The Challenge of Digital Sources in the Web Age

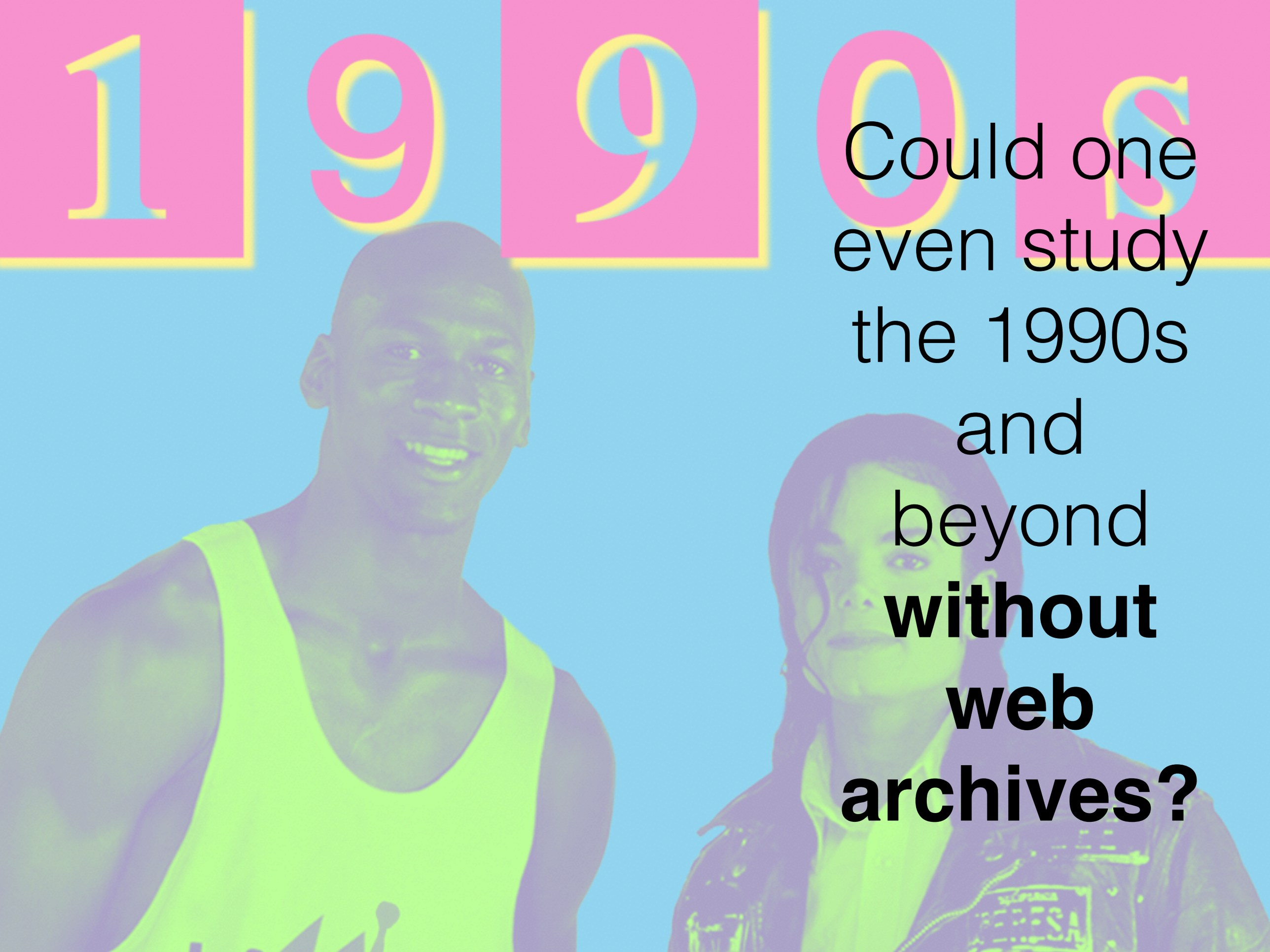
Common Tensions Across Three Web Histories, 1994-2015

Ian Milligan
Assistant Professor

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Why?

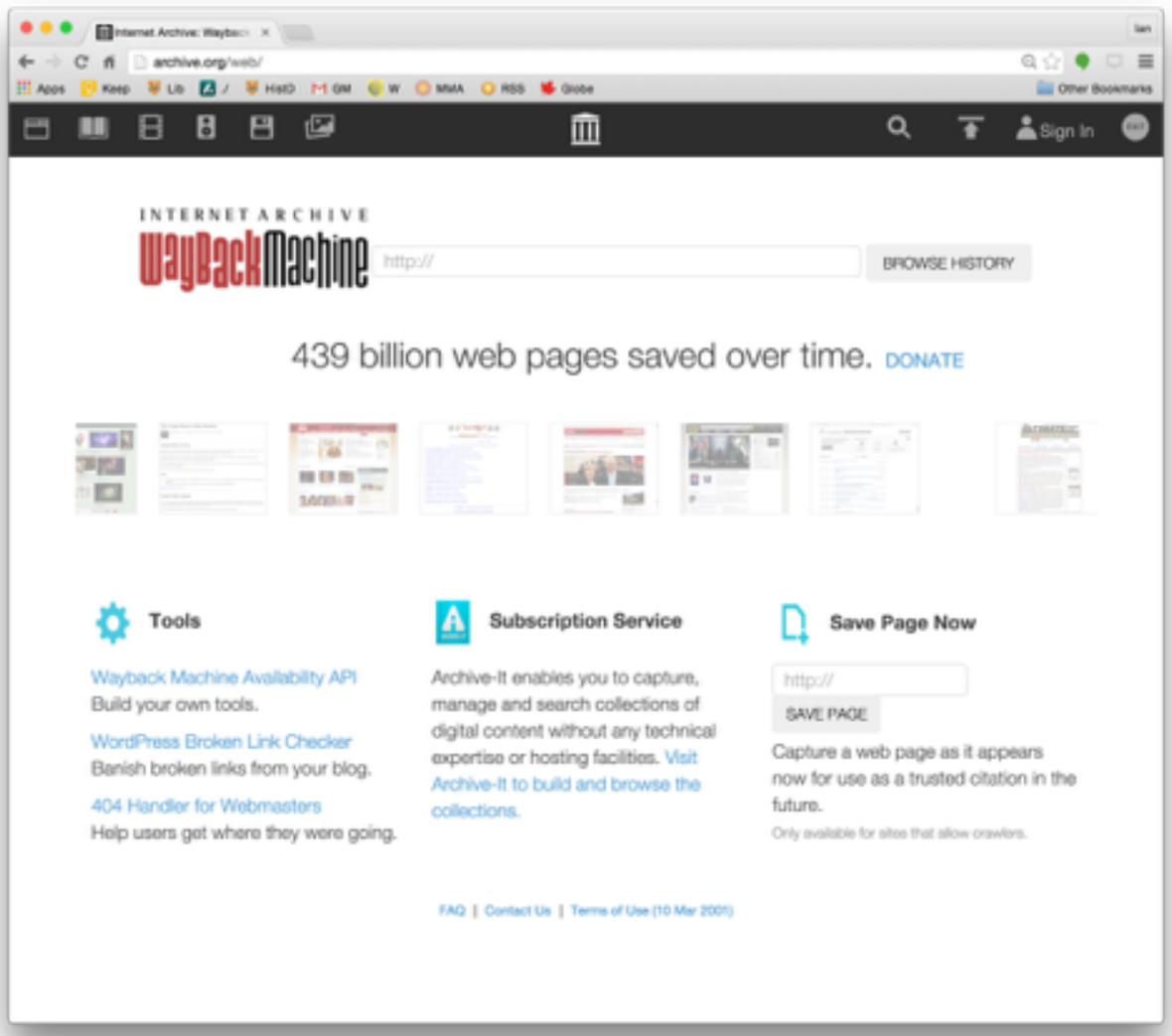
The sheer amount of social, cultural, and political information generated every day presents new opportunities for historians.


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

No.

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

Nightmare Scenario



This won’t be enough!

… but what will our search engines look like?

Nightmare Scenario

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

- 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

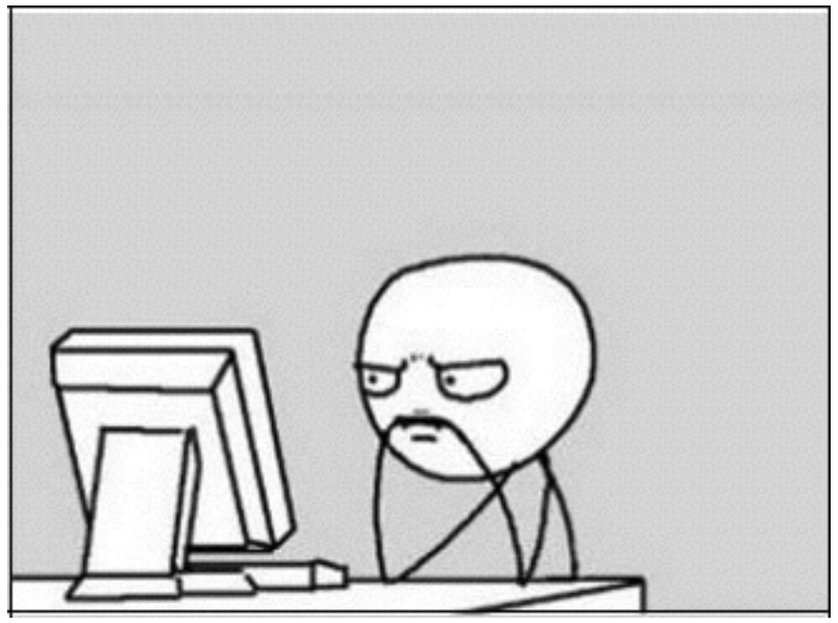
  - **Wide Web Scrape**, March - December 2011 (Internet Archive) (sample of 80TB WARC collection);

  - **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 -

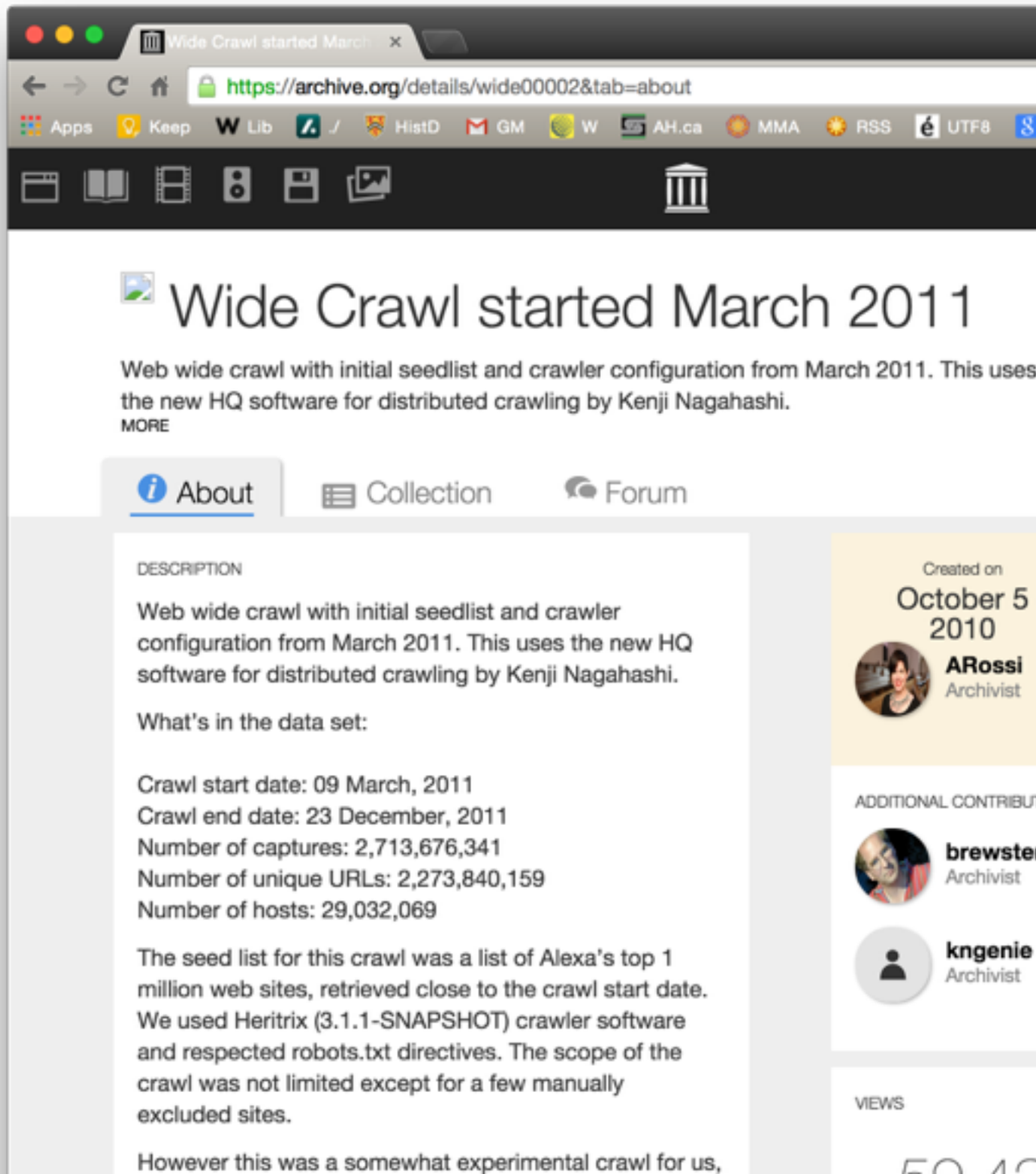
Incredible range of technical skills/no common platform!

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.

