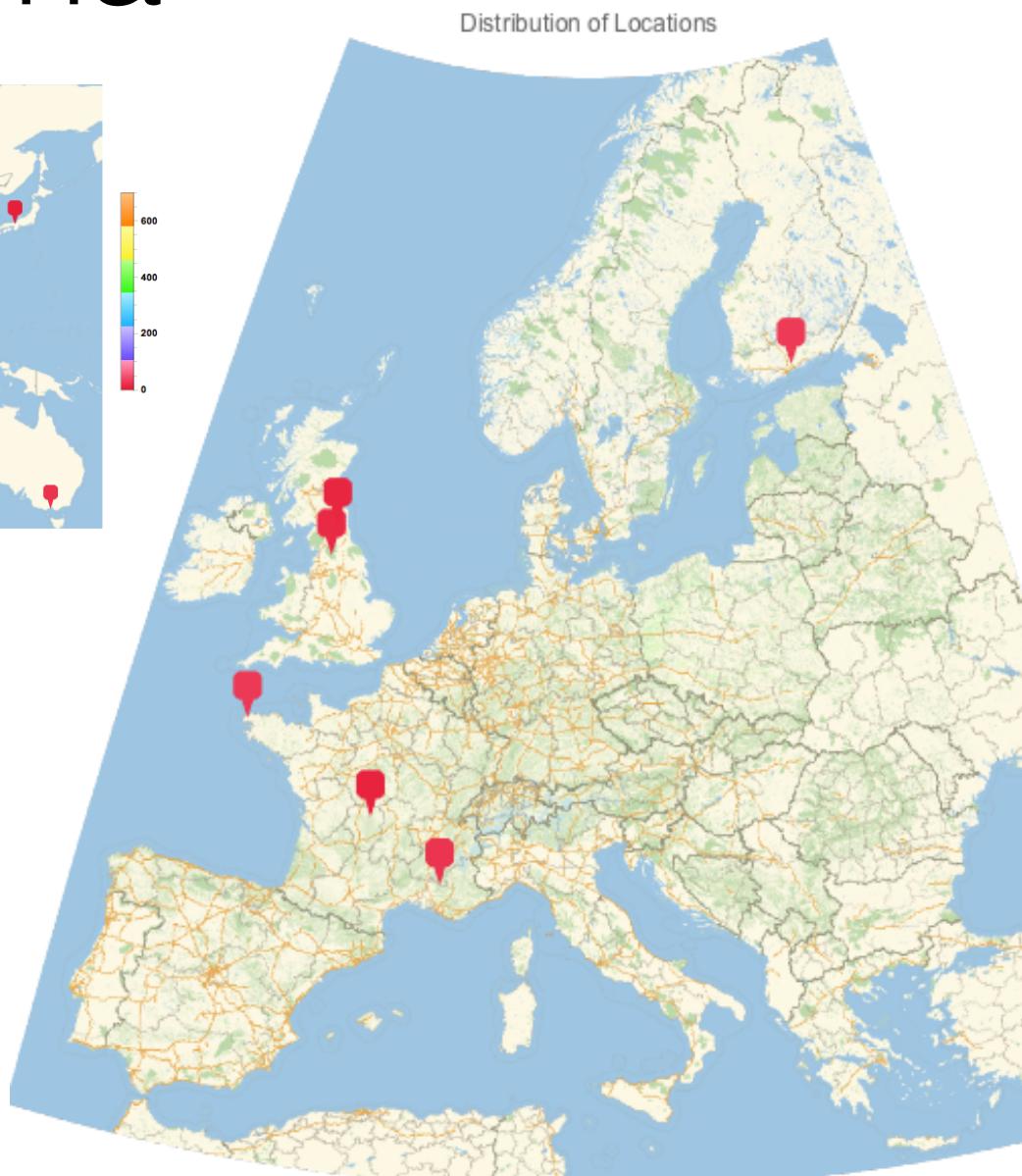
Visualizing NER



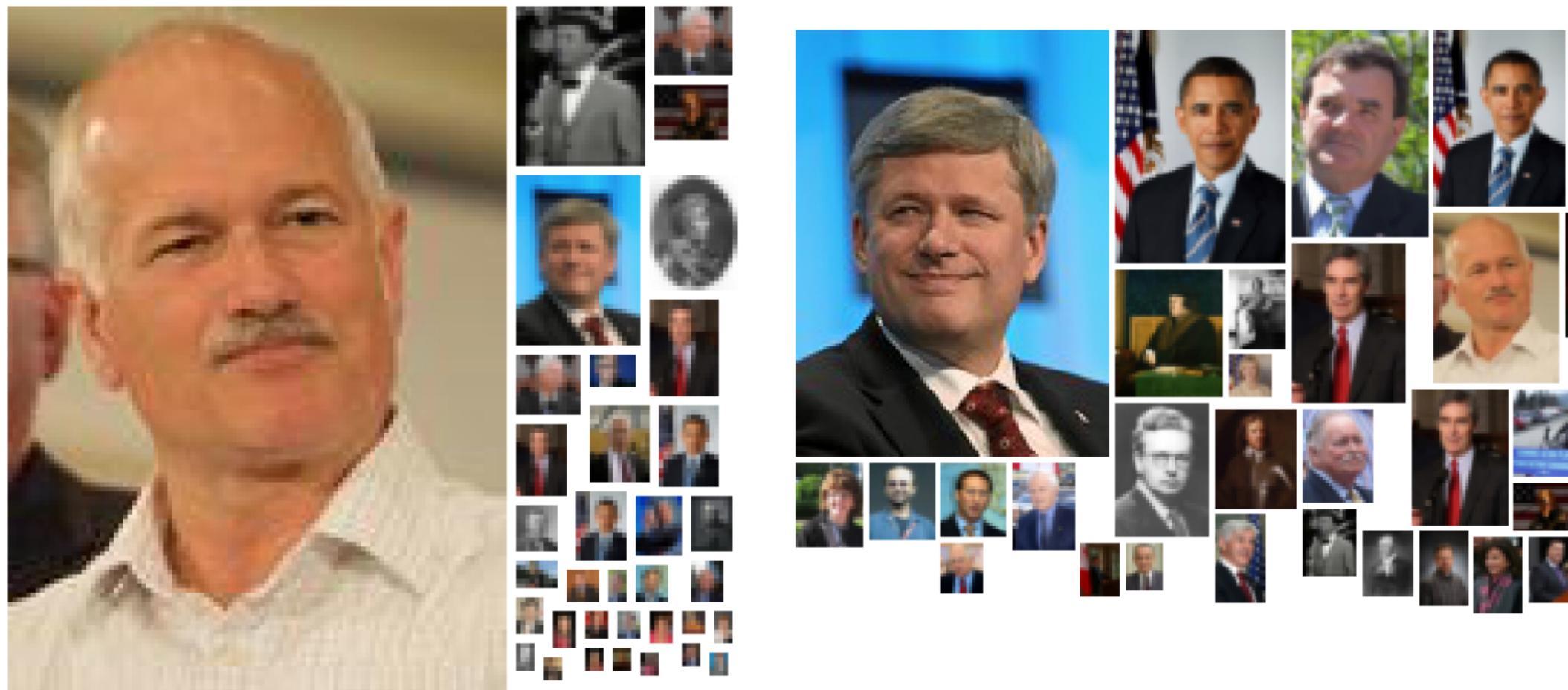
Integration with Wolfram|Alpha



```
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,
```



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 Canadian Political Parties and Political Interest Groups Portal

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

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 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 within 25 words of each other]

Below, here are all of the links for the entire time period, visualized below.

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

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.

