Understanding early Web history through three case studies

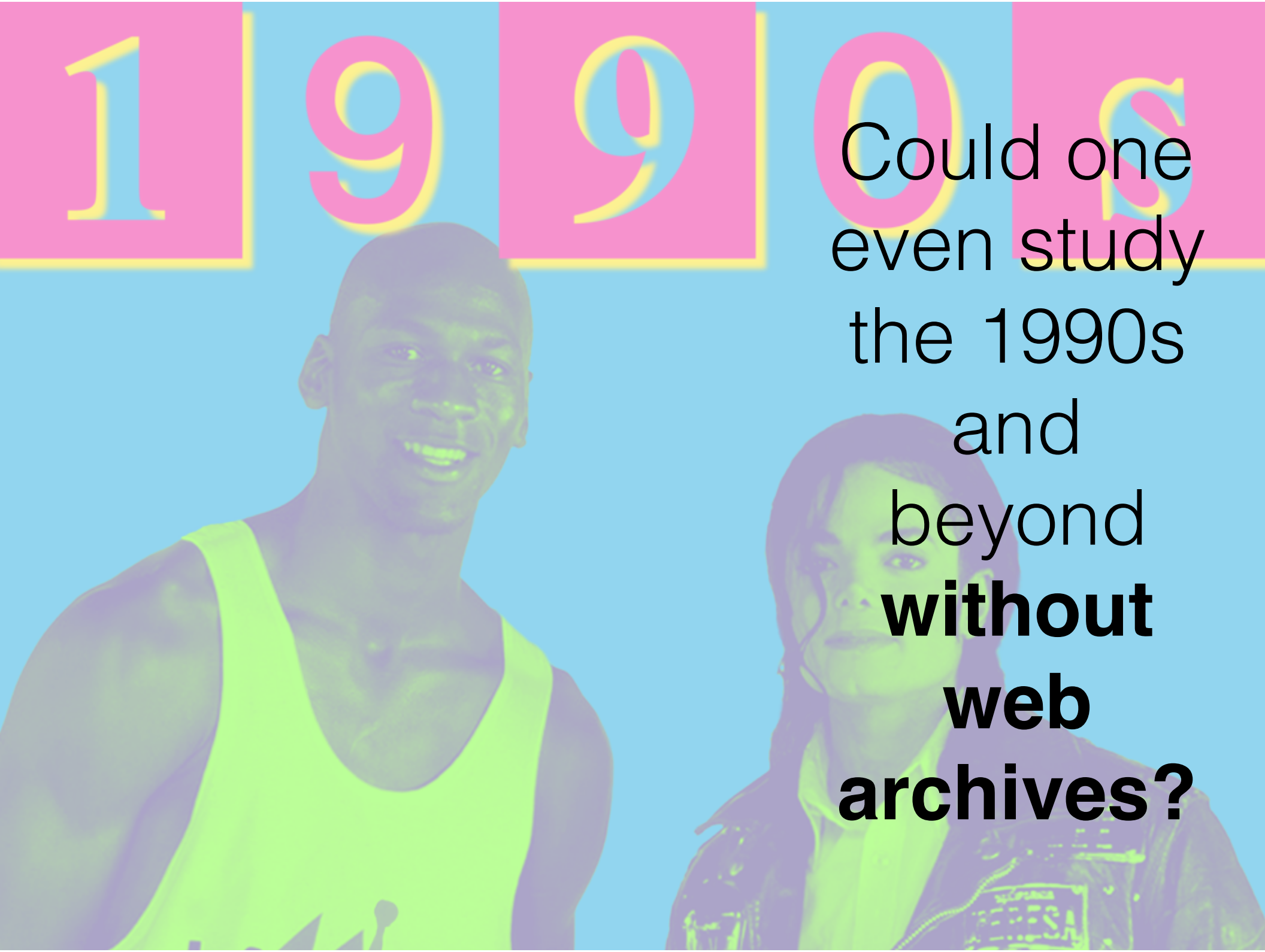
Methodological and technical challenges

Ian Milligan
Assistant Professor
@ianmilligan1

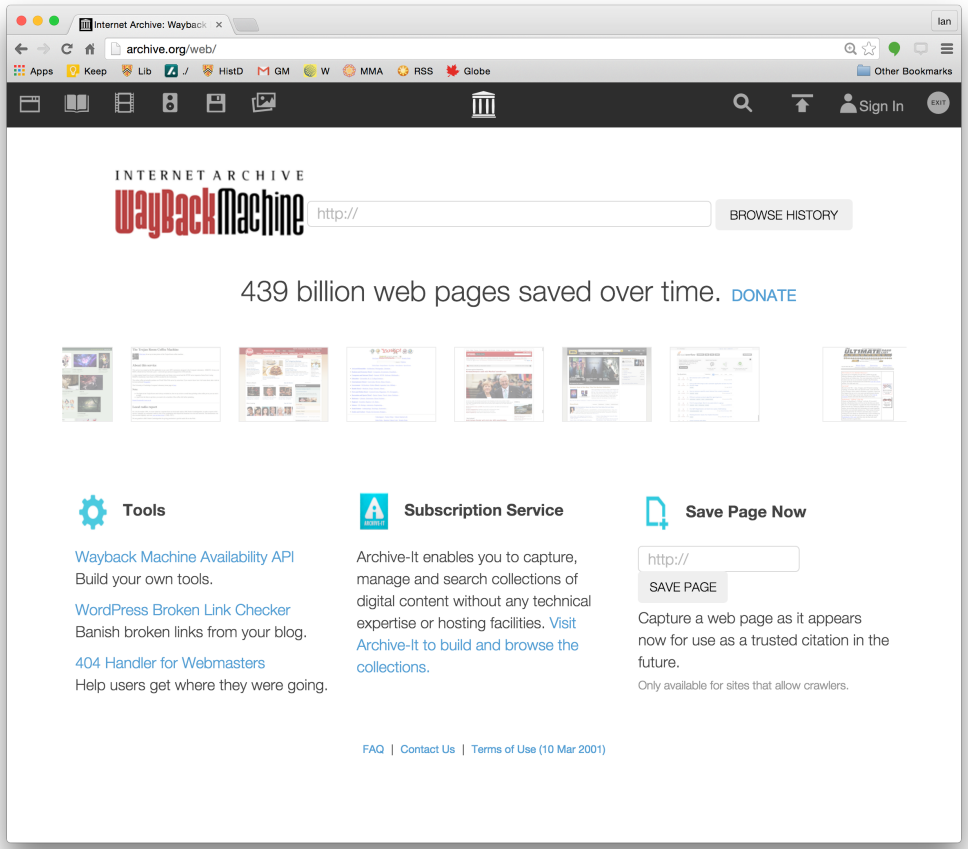
UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Why?