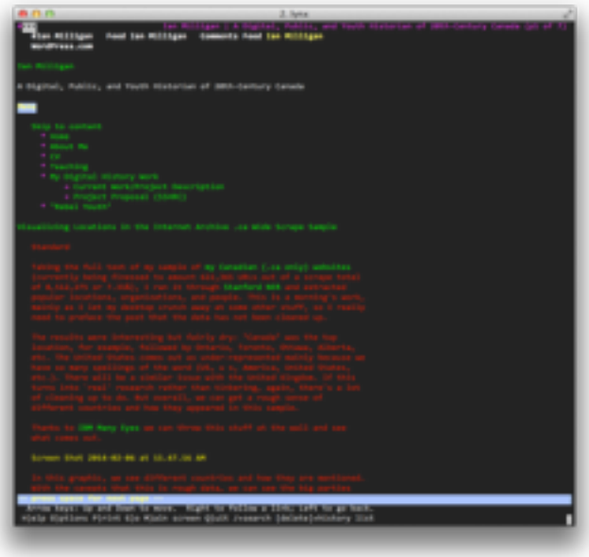


ca,yorku,justlabour)/ 20110714073726
http://www.justlabour.yorku.ca/ text/html
302 3I42H3S6NNFQ2MSVX7XZKYAYSCX5QBYJ
http://www.justlabour.yorku.ca/index.php?
page=toc&volume=16 - 462 880654831
WIDE-20110714062831-crawl416/
WIDE-20110714070859-02373.warc.gz

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)

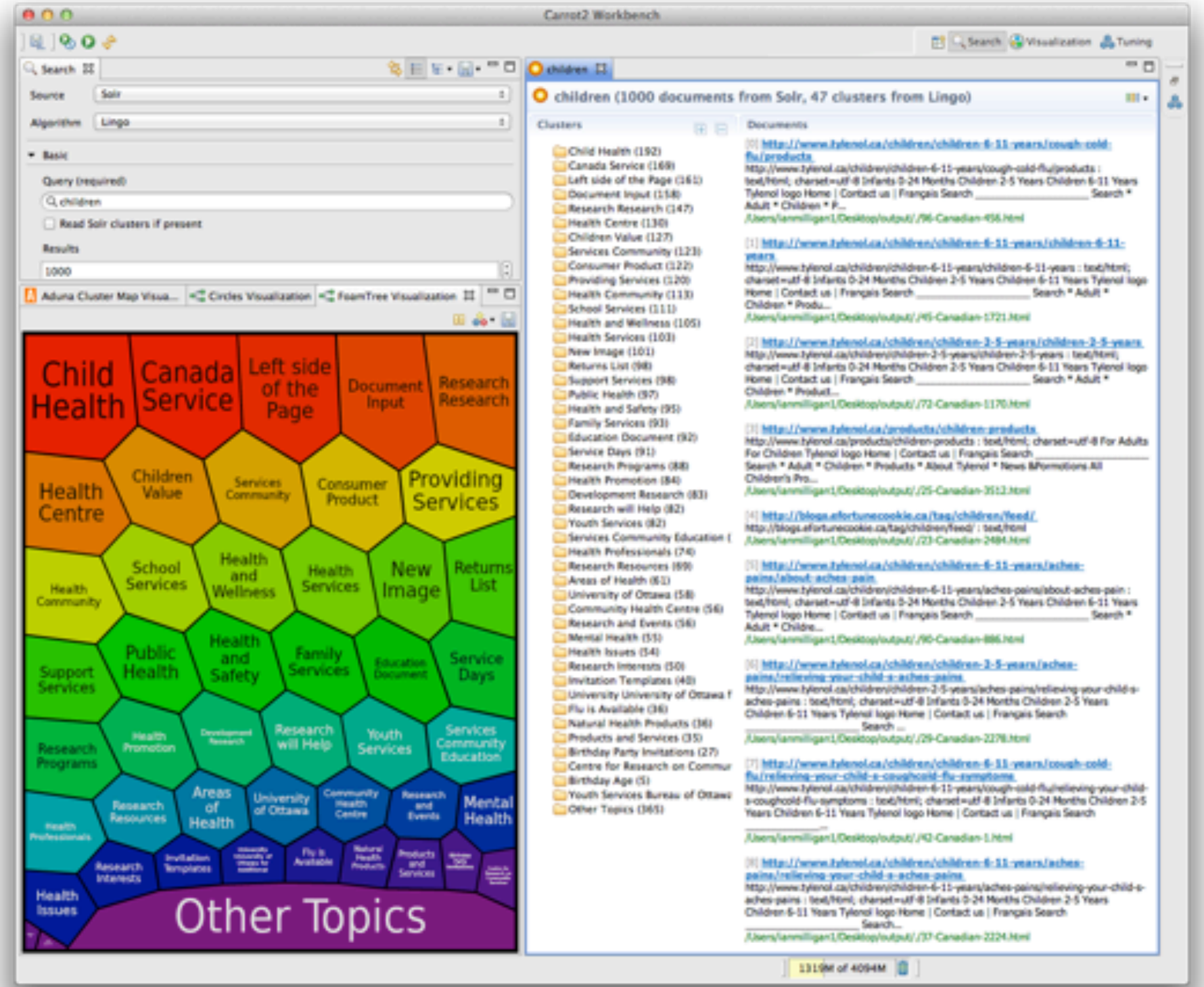
WARC File

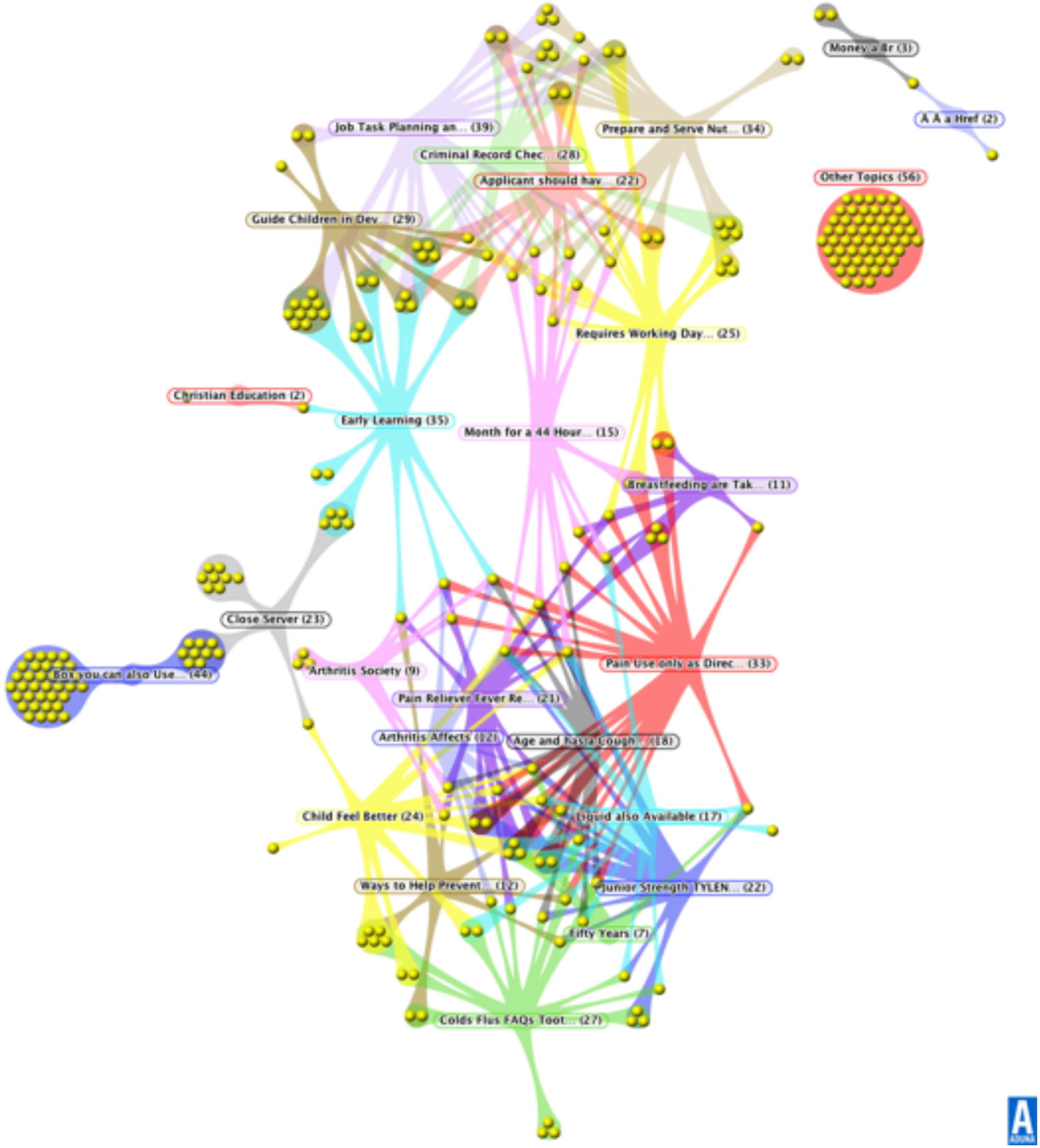


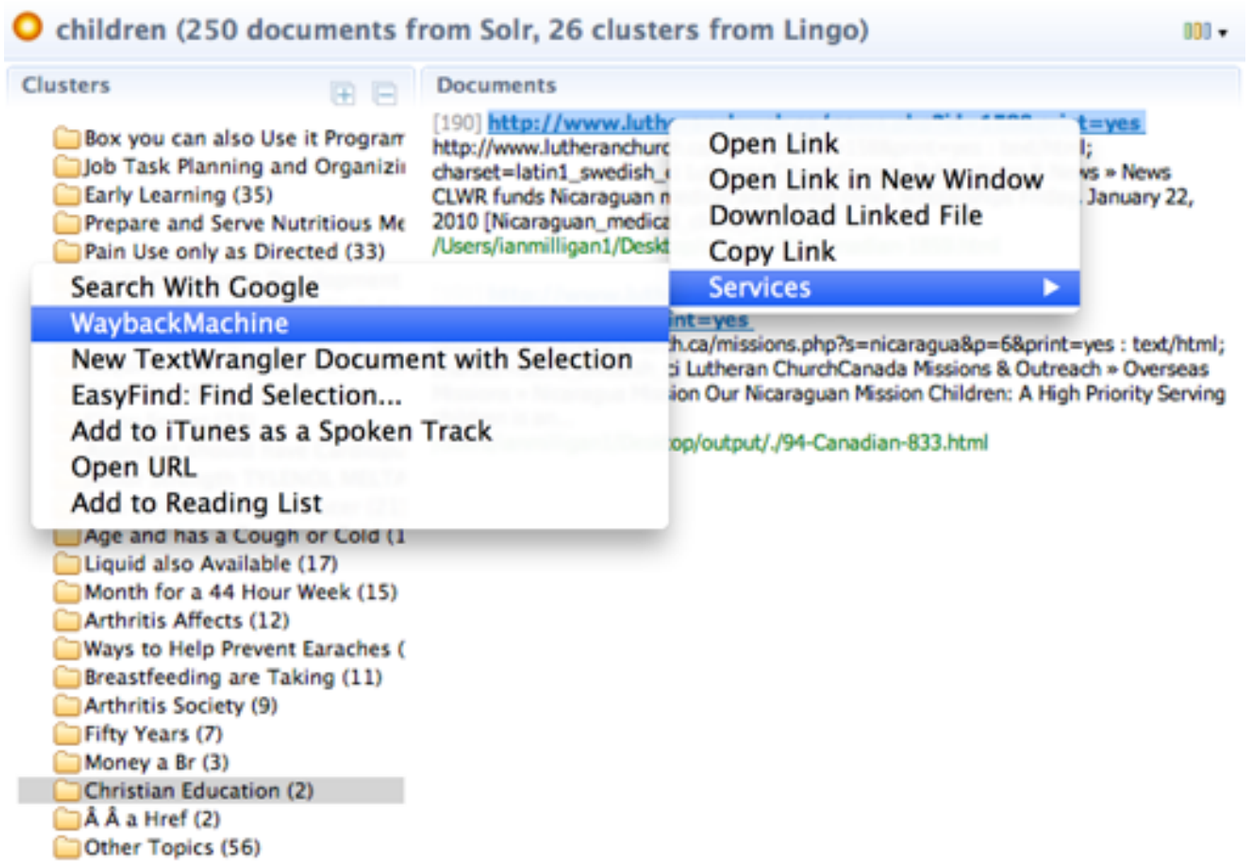
WARC-Tools/Lynx
(`warcfilter.py`, `warchtmlindex.py` and `filesdump.py`)