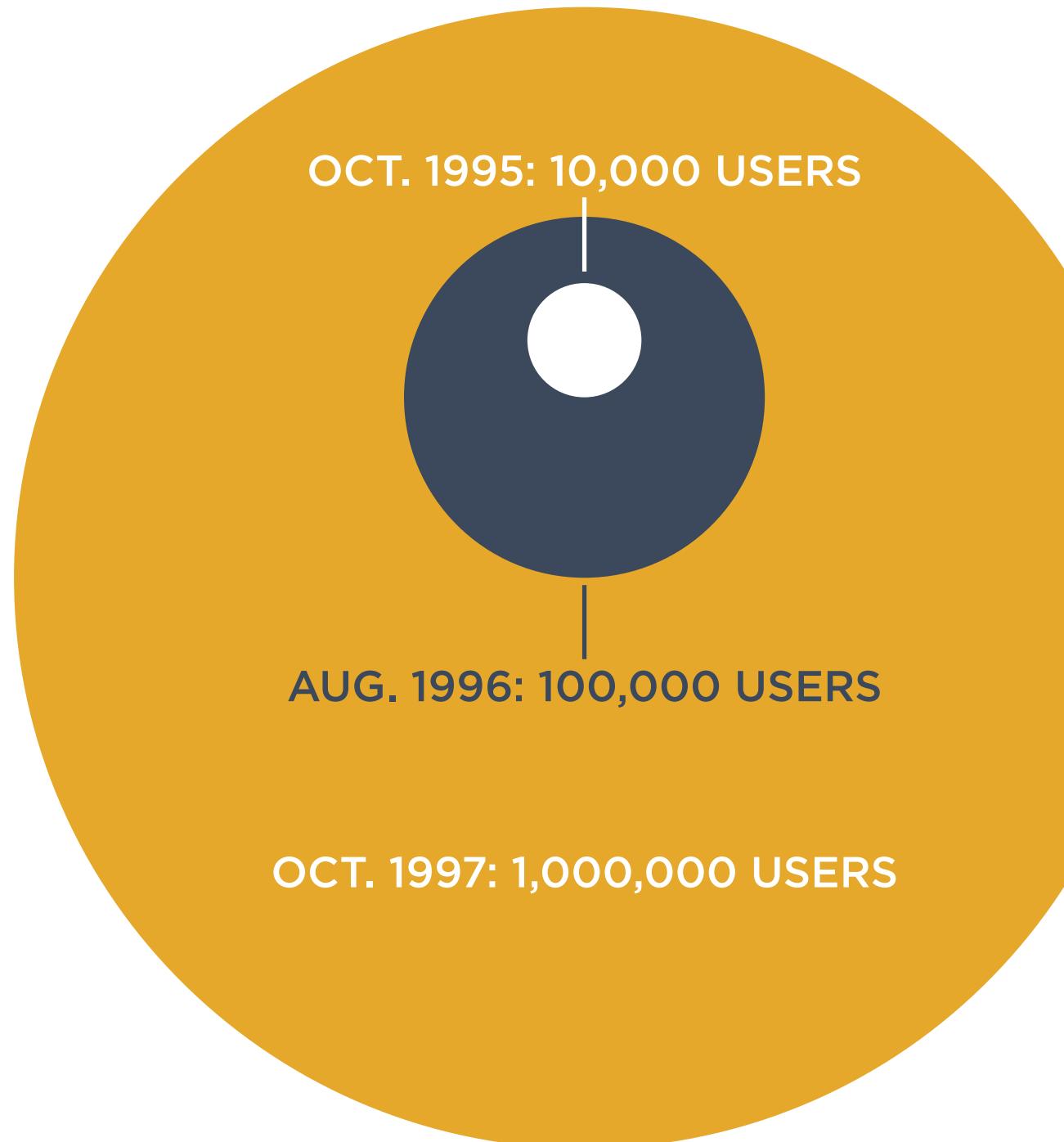
The screenshot shows a web browser window with the URL <https://web.archive.org/web/19961022173245/http://www.geocities.com/>. The page title is "Welcome to GeoCities Home". The main content area features the "Wayback Machine" logo and a banner stating "1,669 captures" from "22 Oct 96 – 13 Oct 14". Below 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 large "GEOCITIES" logo with neighborhood names like "AREAS!", "PARIS", "HEARTLAND", "ATHENS", and "TIMES SQUARE". A sidebar on the right lists links such as "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?