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


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

Nightmare Scenario



This won't be enough!

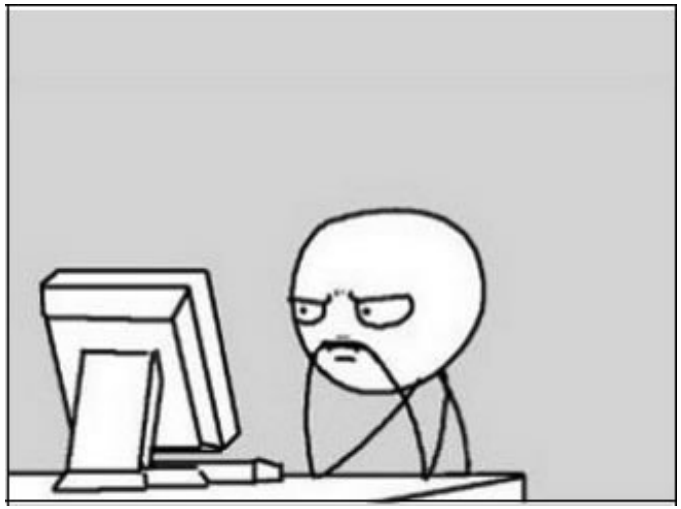
… but what will our search engines look like?

But what will web archives look like?

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

Similarities -

Windows into the lives of everyday people.

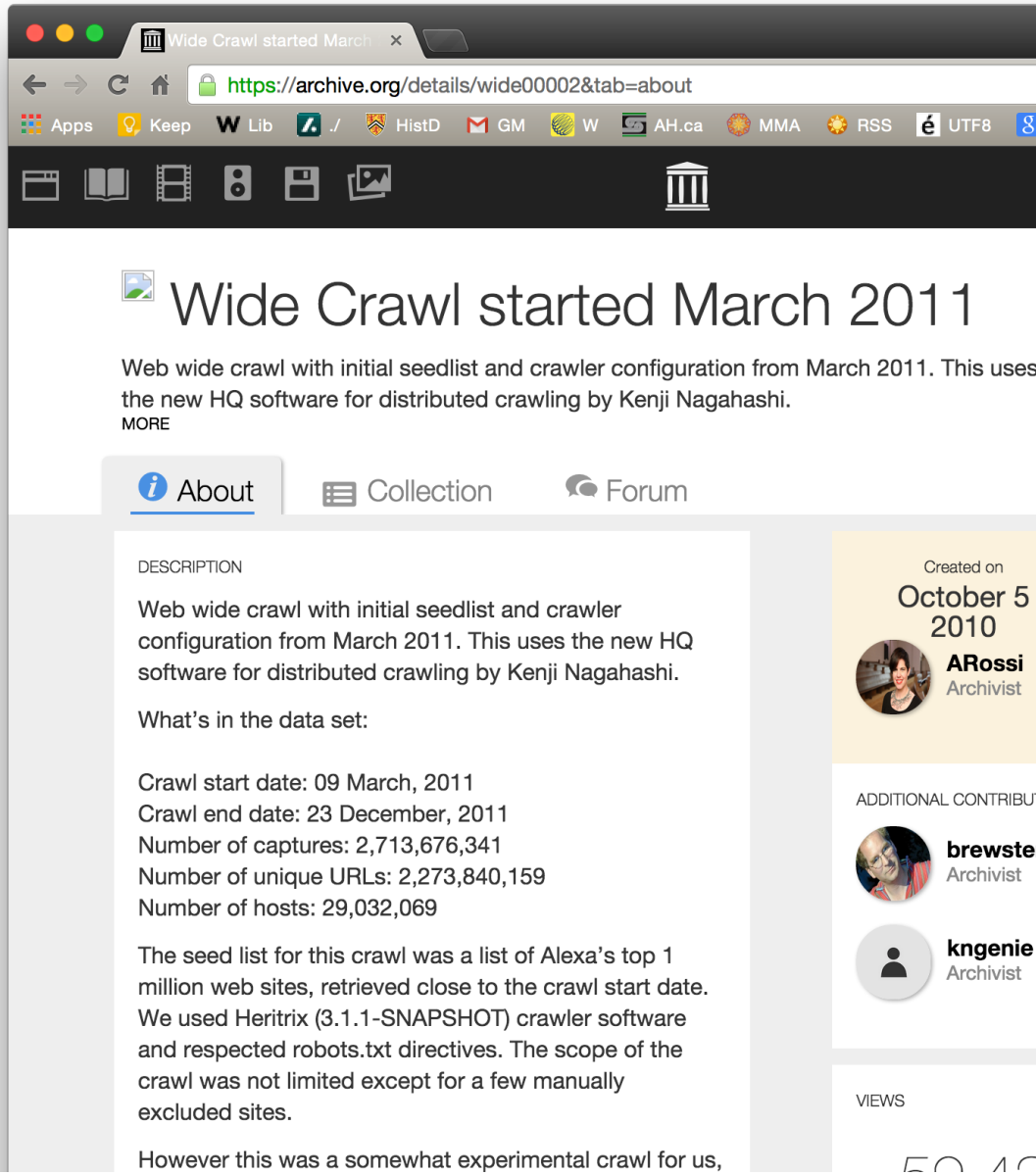


Differences -

Incredible range of technical skills/no common platform!