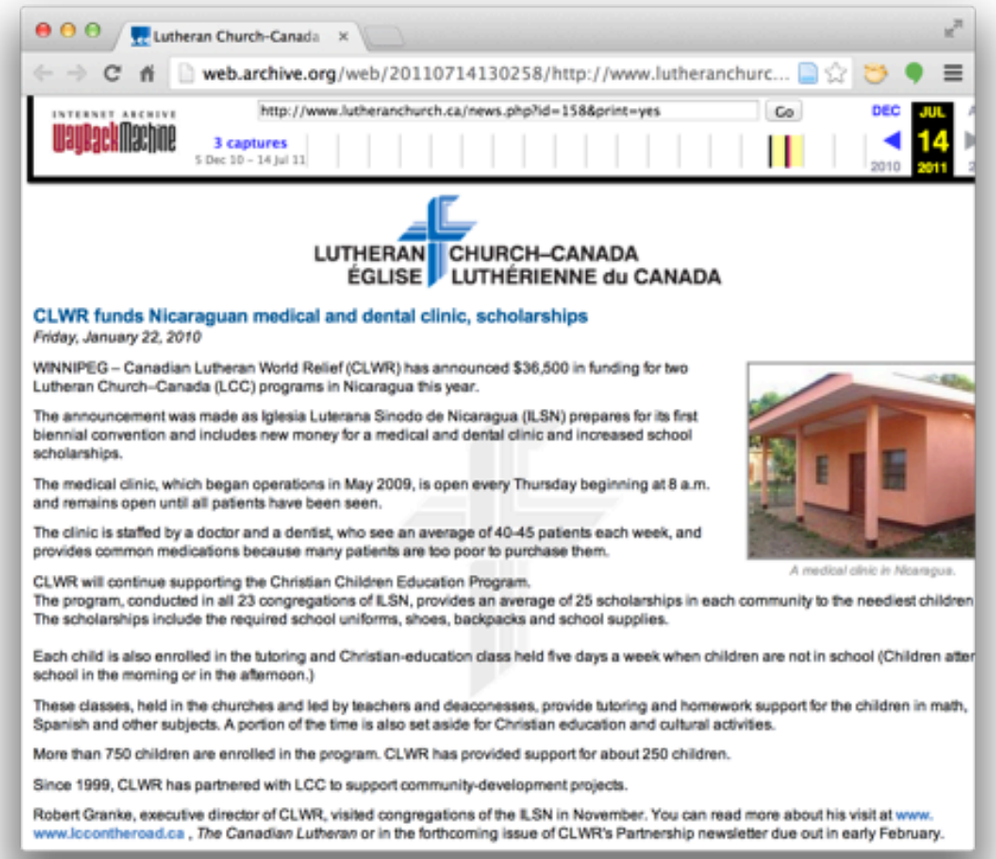
Apache Solr

Indexing


Carrot2 Workbench
Search
Visualization
Tuning
Source
Solr
Algorithm
Lingo
Basic
Query (required)
children
Read Solr clusters if present
Results
1000
Aduna Cluster Map Visua...
Circles Visualization
FoamTree Visualization
Child Health
Canada Service
Left side of the Page
Document Input
Research Research
Health Centre
Children Value
Services Community
Consumer Product
Providing Services
Health Community
School Services
Health and Wellness
Health Services
New Image
Returns List
Support Services
Public Health
Health and Safety
Family Services
Education Document
Service Days
Research Programs
Health Promotion
Development Research
Research will Help
Youth Services
Services Community Education
Health Professionals
Research Resources
Areas of Health
University of Ottawa
Community Health Centre
Research and Events
Mental Health
Research Interests
Invitation Templates
Fliu is Available
Natural Health Products
Products and Services
Health Issues
Other Topics
children
children (1000 documents from Solr, 47 clusters from Lingo)
Clusters
Child Health (192)
Canada Service (169)
Left side of the Page (161)
Document Input (158)
Research Research (147)
Health Centre (130)
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 (
Health Professionals (74)
Research Resources (69)
Areas of Health (61)
University of Ottawa (58)
Community Health Centre (56)
Research and Events (56)
Mental Health (55)
Health Issues (54)
Research Interests (50)
Invitation Templates (40)
University University of Ottawa f
Flu is Available (36)
Natural Health Products (36)
Products and Services (35)
Birthday Party Invitations (27)
Centre for Research on Commur
Birthday Age (5)
Youth Services Bureau of Ottawa
Other Topics (365)
Documents
[0] http://www.tylenol.ca/children/children-6-11-years/cough-cold-flu/products
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 | Français Search _____ Search * Adult * Children * P...
/Users/ianmilligan1/Desktop/output/./96-Canadian-456.html
[1] http://www.tylenol.ca/children/children-6-11-years/children-6-11-years
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 * Produ...
/Users/ianmilligan1/Desktop/output/./45-Canadian-1721.html
[2] http://www.tylenol.ca/children/children-2-5-years/children-2-5-years
http://www.tylenol.ca/children/children-2-5-years/children-2-5-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 * Product...
/Users/ianmilligan1/Desktop/output/./72-Canadian-1170.html
[3] http://www.tylenol.ca/products/children-products
http://www.tylenol.ca/products/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 &Promotions All Children's Pro...
/Users/ianmilligan1/Desktop/output/./25-Canadian-3512.html
[4] http://blogs.efortunecookie.ca/tag/children/feed/
http://blogs.efortunecookie.ca/tag/children/feed/ : text/html
/Users/ianmilligan1/Desktop/output/./23-Canadian-2484.html
[5] http://www.tylenol.ca/children/children-6-11-years/aches-pains/about-aches-pain
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 * Childre...
/Users/ianmilligan1/Desktop/output/./90-Canadian-886.html
[6] http://www.tylenol.ca/children/children-2-5-years/aches-pains/relieving-your-child-s-aches-pains
http://www.tylenol.ca/children/children-2-5-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 | Français Search _____ Search ...
/Users/ianmilligan1/Desktop/output/./29-Canadian-2278.html
[7] http://www.tylenol.ca/children/children-6-11-years/cough-cold-flu/relieving-your-child-s-coughcold-flu-symptoms
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 | Français Search _____
/Users/ianmilligan1/Desktop/output/./42-Canadian-1.html
[8] http://www.tylenol.ca/children/children-6-11-years/aches-pains/relieving-your-child-s-aches-pains
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 | Français Search _____ Search...
/Users/ianmilligan1/Desktop/output/./37-Canadian-2224.html
131M of 4094M

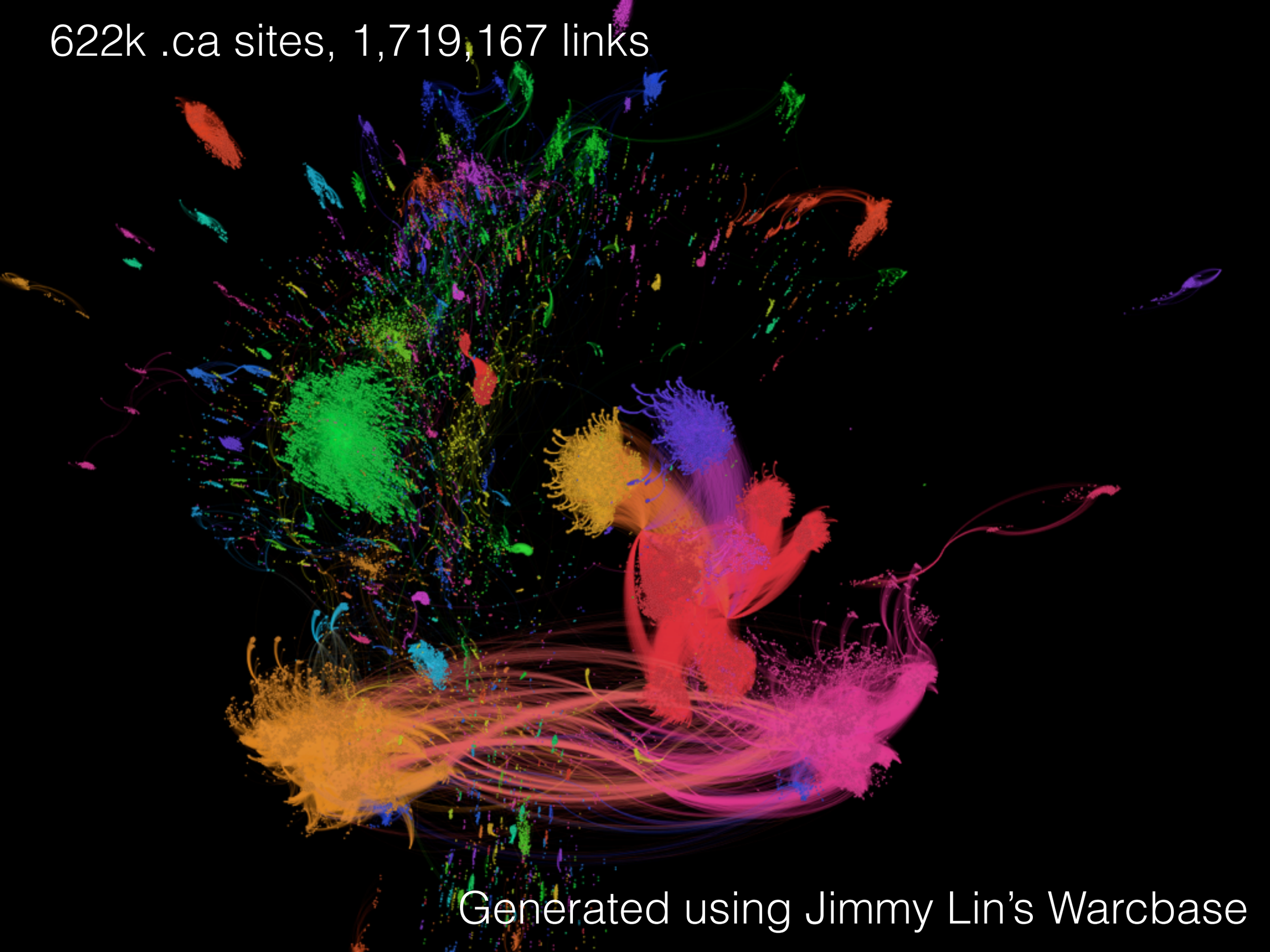

Money a Br (3)
A A a Href (2)
Other Topics (56)
Job Task Planning an... (39)
Prepare and Serve Nut... (34)
Criminal Record Chec... (28)
Applicant should hav... (22)
Guide Children in Dev... (29)
Requires Working Day... (25)
Christian Education (2)
Early Learning (35)
Month for a 44 Hour... (15)
Breastfeeding are Tak... (11)
Close Server (23)
So, you can also Use... (44)
Arthritis Society (9)
Pain Use only as Direc... (33)
Pain Reliever Fever Re... (21)
Arthritis Affects (12)
Age and hasta Cough ... (18)
Child Feel Better (24)
Liquid also Available (17)
Ways to Help Prevent... (12)
Junior Strength TYLEN... (22)
Fifty Years (7)
Colds Flus FAQs Toot... (27)


children (250 documents from Solr, 26 clusters from Lingo)
Clusters
Documents
Box you can also Use it Program
Job Task Planning and Organizi
Early Learning (35)
Prepare and Serve Nutritious Me
Pain Use only as Directed (33)
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)
Â Â a Href (2)
Other Topics (56)
[190] http://www.luth
http://www.lutheranchurc
charset=latin1_swedish_
CLWR funds Nicaraguan m
2010 [Nicaraguan_medica
/Users/ianmilligan1/Deskt
Open Link
Open Link in New Window
Download Linked File
Copy Link
Services
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
t=yes
January 22,
int=yes
h.ca/missions.php?s=nicaragua&p=6&print=yes : text/html;
i Lutheran ChurchCanada Missions & Outreach » Overseas
ion Our Nicaraguan Mission Children: A High Priority Serving
op/output/./94-Canadian-833.html


Lutheran Church-Canada
web.archive.org/web/20110714130258/http://www.lutheranchurc...
http://www.lutheranchurch.ca/news.php?id=158&print=yes
Go
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.
A medical clinic in Nicaragua.
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.
www.lccontheroad.ca , The Canadian Lutheran or in the forthcoming issue of CLWR's Partnership newsletter due out in early February.

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

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

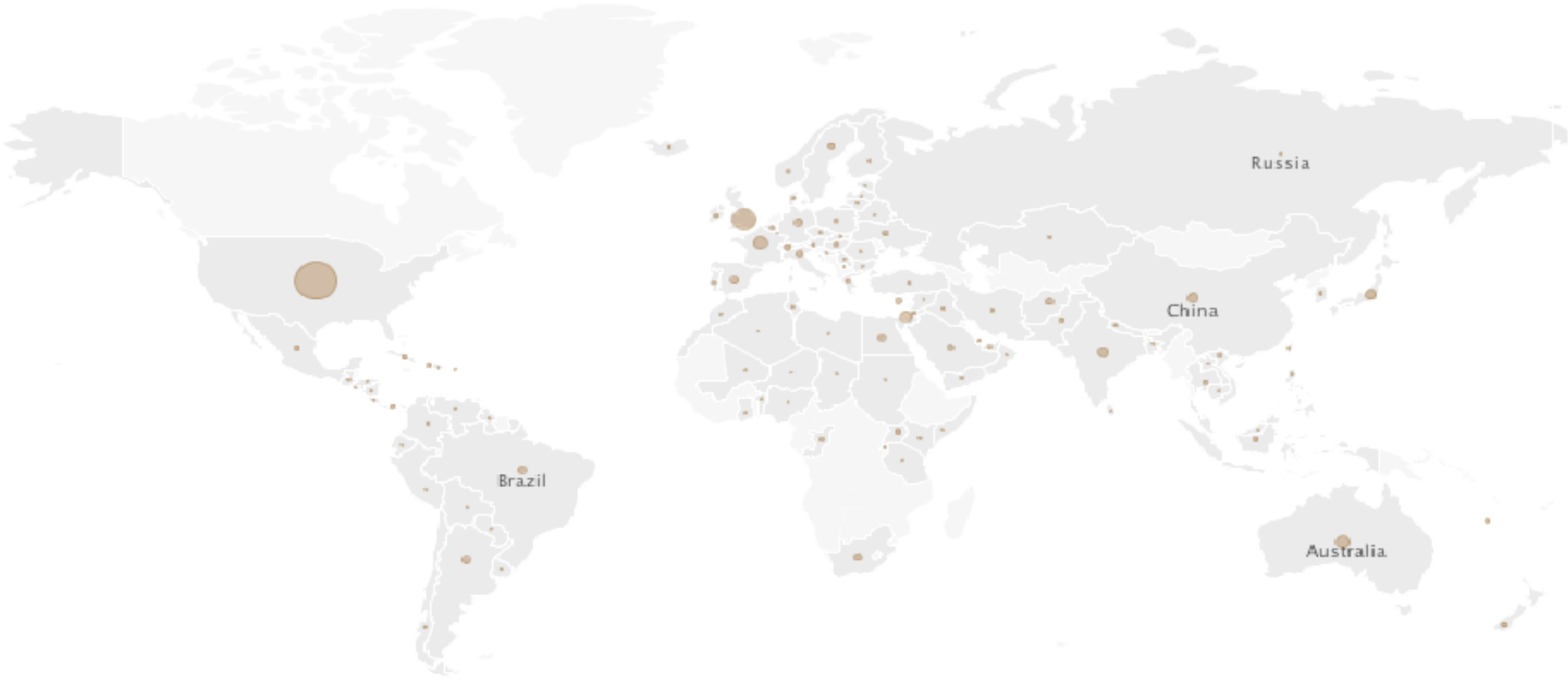


Generated using Jimmy Lin's Warcbase

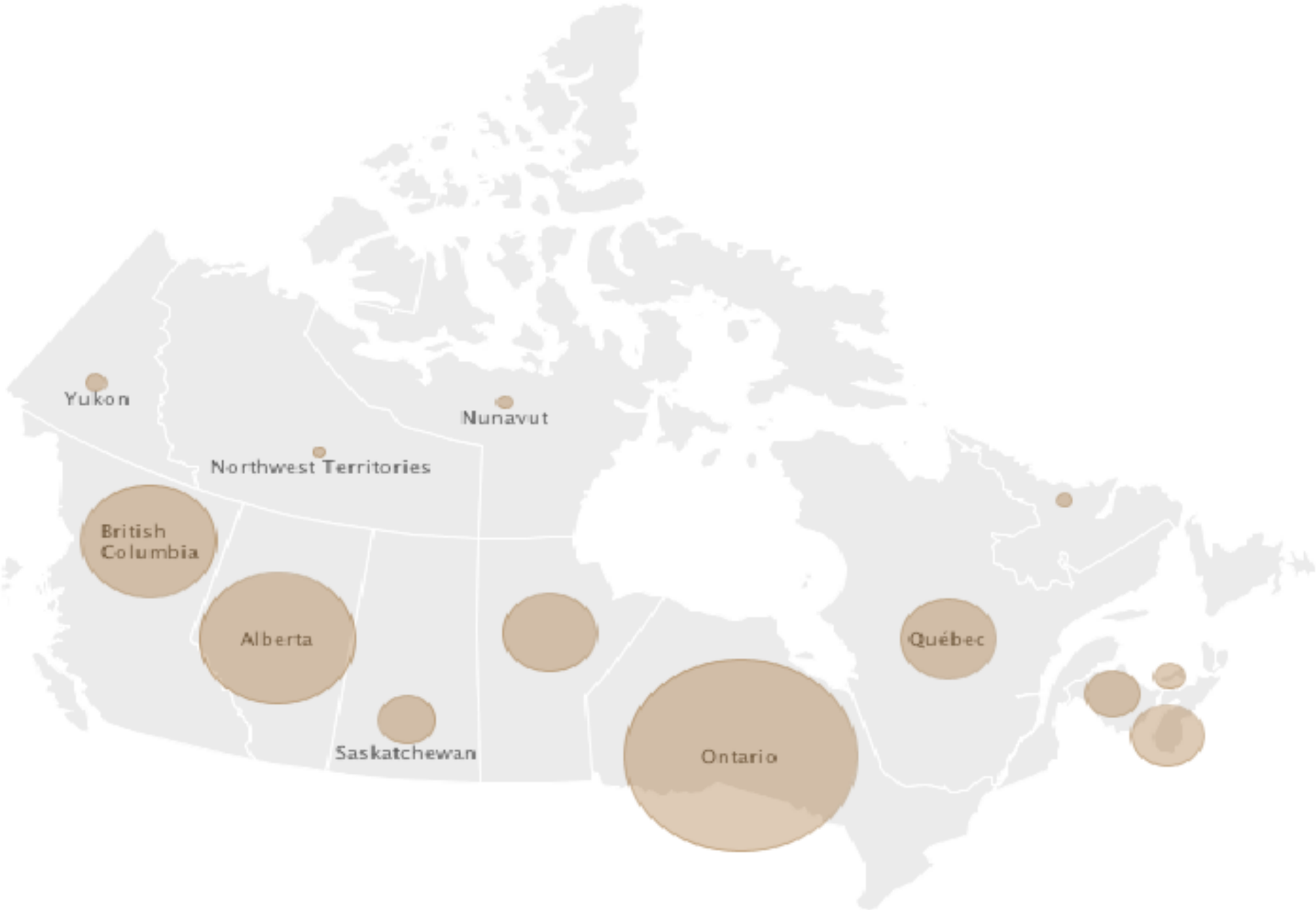

http://www.jobs-open.ca/
http://www.jobs-open.ca/info-privacy.php
http://www.jobs-open.ca/info-disclaimers.php
http://www.jobs-open.ca/info-advertising.php
mailto:webmaster@jobs-open.ca
http://www.jobs-open.ca/info-about.php