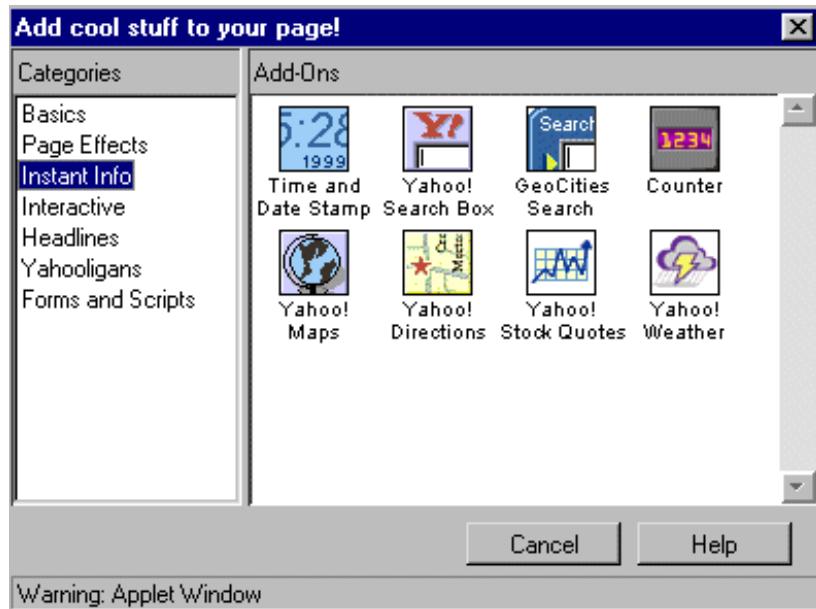
GEOCITIES USERS:

**What was
GeoCities?**

**Why does it
matter?**



“largest body of texts
detailing the lives of non-elite
people ever published”?



Massive experiment in user-generated content

Web Hosting from Yahoo S X

lan

← → C ⌘ ⌘ http://geocities.com/

Home Mail Search News Sports Finance Weather Games Answers Screen Flickr Mobile More Insta

YAHOO! SMALL BUSINESS

Search Web Sign In Mail Gear

Website Hosting Sell Online Local Marketing Commerce Central Domains Business Email Advisor

Create your professional website in no time

Easy-to-use tools, included business mail accounts, and search ad credits help you succeed.

cinefex

Need help ordering? 1-866-781-9246 (Mon-Fri, 7:00 AM to 5:00 PM PT)