Case Study One

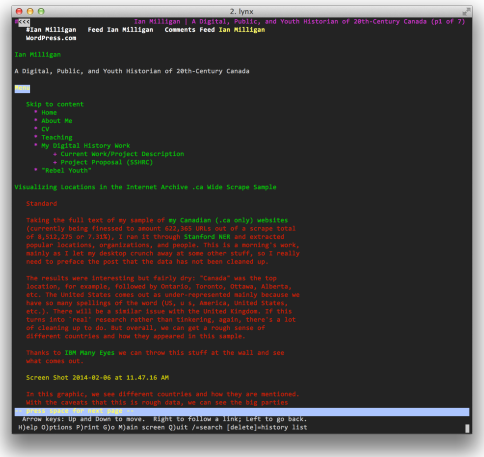
- A handy introduction to WARCs and CDXs
- The **Wide Web Scrape** (~ 80TB)
- **85,570** WARC files, CDX metadata



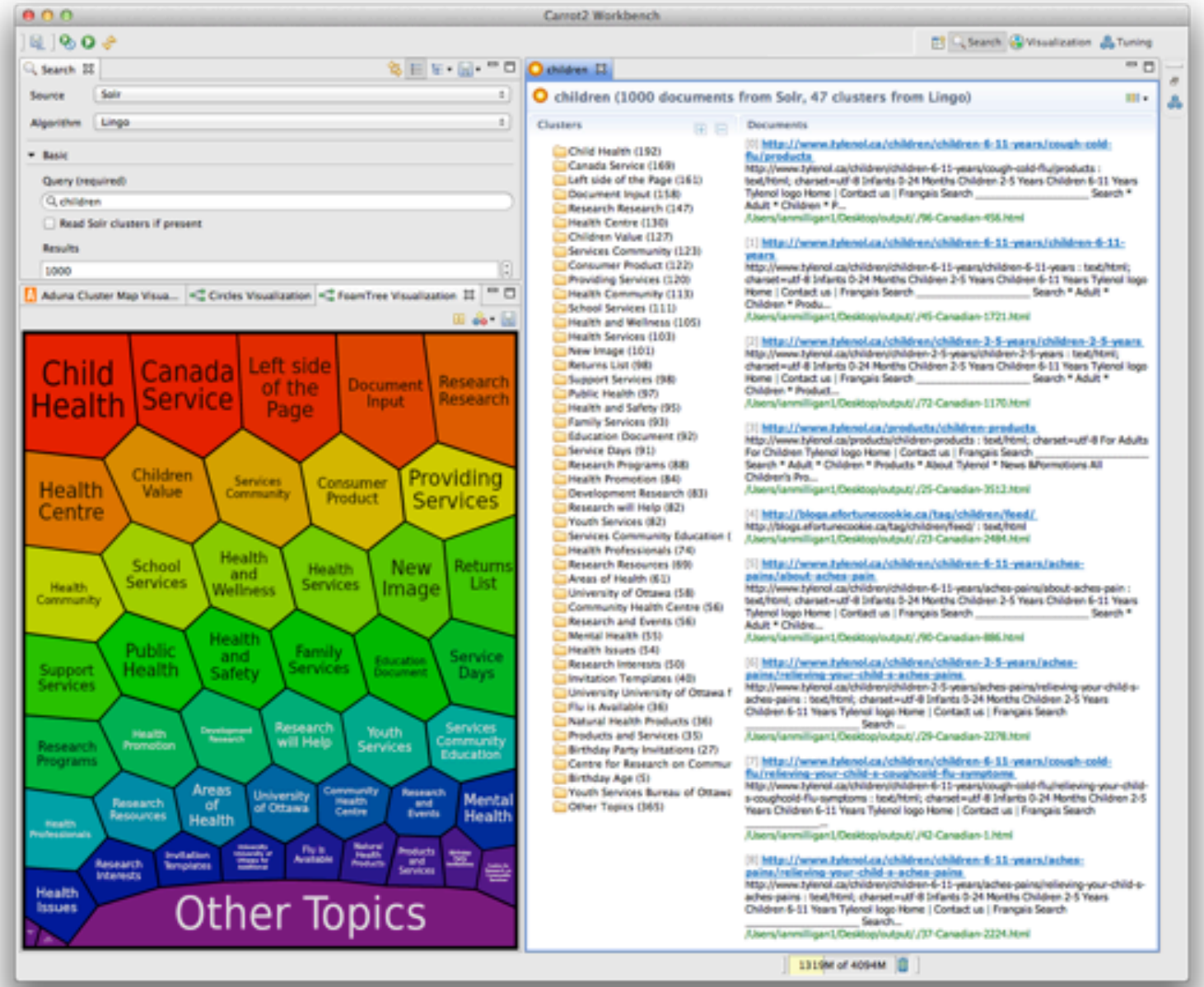
ca,yorku,justlabour)/ 20110714073726
http://www.justlabour.yorku.ca/ text/html
302 3I42H3S6NNFQ2MSVX7XZKYAYSCX5QBYJ
http://www.justlabour.yorku.ca/index.php?
page=toc&volume=16 - 462 880654831
WIDE-20110714062831-crawl416/
WIDE-20110714070859-02373.warc.gz

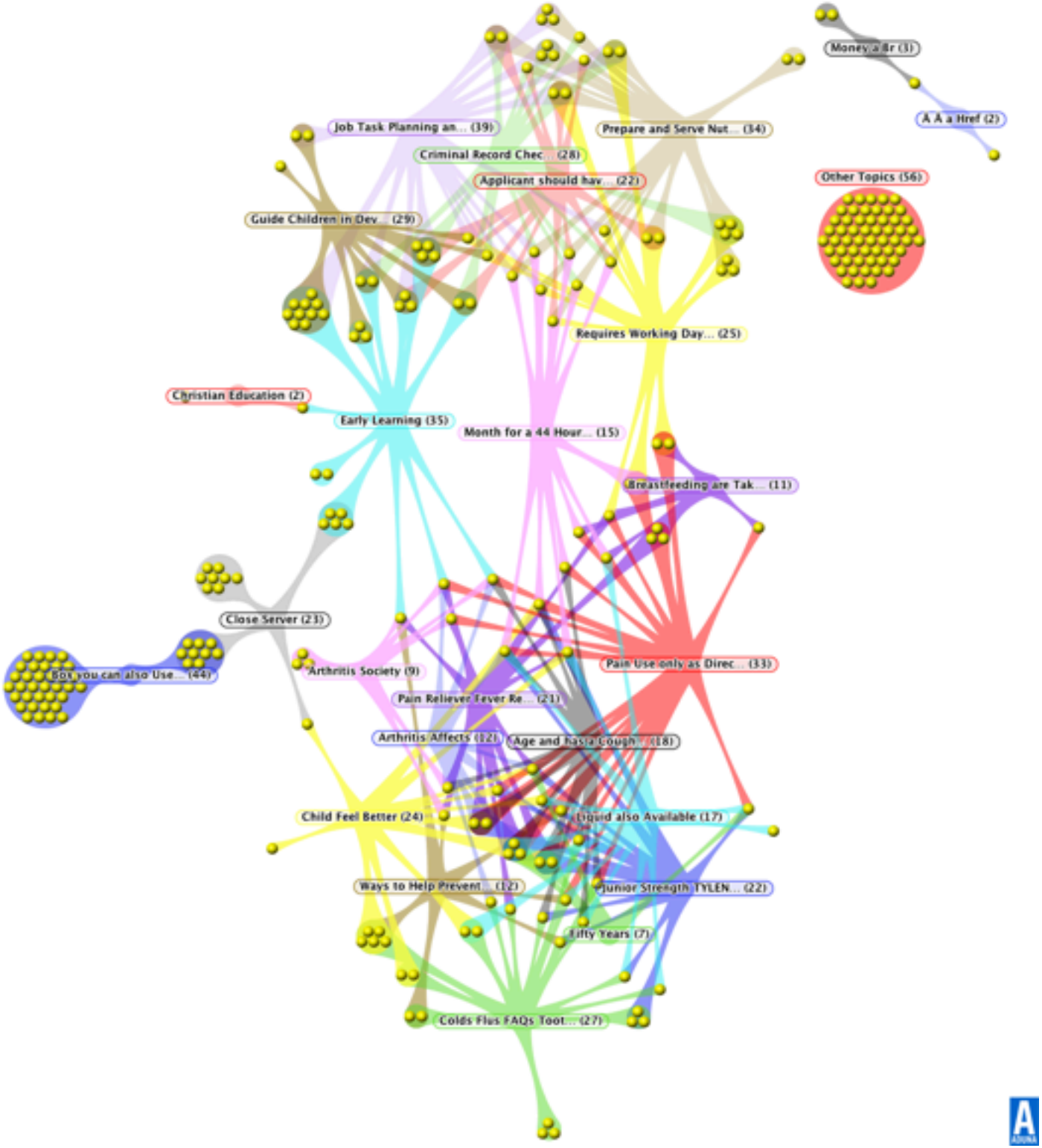
Top-Level Domain	Number of Distinct URLs Downloaded in Sample	Number of Overall URLs in Wide Web Scrape (selected domains)	Percentage of URLs Captured
.com	29,219,706	1,260,409,874	2.32%
.org	2,489,050	96,681,268	2.57%
.net	2,438,903	140,726,805	1.73%
.edu	350,482	6,620,283	5.29%
.gov	97,484	2,205,332	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%**