http://nova-scotia.jobs-open.ca/
http://www.uottawa.ca/cartes
http://www.biblio.uottawa.ca/index-f.php
http://www.uottawa.ca/infoscolaire/
http://www.uottawa.ca/bienvenue.html
https://web3.uottawa.ca/InfoWeb/logon/fr.html
http://www.ressourcesfinancieres.uottawa.ca/etudiant/payment-university-fees-fr.php
http://www.uottawa.ca/academic/info/regist/crs/programmes.html
http://www.admission.uottawa.ca/Default.aspx?tabid=2548&source=qp
http://www.uottawa.ca/contacteznous.html
http://www.uottawa.ca/requirer.html
http://www.admission.uottawa.ca/Default.aspx?tabid=2672
http://www.etudesup.uottawa.ca/Default.aspx?tabid=1726
https://web9.uottawa.ca/services/SiteEvaluation/
http://www.uottawa.ca/icone-recherche/

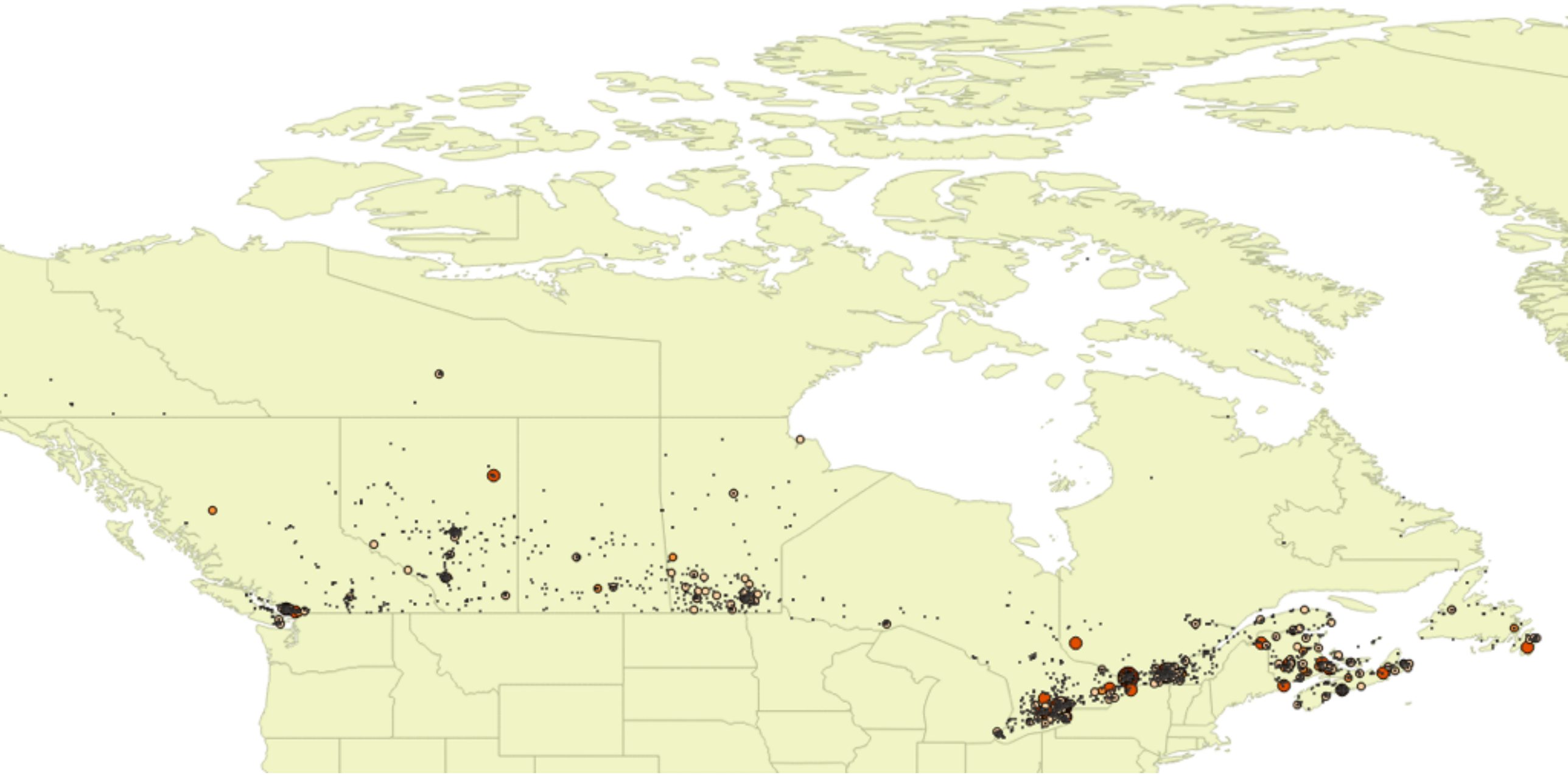
Countries Mentioned in .ca TLD (excluding Canada)



Provinces Mentioned in .ca TLD



Canadian Postal Codes visualized



Need longitudinal, but the size/intensity = extreme.

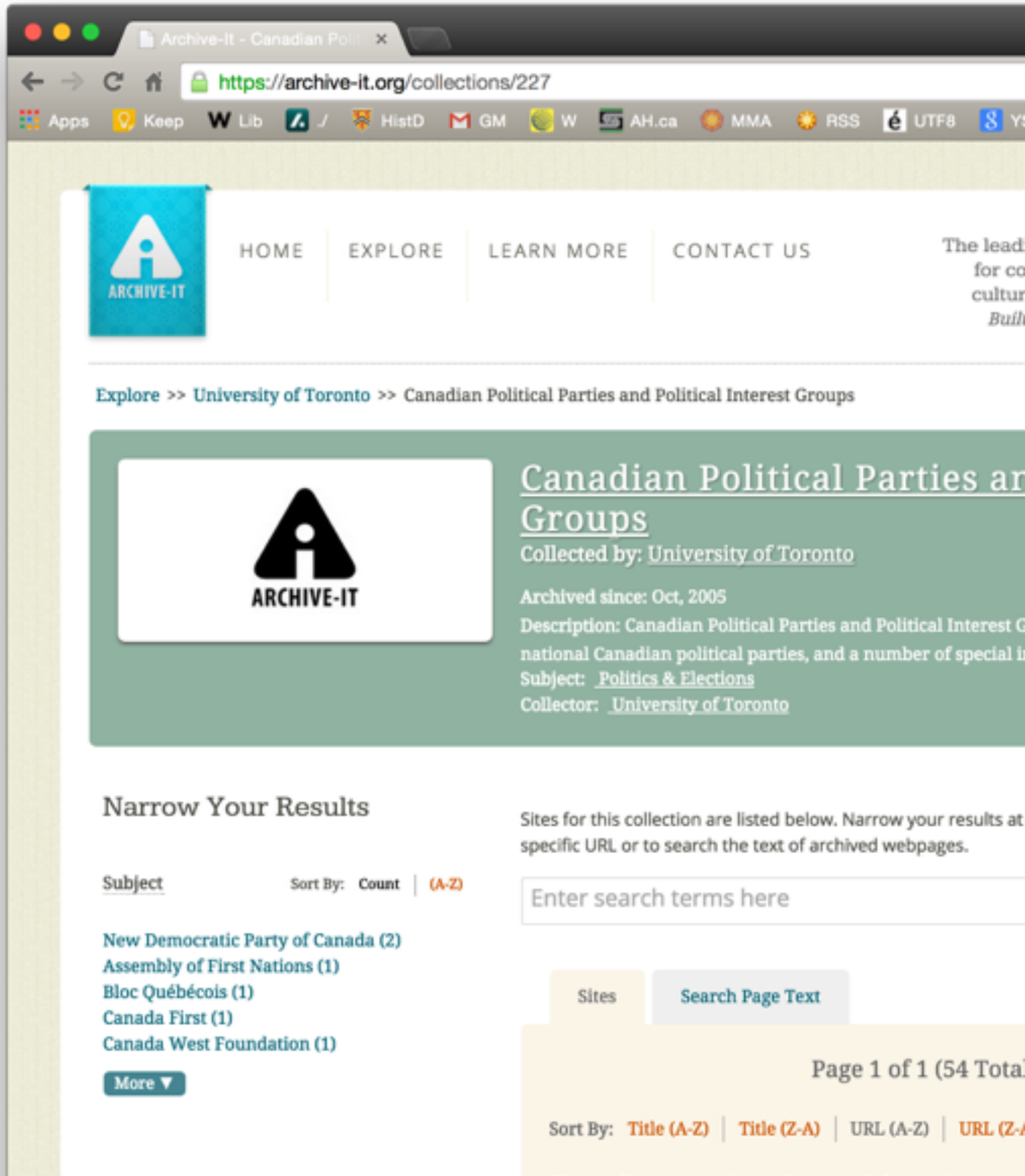
Wide Web Scrapes and the **Dream of Social History**.

Case Study Two

- **Archive-It Research Services**: "Canadian Political Parties and Political Interest Groups" and "Canadian Labour Unions."

- 2005 - 2015

- WAT & WARC files



Pivotal Changes in Canadian Politics, 2005-2015

- Militarization of Canadian society?

- Change from ‘natural governing party’ of Liberals to Conservatives

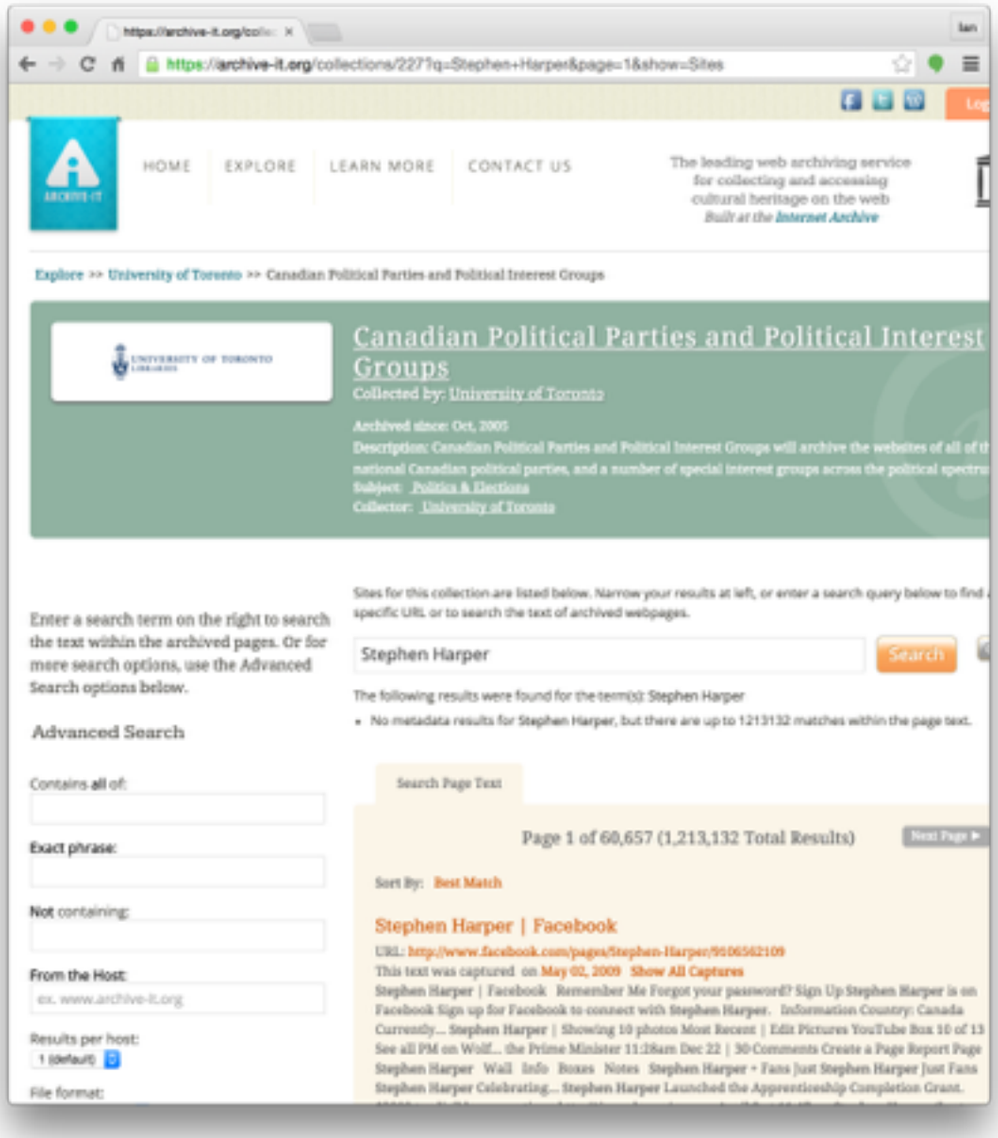
- Major policy changes on foreign policy, environment, etc.

- How to measure?

Current Interface

- Very limited - simple search engine, some advanced options; no facets

- Great collections.. but nobody uses them!



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**?