Basic plan

Ideal for small personal sites

Advanced plan

Ideal for business sites

POPULAR

Premier p

Ideal for heavy tra

\$2.75

\$5.00

\$8.00

ARCHIVE TEAM



WE ARE GOING TO RESCUE YOUR SHIT

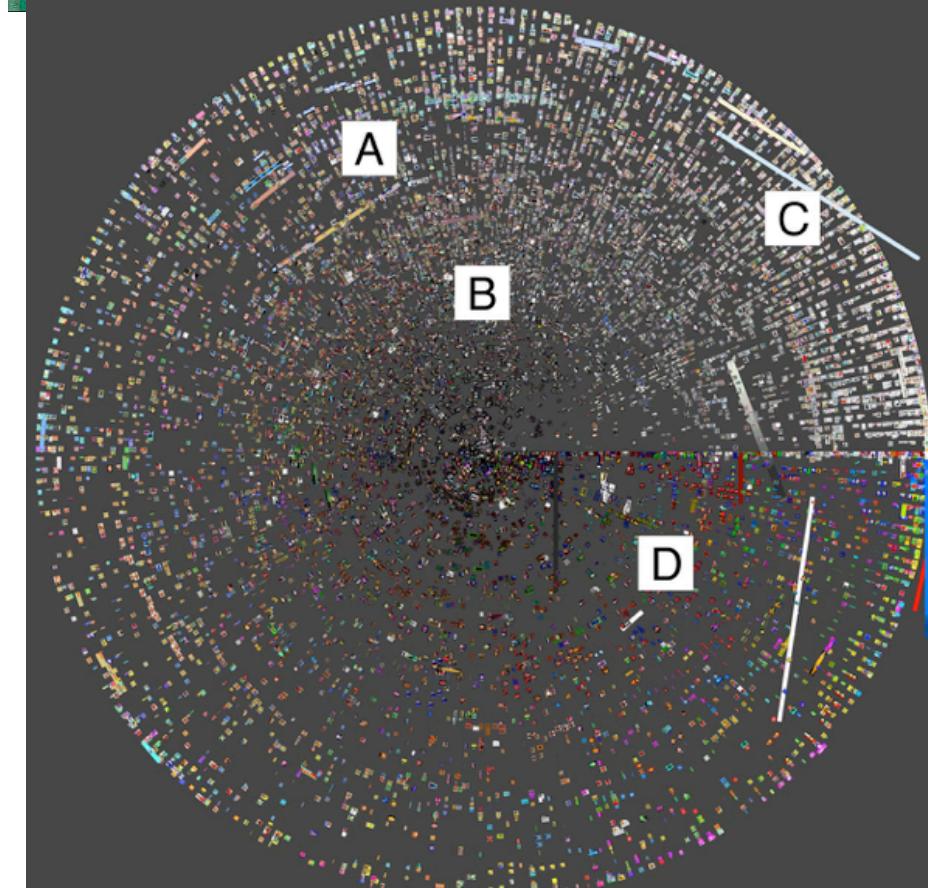
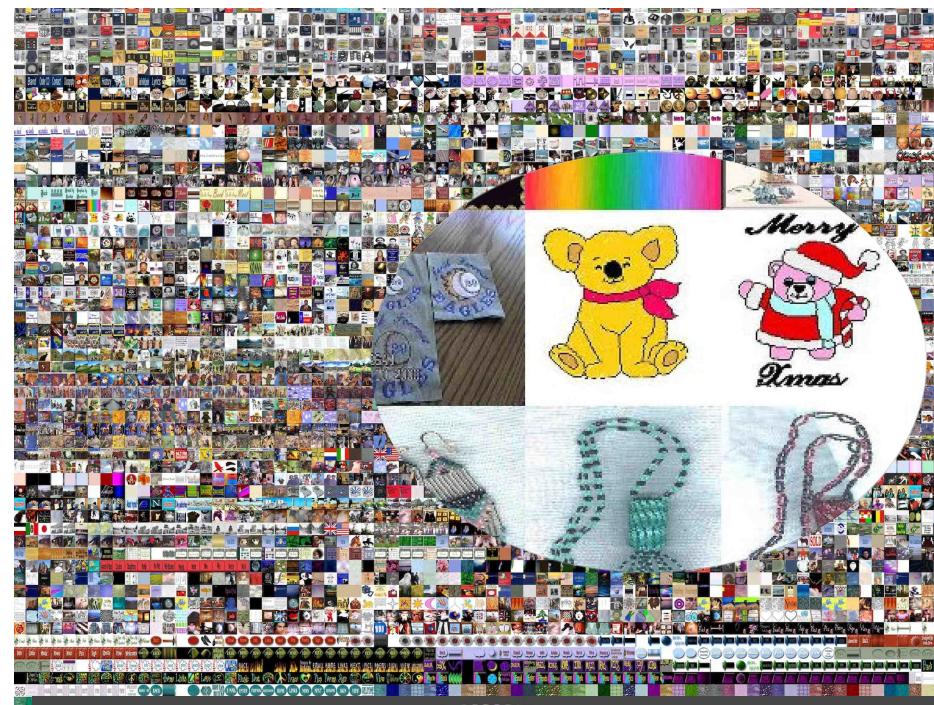