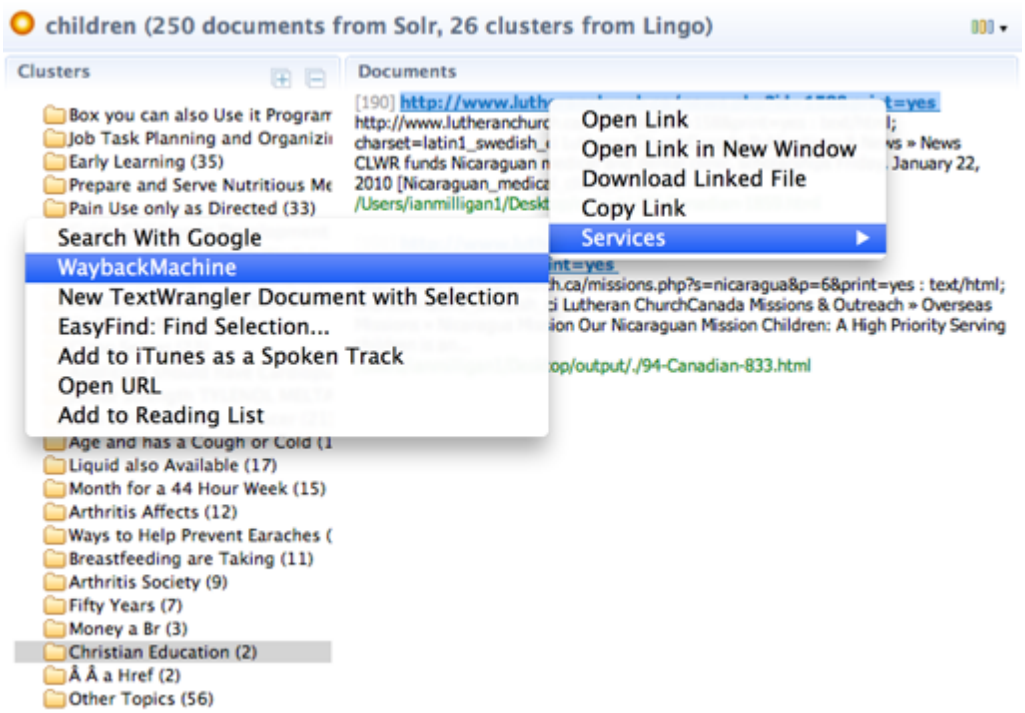
CDX Files (finding aids)