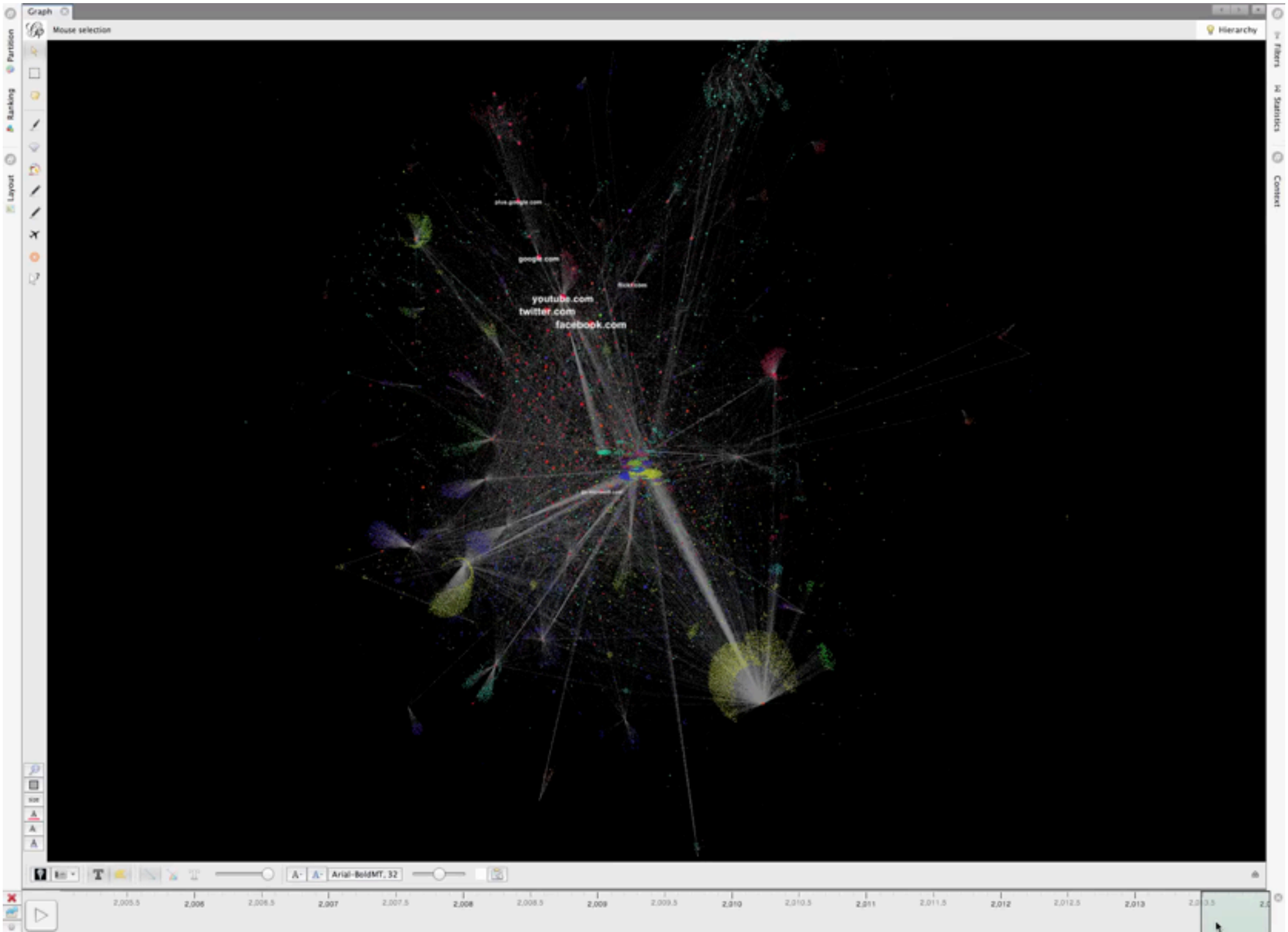
Two problems

Problem One:

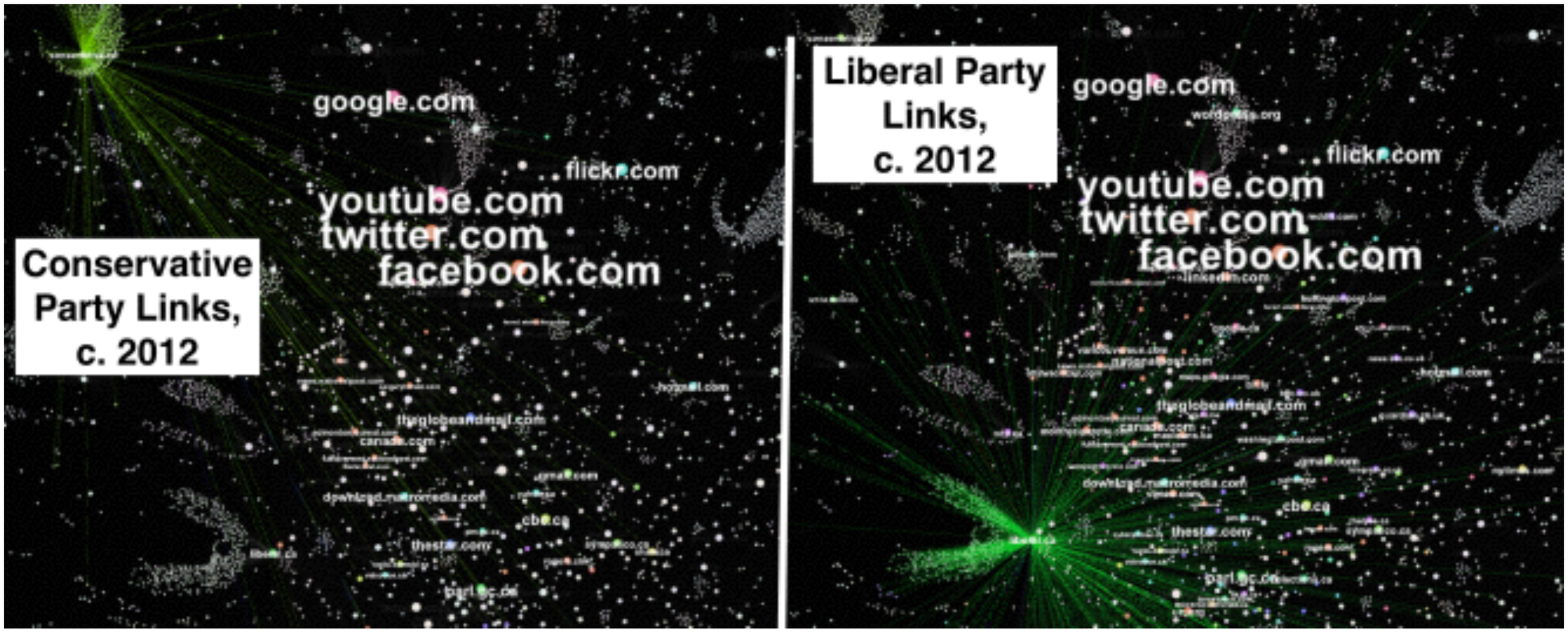
Historians want to work with content, but we can only use metadata on most computer systems.

(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

colincarriemp.ca	12
colincarriemp.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**;

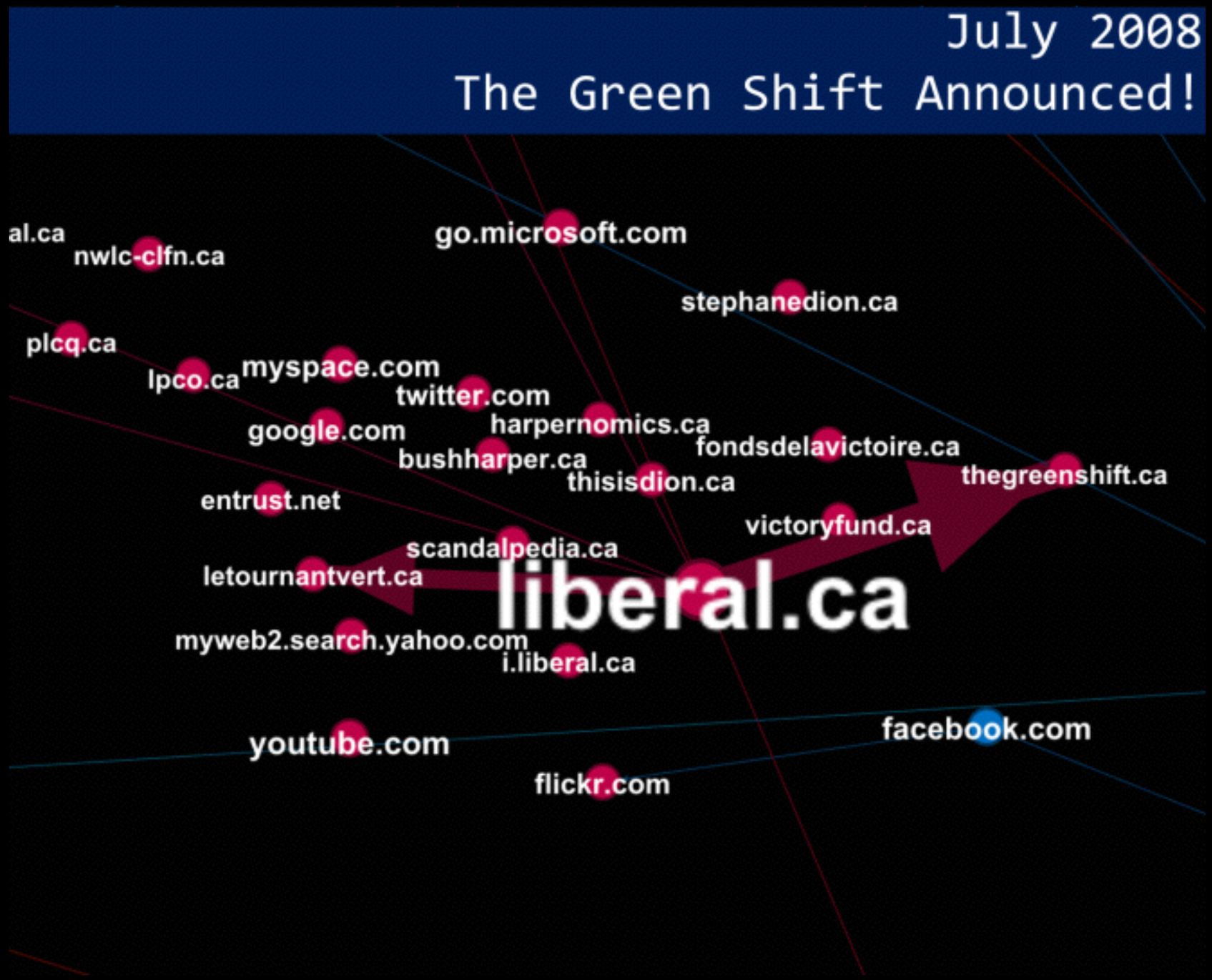
- Liberal themes (2006): community questions, electoral topics, universities, human rights, child care support.

As well as short stories...