Ethically Navigating the Records of Seven Million People

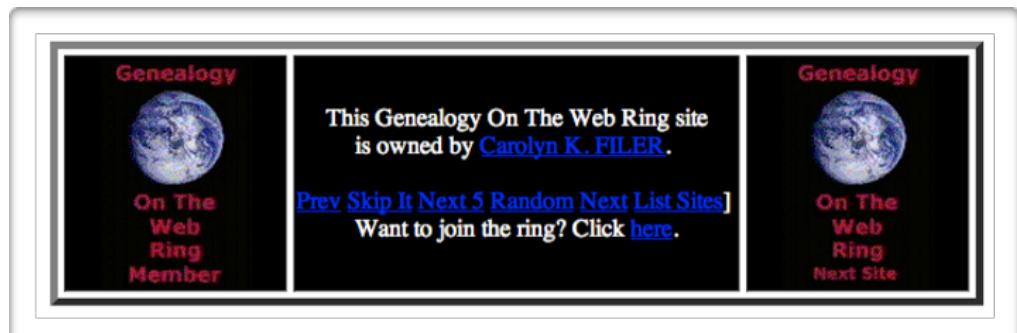
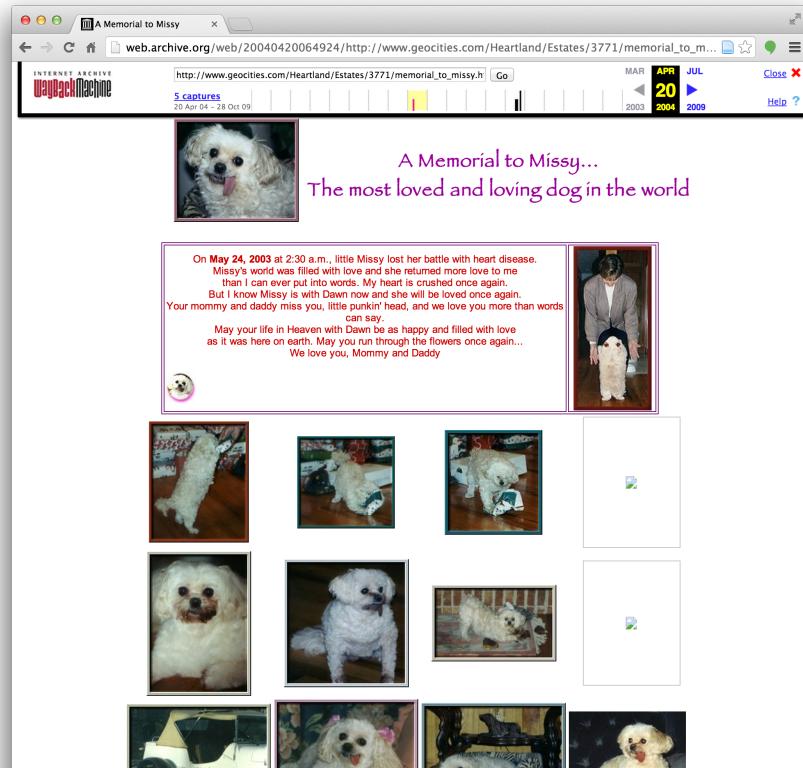
Topic Modelling Community to Test Coherence

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 human good mind soul make nature body case made point part parts goddess witch healing incense witchcraft love energy pagan shaman witches sun spirit protection light circle earth religion
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 party friends jq battalion show st jonny horse battery armored lt artillery camp sailor army field col pingu war area quest
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 year work make life school person system state world books government good family county church home years information st city born state war 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 guy guys rock frasier niles martin daphne roz don back ll door room scene ve dad turns takes crane good walks yeah
Pentagon <i>Military men and women.</i>	war people president government american world states power state united general military public soviet political clinton america make army fort war civil island iran world adams army british history badge rhode 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 feminine society identity transgendered marriage man children transsexual

Looking at millions of user- contributed & generated images



And the stories of significant users and meaningful experiences.



The possibilities of
such digital scholarship

Shared Problems

- Never have enough processing power or memory;
- Web archive tools often designed for clusters - less than ten historians in North America probably can use one...
- **Tools**
 - Some work on **WARCs**;
 - Some work on **ARCs**;
 - Some work on **WATs**;
 - And some work on **live-web material**;

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 'README.md' file is shown below, containing the following text:

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:

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