WARC File **Plain Text** **Indexing**

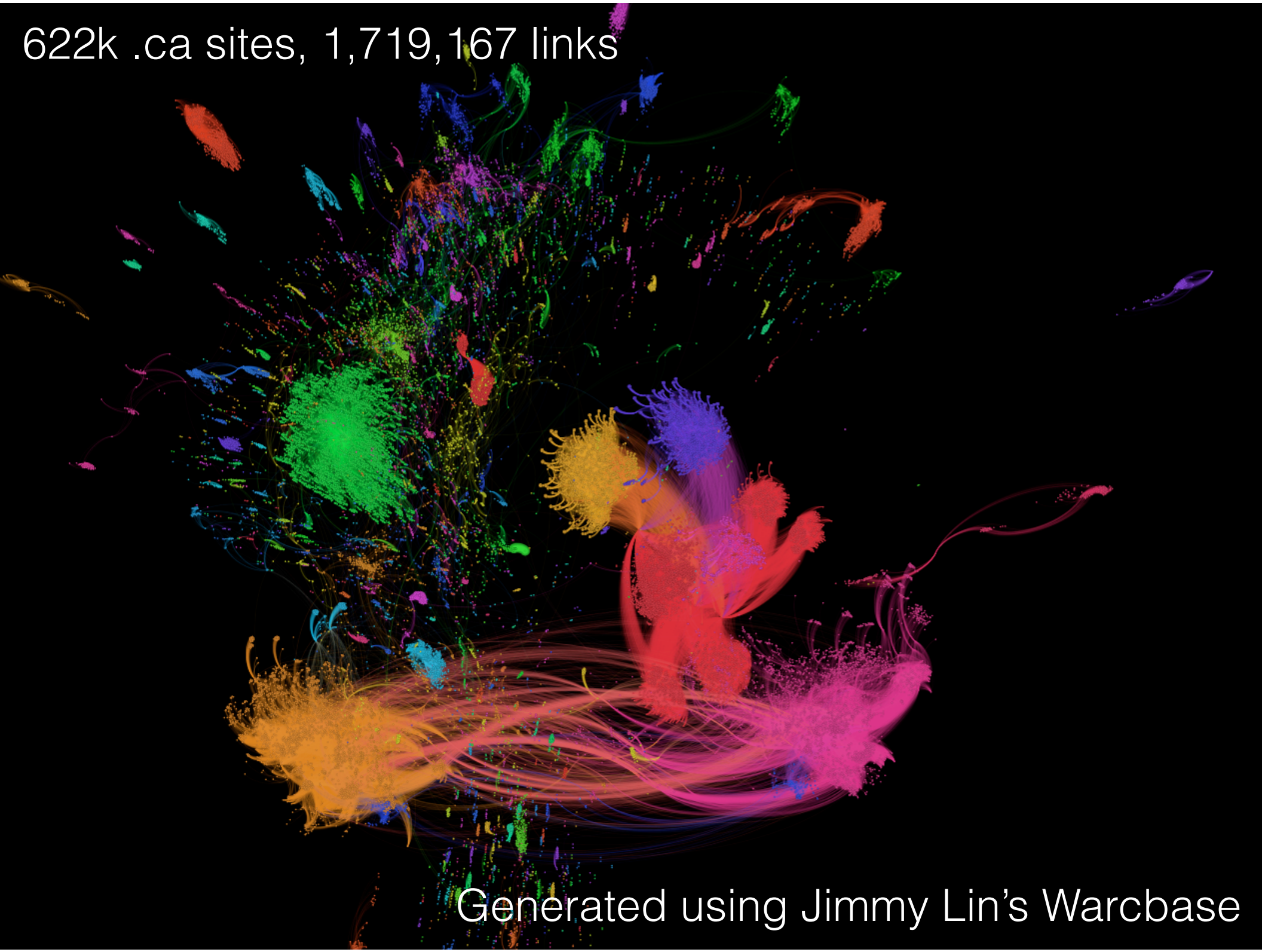

Carrot2 Workbench
Search
Visualization
Tuning
Source
Solr
Algorithm
Lingo
Basic
Query (required)
children
Read Solr clusters if present
Results
1000
Aduna Cluster Map Visua...
Circles Visualization
FoamTree Visualization
Child Health
Canada Service
Left side of the Page
Document Input
Research Research
Health Centre
Children Value
Services Community
Consumer Product
Providing Services
Health Community
School Services
Health and Wellness
Health Services
New Image
Returns List
Public Health
Health and Safety
Family Services
Education Document
Service Days
Support Services
Research Programs
Health Promotion
Development Research
Research will Help
Youth Services
Services Community Education
Areas of Health
University of Ottawa
Community Health Centre
Research and Events
Mental Health
Research Resources
Health Professionals
Research Interests
Invitation Templates
Flu 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
1319M of 4094M

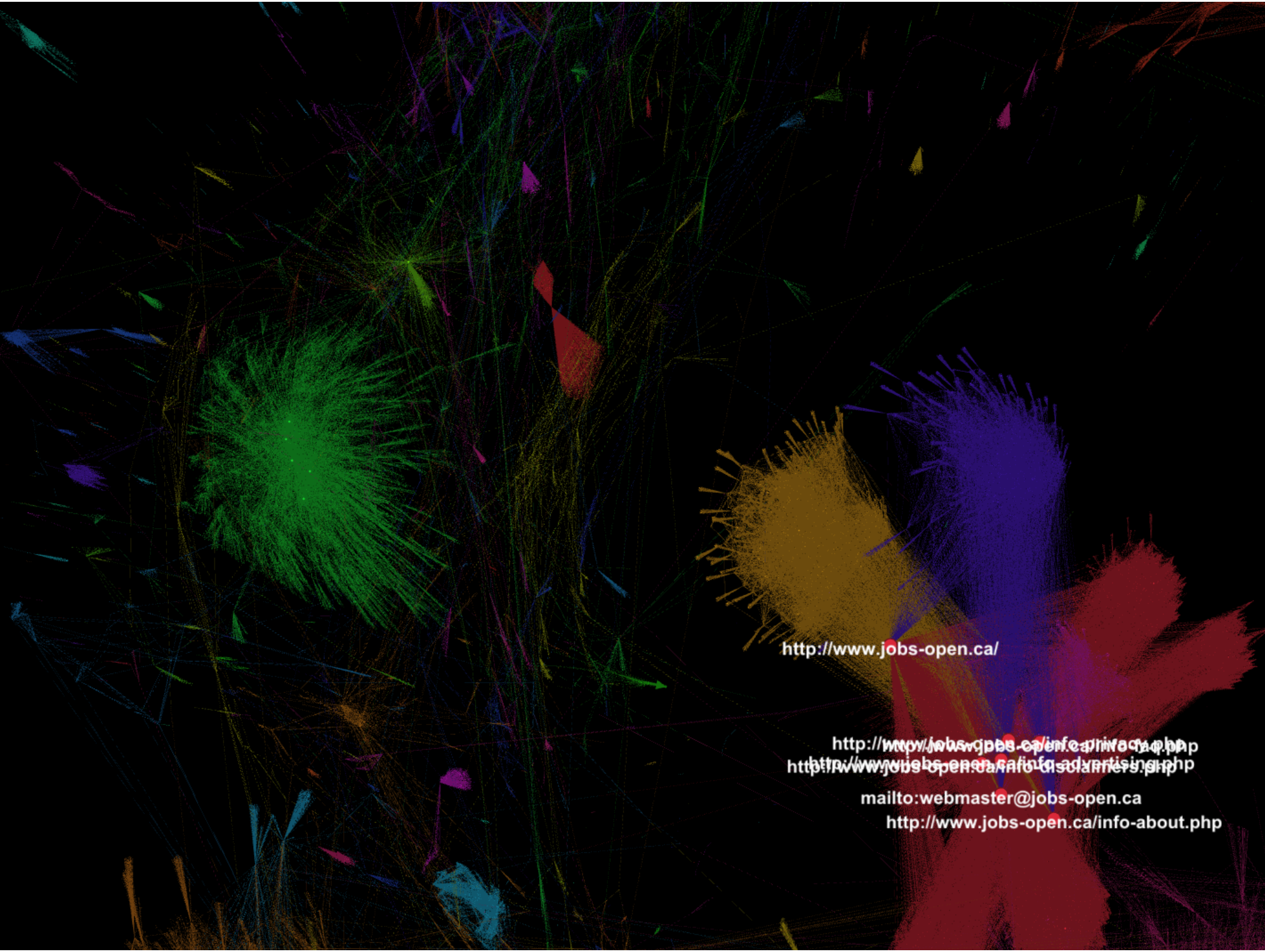

Money a Br (3)
A A a Href (2)
Other Topics (56)
Job Task Planning an... (39)
Prepare and Serve Nut... (34)
Criminal Record Chec... (28)
Applicant should hav... (22)
Guide Children in Dev... (29)
Requires Working Day... (25)
Christian Education (2)
Early Learning (35)
Month for a 44 Hour... (15)
Breastfeeding are Tak... (11)
Close Server (23)
Bot 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 n
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
January 22,
h.ca/missions.php?s=nicaragua&p=6&print=yes : text/html;
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
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.