December 2006
Stephane Dion Elected Leader of Party

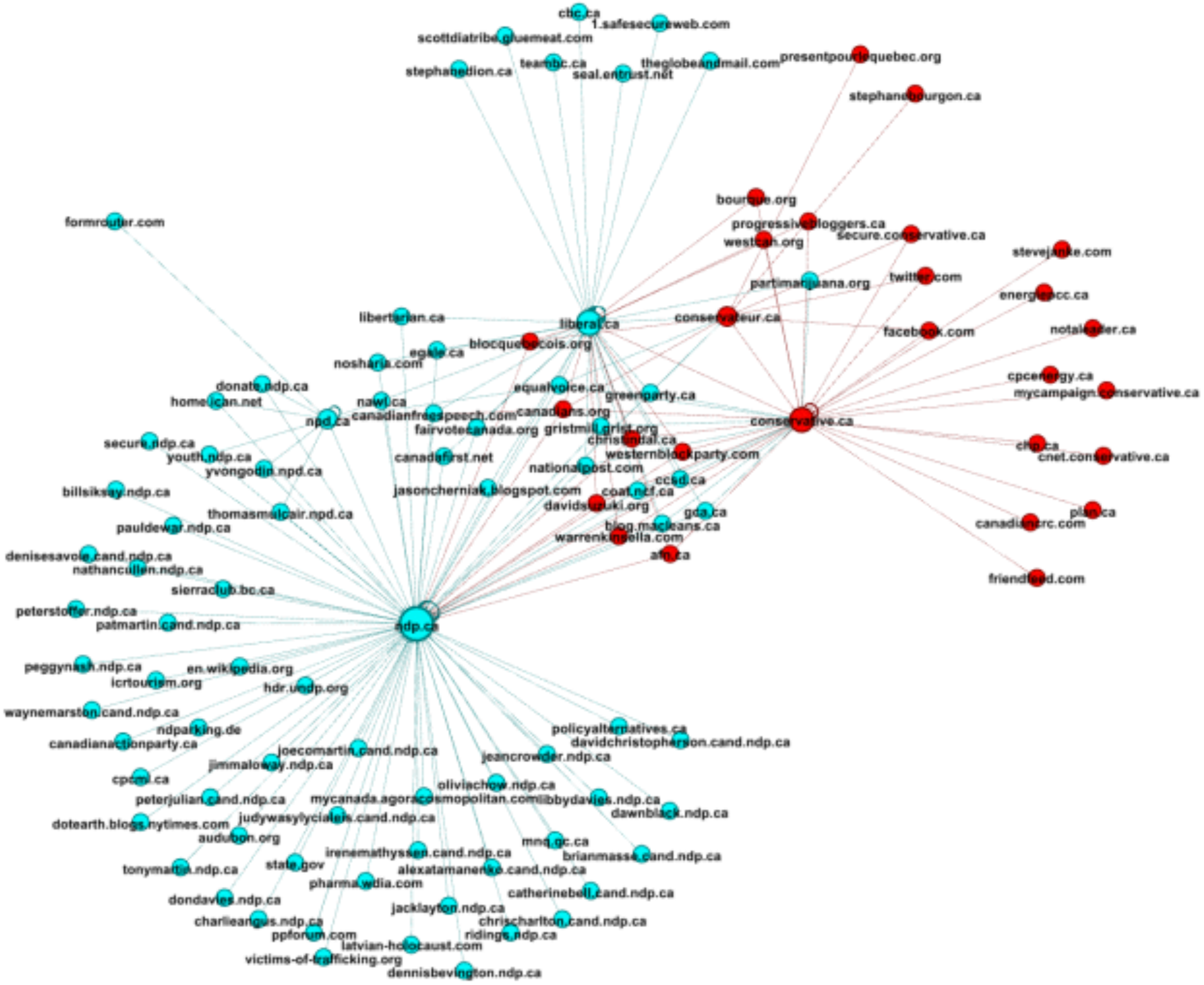
December 2007
Rise of Social Media

April 2008
Fundraising with the Victory Fund/ Fonds de la Victoire

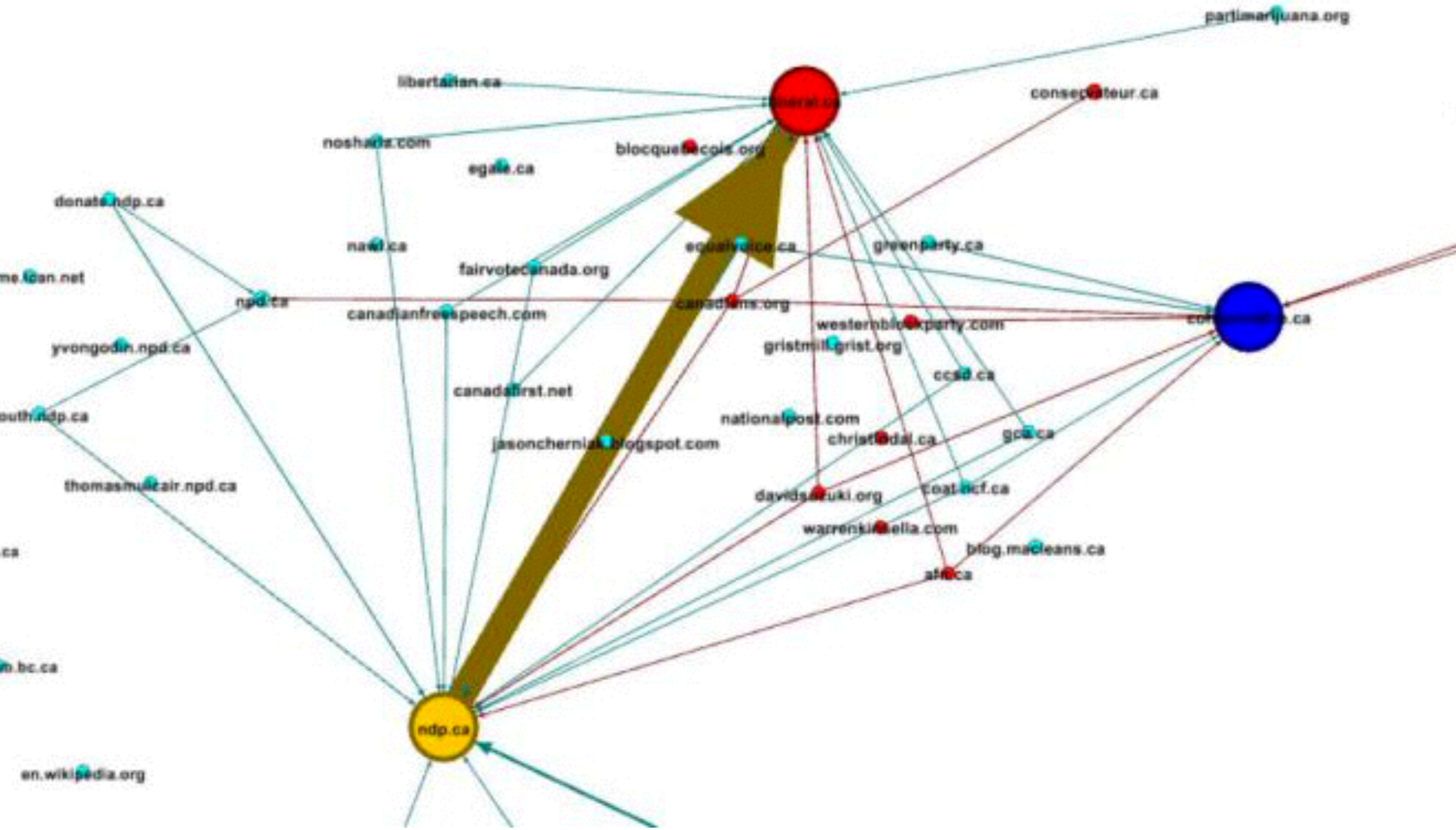

July 2008
The Green Shift Announced!
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
youtube.com
facebook.com
flickr.com

October 2008
Election Campaign - Advertisement Sites

December 2008
Election campaign Ends; Attacking Harper on Anti-American Grounds (bushharper)


cbc.ca
1.safesecureweb.com
scottdiatribe.gluemeat.com
stephanedion.ca
teambc.ca
seal.entrust.net
theglobeandmail.com
presentpourlequebec.org
stephanebourgon.ca
bourque.org
progressivebloggers.ca
westcan.org
secure.conservative.ca
formrouter.com
partimarijuana.org
twitter.com
stevejanke.com
energyprcc.ca
libertarian.ca
liberal.ca
conservateur.ca
facebook.com
notaleader.ca
egale.ca
blocquebecois.org
noshame.com
donate.ndp.ca
equalvoice.ca
greenparty.ca
cpcenergy.ca
homelcan.net
nawl.ca
canadians.org
mycampaign.conservative.ca
canadianfreespeech.com
npd.ca
fairvotecanada.org
gristmill.grist.org
conservative.ca
christendial.ca
secure.ndp.ca
youth.ndp.ca
canadafirst.net
westernblockparty.com
cbp.ca
cnet.conservative.ca
yvongodin.npd.ca
nationalpost.com
ccsd.ca
billsiksay.ndp.ca
jasoncherniak.blogspot.com
coat.ncf.ca
davidsuzuki.org
gda.ca
plan.ca
thomasmulcair.npd.ca
blog.macleans.ca
canadiancrc.com
pauldewar.ndp.ca
warrenkinsella.com
afn.ca
denisesavoie.cand.ndp.ca
nathancullen.ndp.ca
friendfeed.com
sierraclub.bc.ca
peterstoffer.ndp.ca
patmartin.cand.ndp.ca
ndp.ca
peggynash.ndp.ca
en.wikipedia.org
icrtourism.org
hdr.undp.org
waynemarston.cand.ndp.ca
ndparking.de
policyalternatives.ca
canadianactionparty.ca
davidchristopherson.cand.ndp.ca
joecomartin.cand.ndp.ca
jeancrowder.ndp.ca
jimmaloway.ndp.ca
cpcml.ca
oliviachow.ndp.ca
peterjulian.cand.ndp.ca
mycanada.agoracosmopolitan.com
libbydavies.ndp.ca
dawnblack.ndp.ca
judywasylycia-leis.cand.ndp.ca
dotearth.blogs.nytimes.com
audubon.org
mnq.qc.ca
irenemathyssen.cand.ndp.ca
brianmasse.cand.ndp.ca
tonymartin.ndp.ca
state.gov
alexatamanenko.cand.ndp.ca
pharmawatch.com
catherinebell.cand.ndp.ca
dondavies.ndp.ca
jacklayton.ndp.ca
charlieangus.ndp.ca
chrischarlton.cand.ndp.ca
ppforum.com
ridings.ndp.ca
latvian-holocaust.com
victims-of-trafficking.org
dennisbevington.ndp.ca

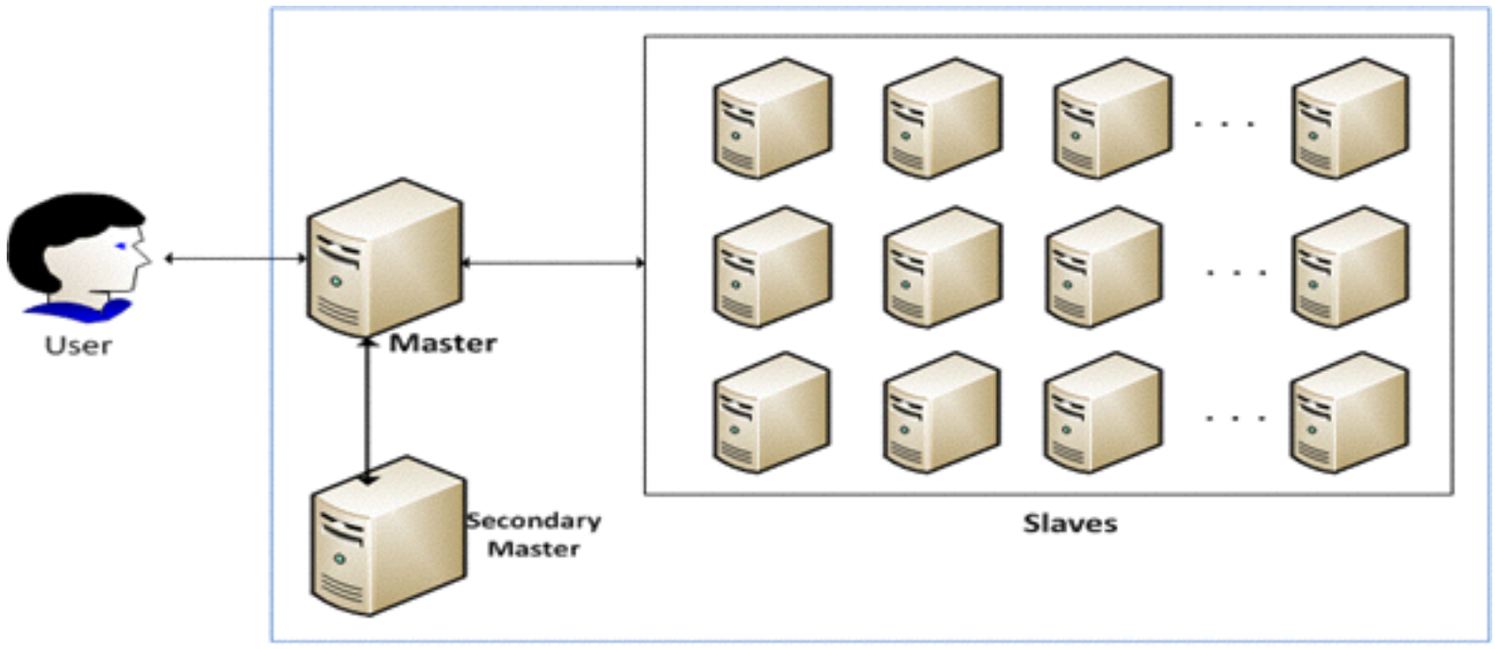
2005 Canadian Federal Election



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

Problem Two:

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



Hadoop Cluster

WARC Analysis

- **2005-2009**: 244 GB of content; 2.9 GB of plain text

- 10,606,822 websites

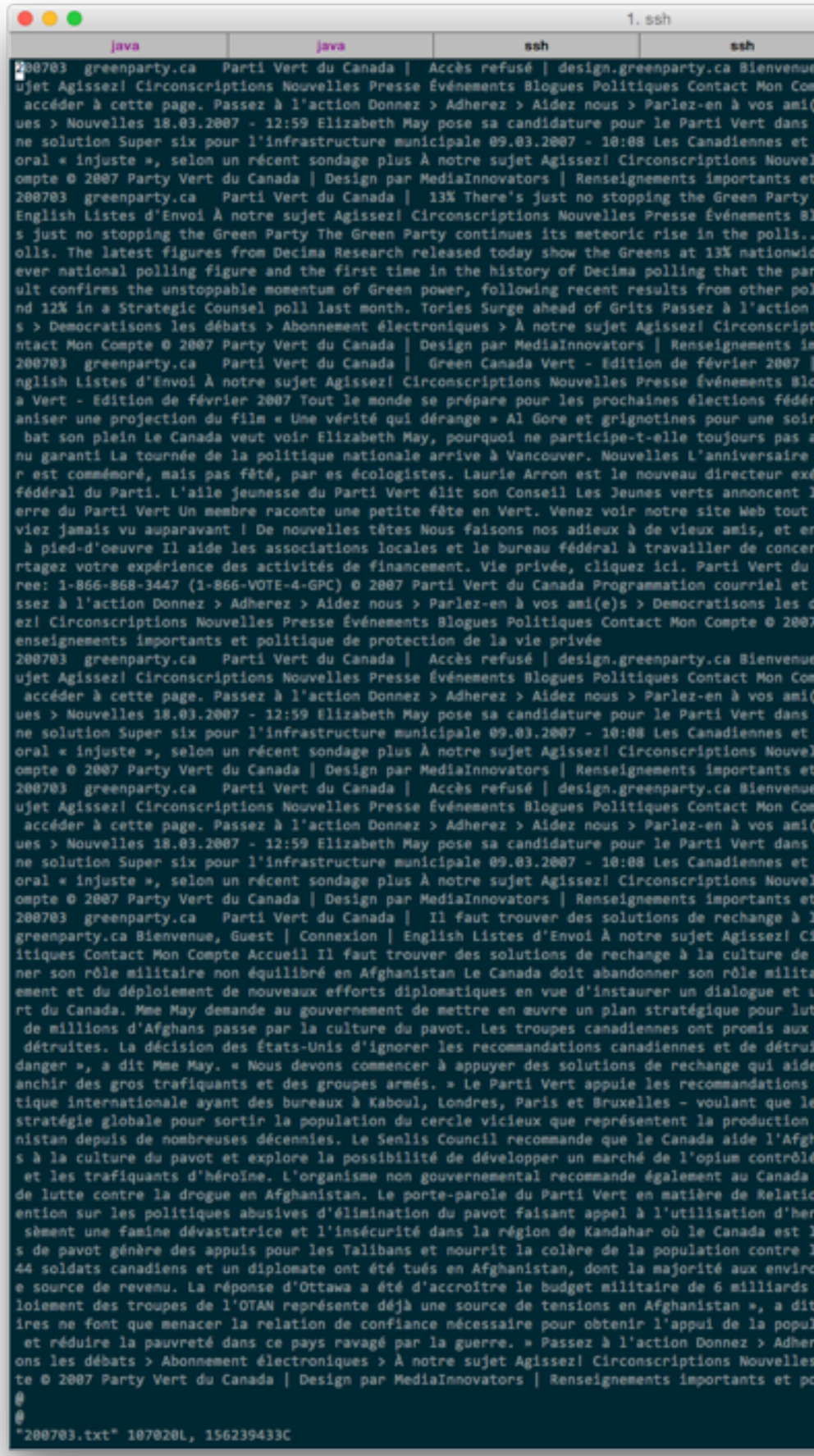
- On a local powerful node (3 Ghz 8-Core Intel Xeon E5/64 GB RAM, data on SSD), about three to four hours per query

- On a cluster, about ~10-20 minutes per query, depending on traffic

Large-Scale Text Analysis

- With Hadoop about 15-20 minutes to extract all plain-text from any specified queries: i.e. all pages belonging to Green Party, Liberal Party, Conservative Party, Council of Canadians, etc.