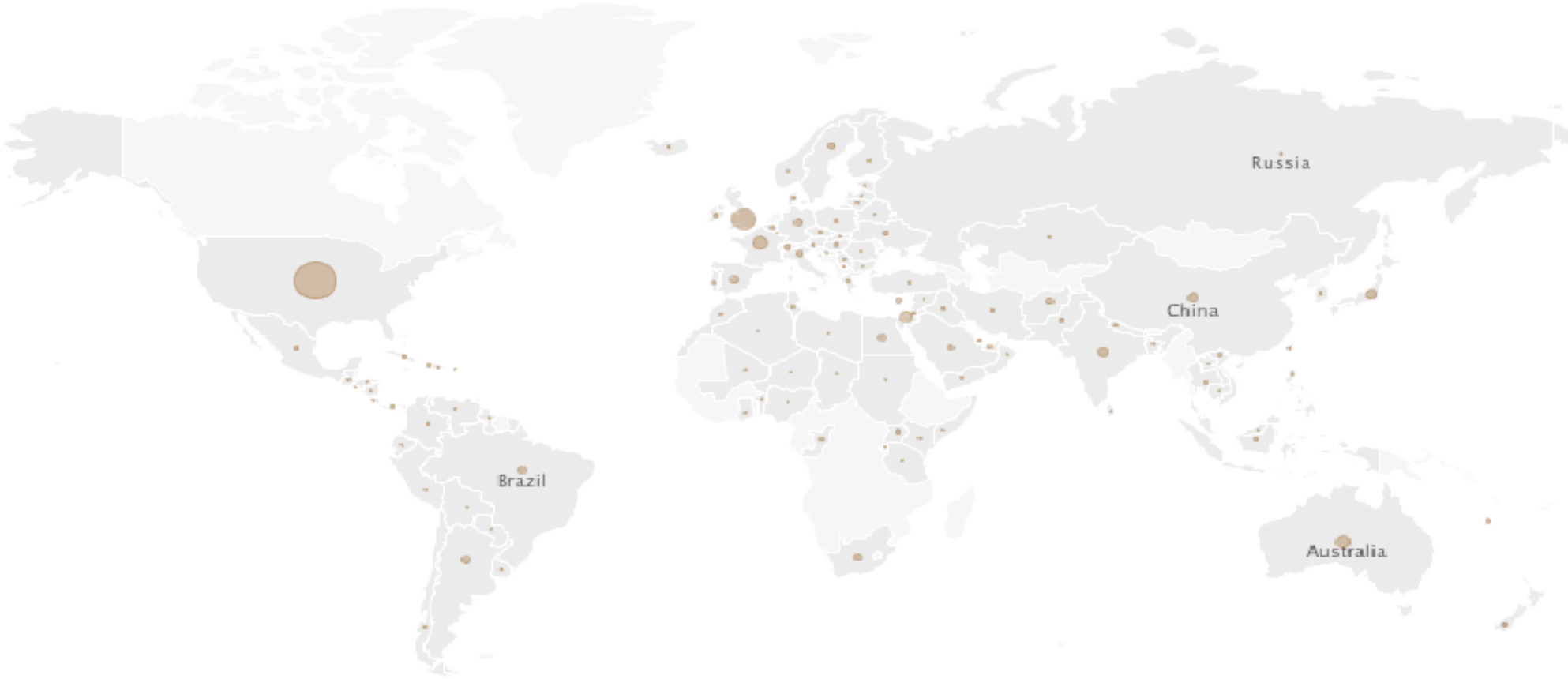
Problem is.. you need to know what you’re looking for!


622k .ca sites, 1,719,167 links
Generated using Jimmy Lin's Warcbase

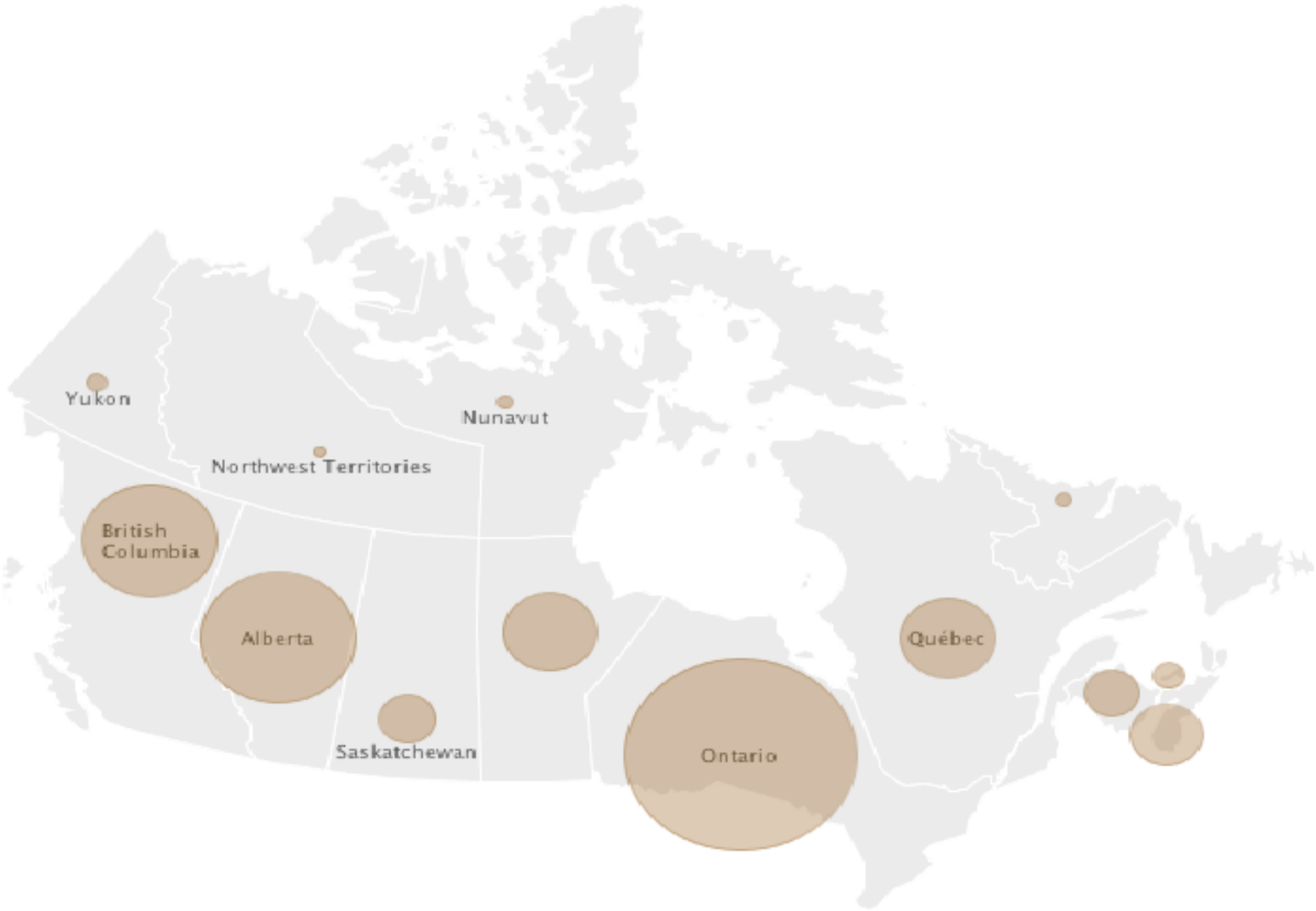

http://www.jobs-open.ca/
http://www.jobs-open.ca/info-privacy.php
http://www.jobs-open.ca/info-disclaimers.php
mailto:webmaster@jobs-open.ca
http://www.jobs-open.ca/info-about.php


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.htm
http://www.admission.uottawa.ca/Default.aspx?tabid=2548&sourceqp
http://www.uottawa.ca/contactez-nous.html
http://www.uottawa.ca/requirer.html
http://www.uottawa.ca/Default.aspx?tabid=2672
https://web9.uottawa.ca/services/SiteValidation/a/student/guide/francais/?section/bourses/
http://www.uottawa.ca/icone-recherche/

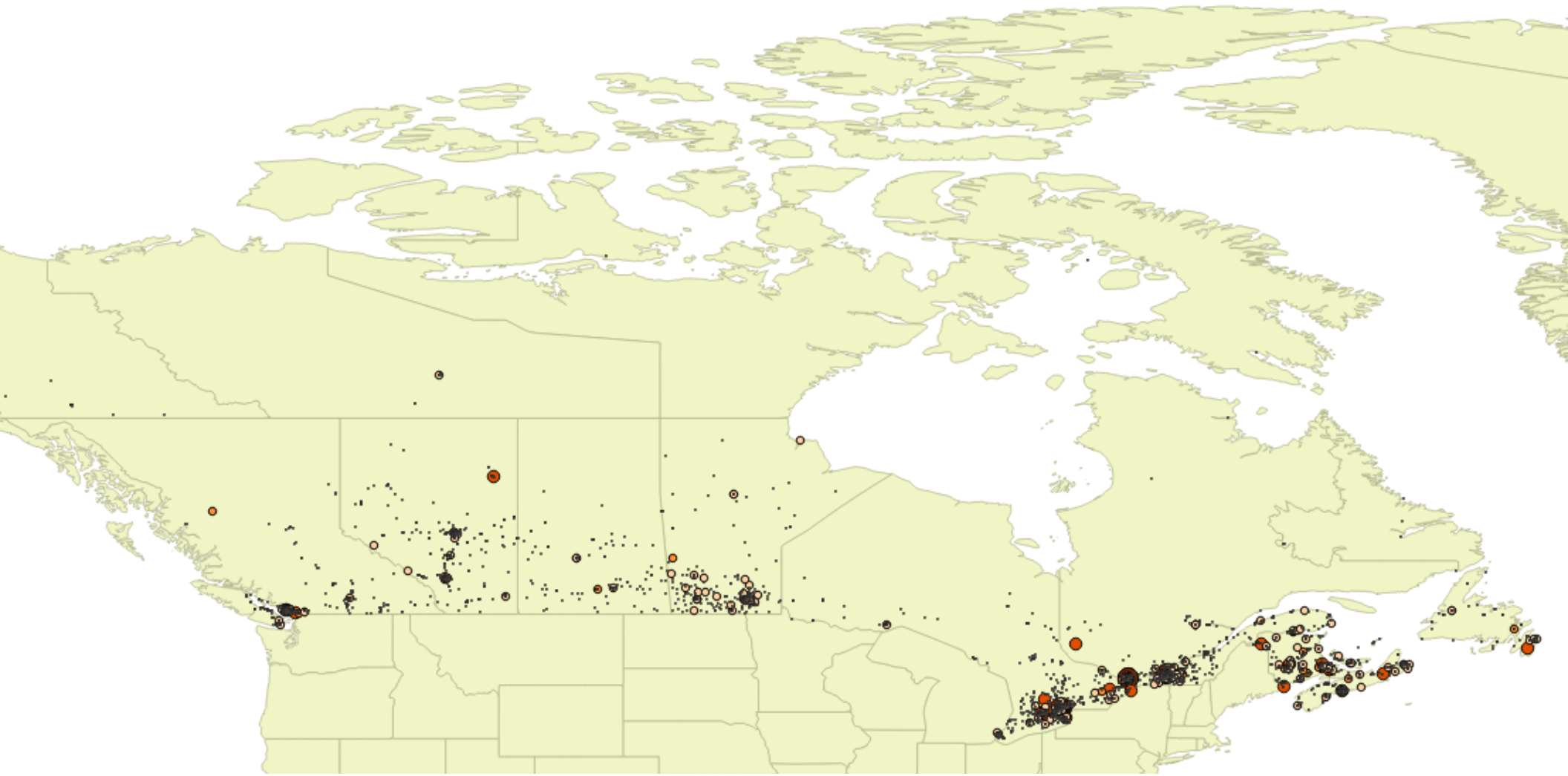
Countries Mentioned in .ca TLD (excluding Canada)