- Compared to “out of memory”/ go home for an extended weekend on a local node



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

Recipe Book Idea

Using Warcbase to analyze links and full-text

Recipe book:

https://github.com/lintool/warcbase/wiki

NER

October 2005

62476 Stephen Harper

30234 Michael Chong

30109 Gwynne Dyer

28011 ami Entrez

26238 Paul Martin

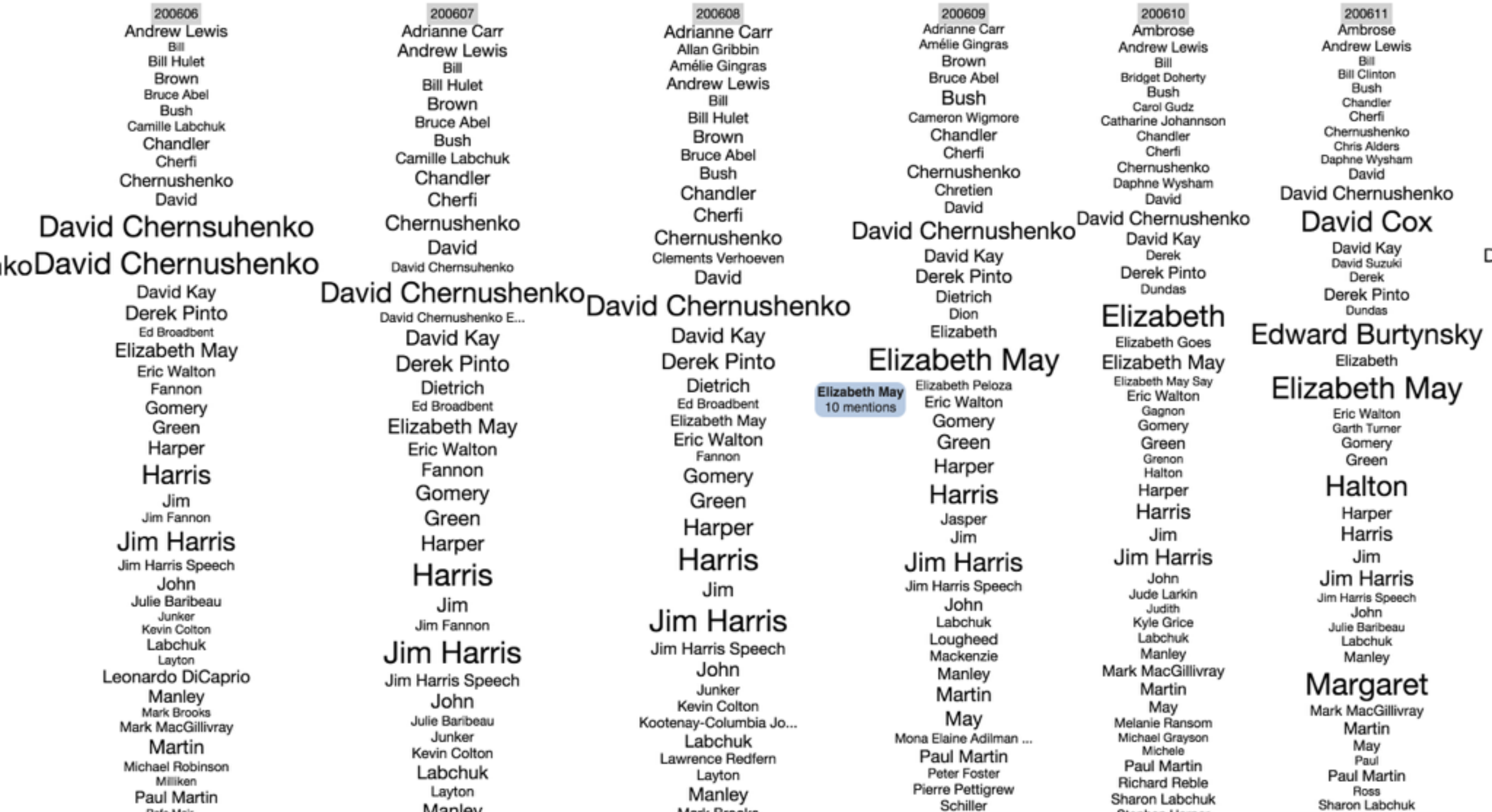
22303 Harper

NER

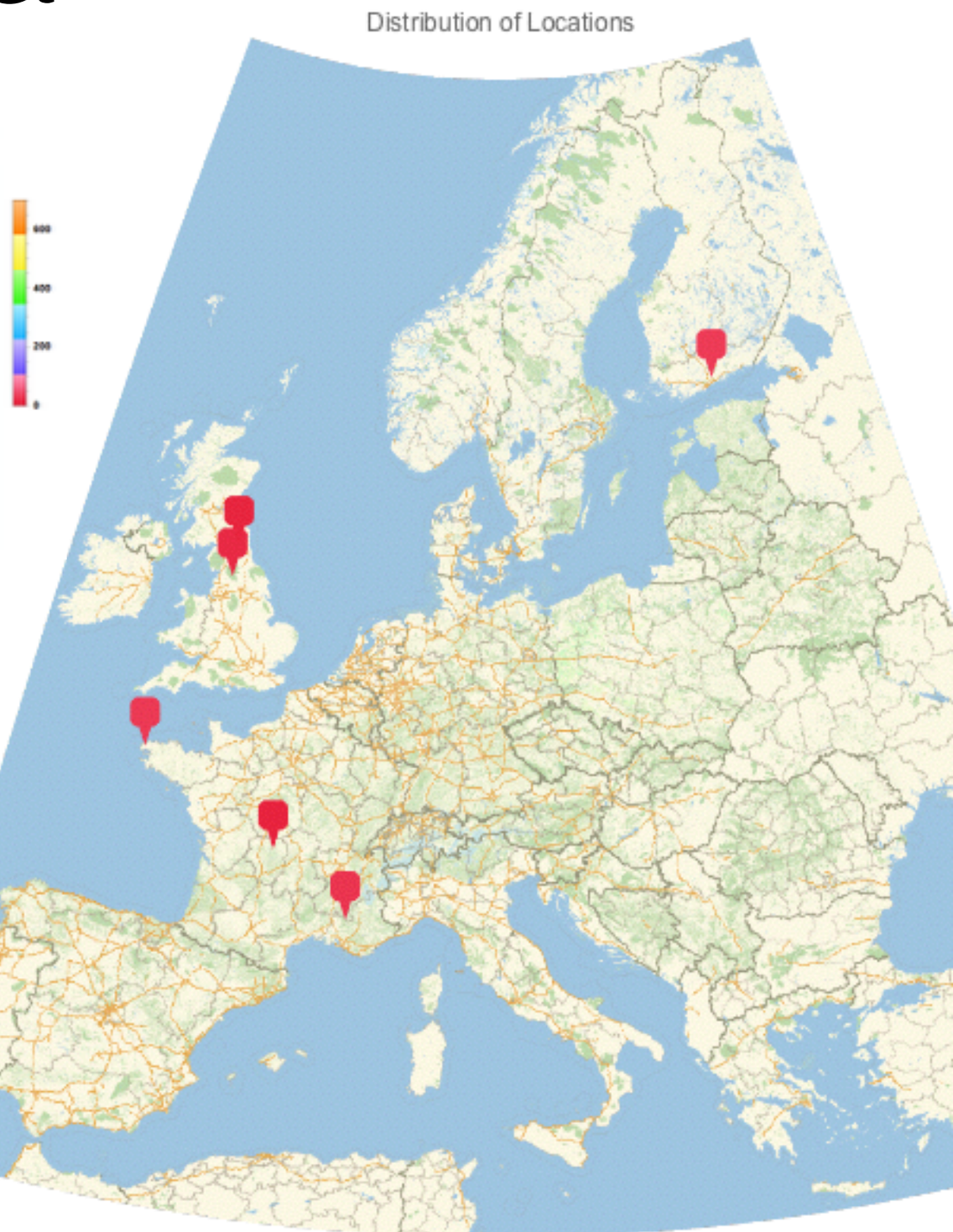
November 2008

3188 Stéphane Dion
2557 Stephen Harper
2471 Stephen HarperLaureen
2410 Dion
2356 Harper

Visualizing NER



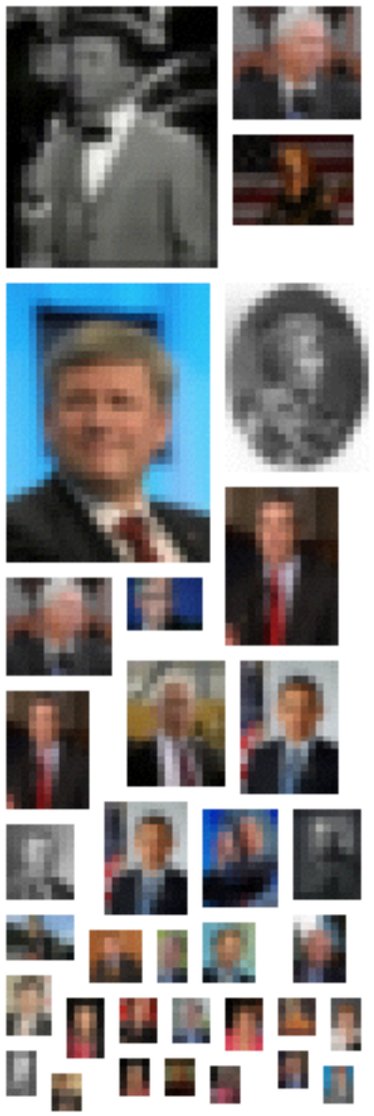
Integration with Wolfram| Alpha

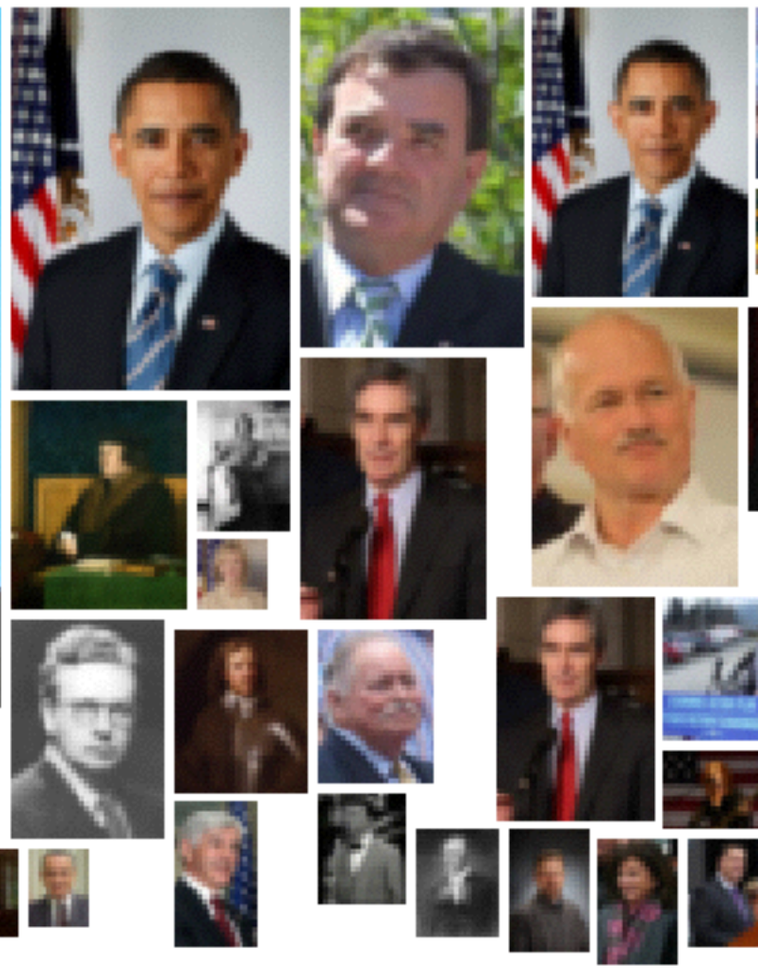



```
In[95]:= int = SemanticInterpretation[#] & /@ processedfreq[[All, 1]]

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

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

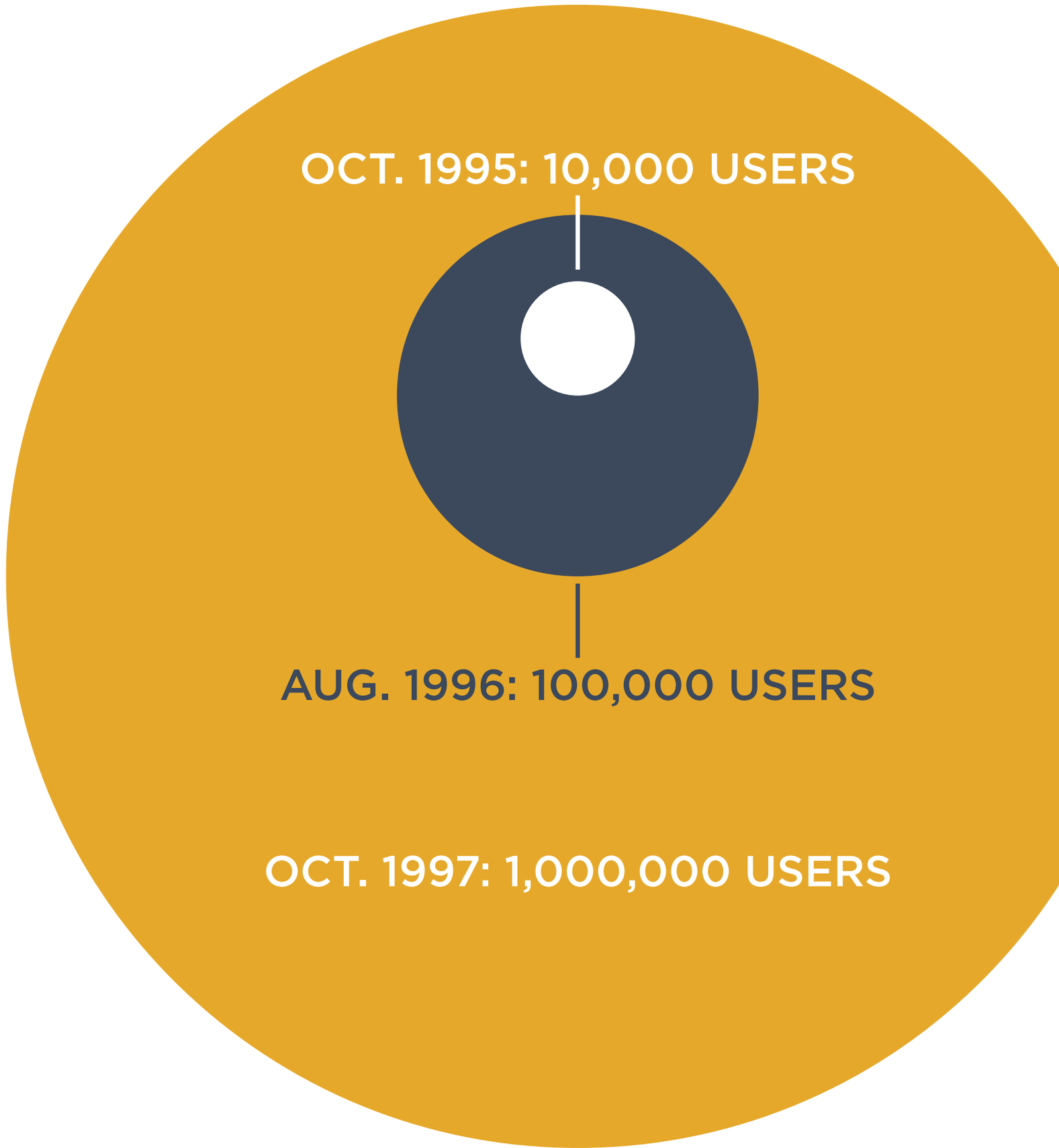
Case Study Three

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

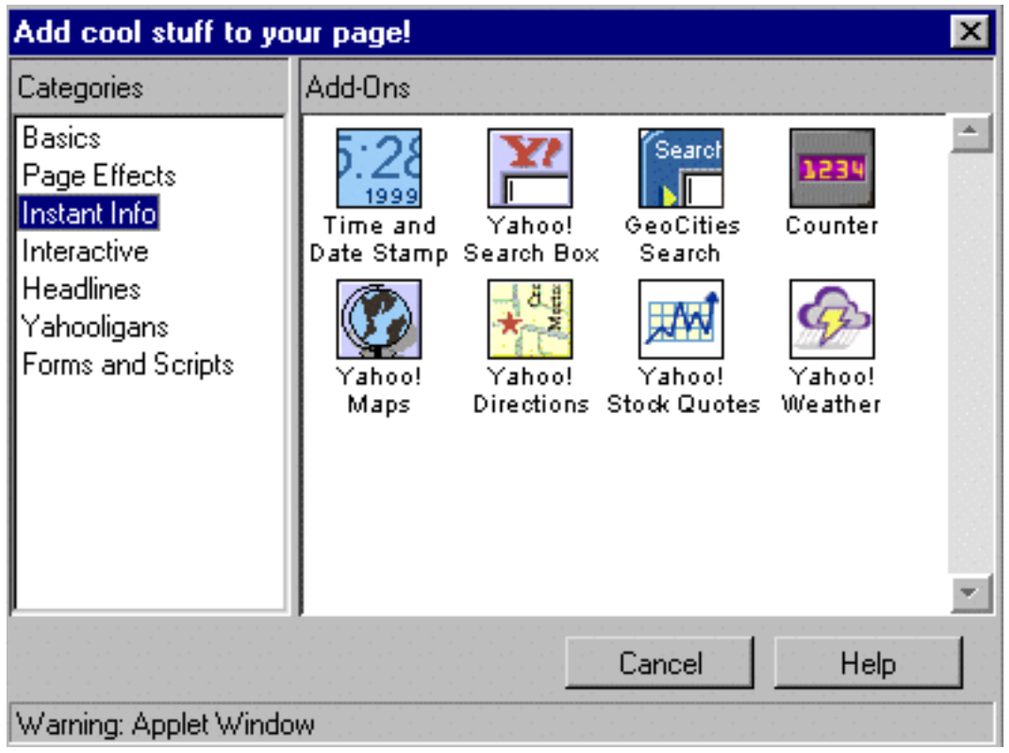
A substantive research question?

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

Ian

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

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.

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

Basic plan	Advanced plan	Premier p
Ideal for small personal sites	Ideal for business sites	Ideal for heavy tra
	POPULAR	

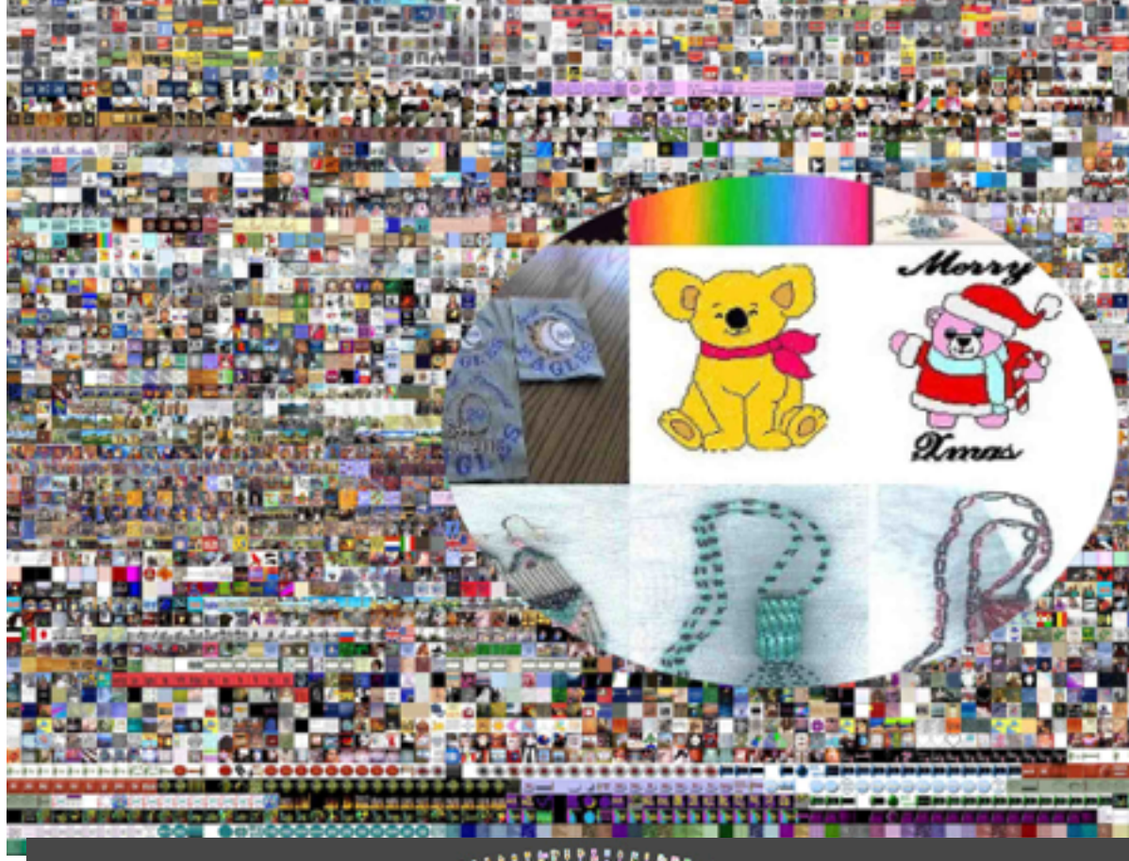

ARCHIVE TEAM
WE ARE GOING TO RESCUE YOUR SHIT

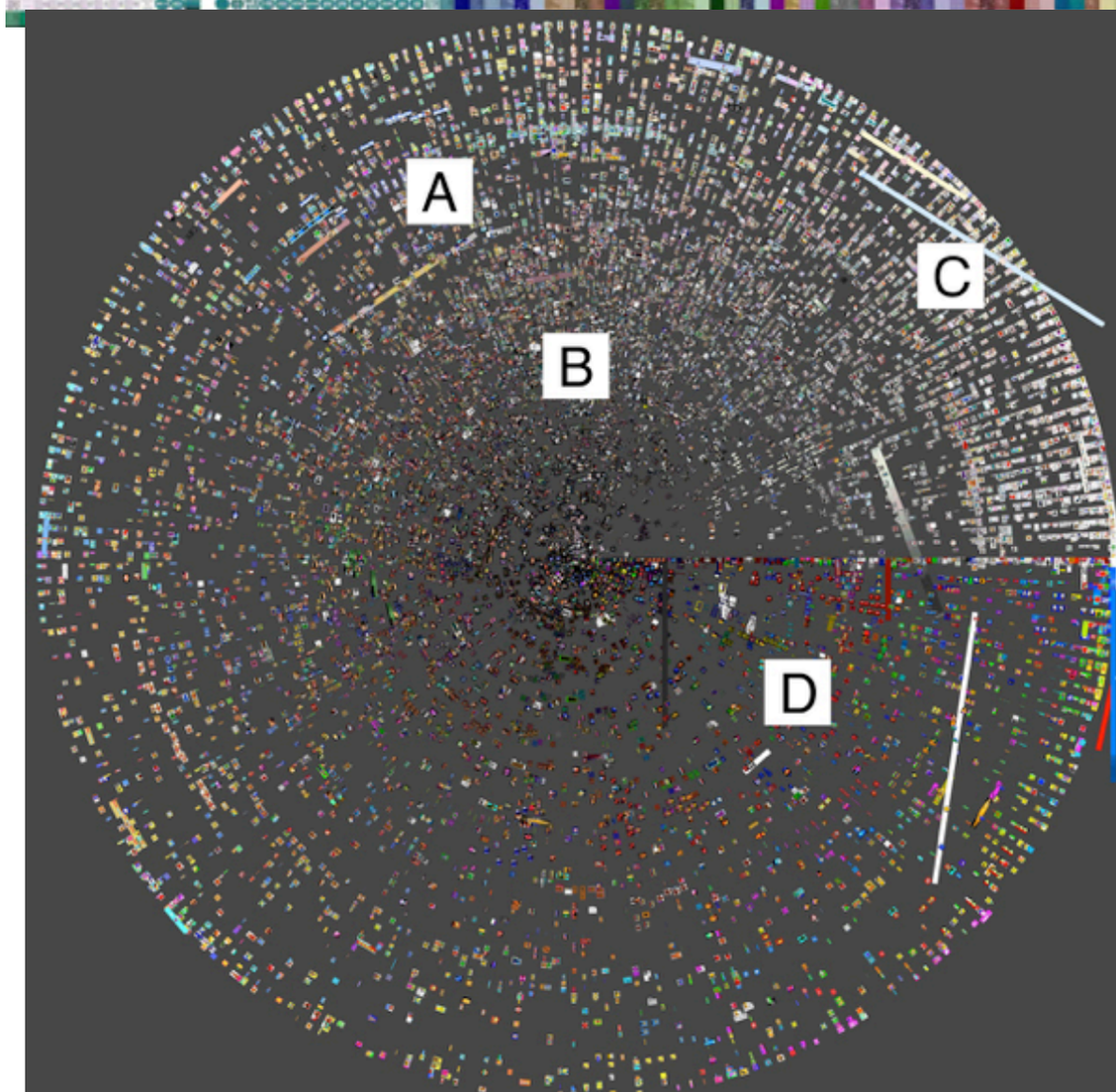
Ethically Navigating the Records of Seven Million People

Topic Modelling Community to Test Coherence

Selected Neighbourhoods	Top Two Topics
Athens *"… based on education, teaching, reading, writing and philosophy".*	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 *"A place for and about kids. Games, stories, educational sites, and homepages created by kids themselves."*	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 *"A family oriented neighborhood that represents Main Street in cyberspace. This is the place to find parenting, pets, and home town values."*	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 *"Entertainment capital of the world. Movies, television, and our live video camera at the corner of Hollywood and Vine!"*	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 *Military men and women.*	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 *"A community with a culture based on gay and lesbian identity."*	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





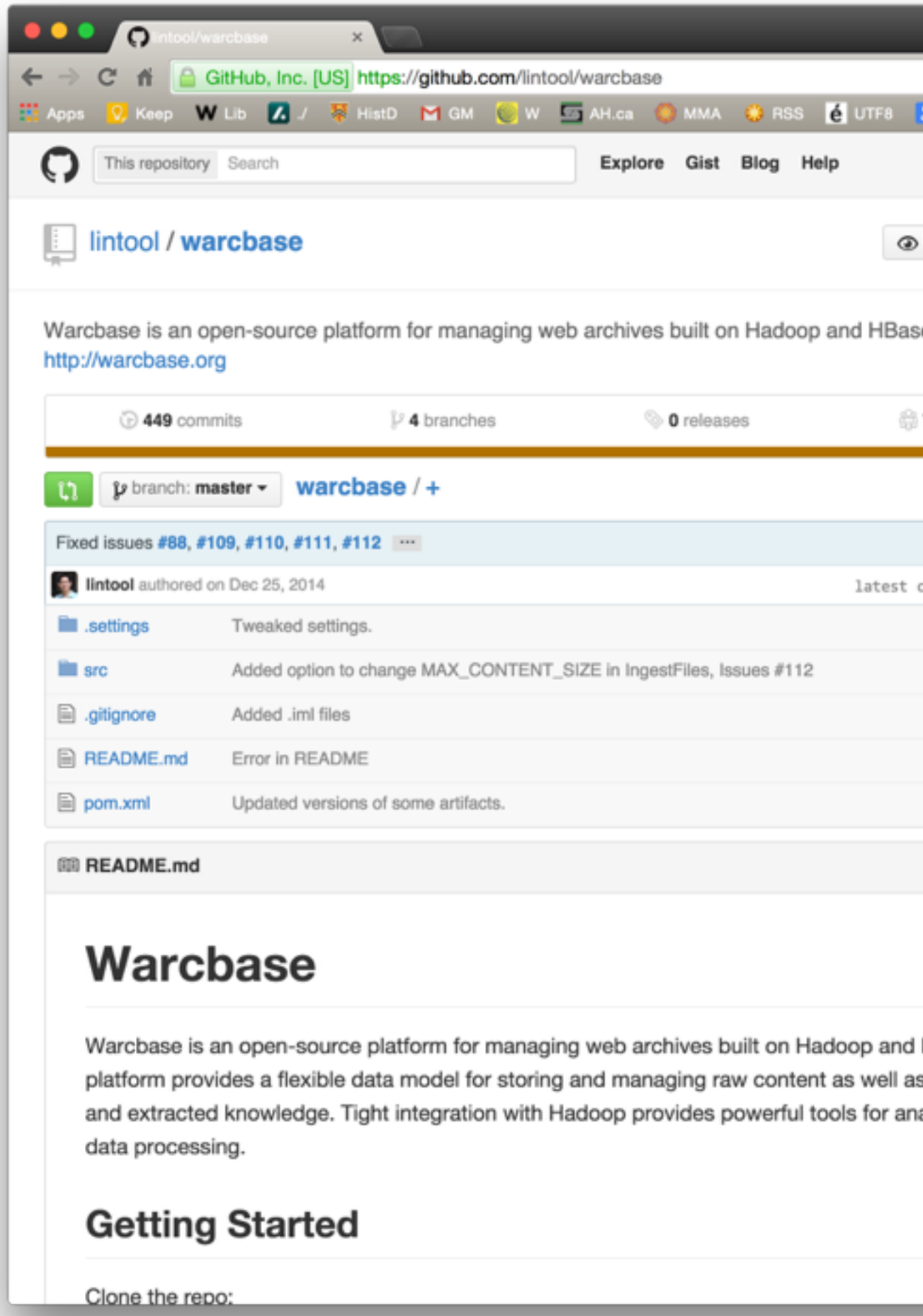
And the stories of significant users and meaningful experiences.

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.



And learning from each other through workshops and hackathons - i.e. Archives Unleashed, 3 -5 March 2016 right next door..

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