Provinces Mentioned in .ca TLD



Canadian Postal Codes visualized

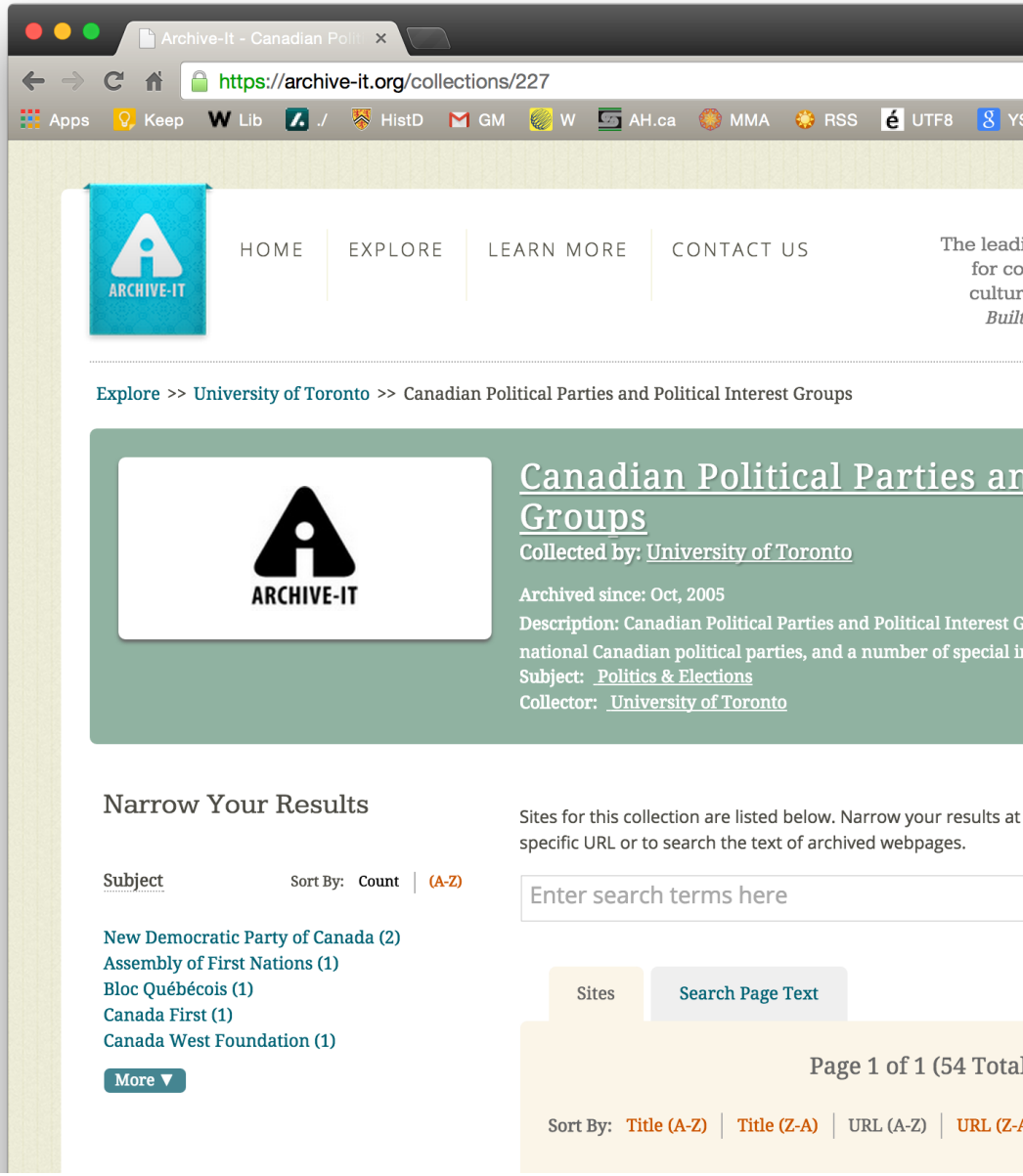


Need longitudinal, but the size/intensity = extreme.

Wide Web Scrapes 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

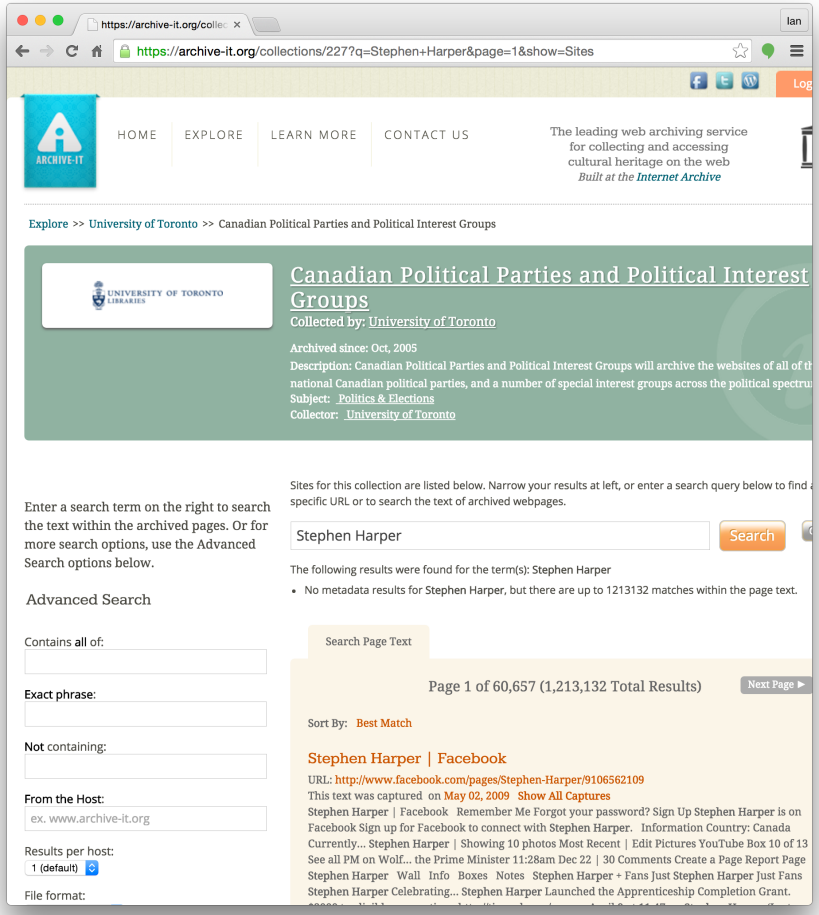


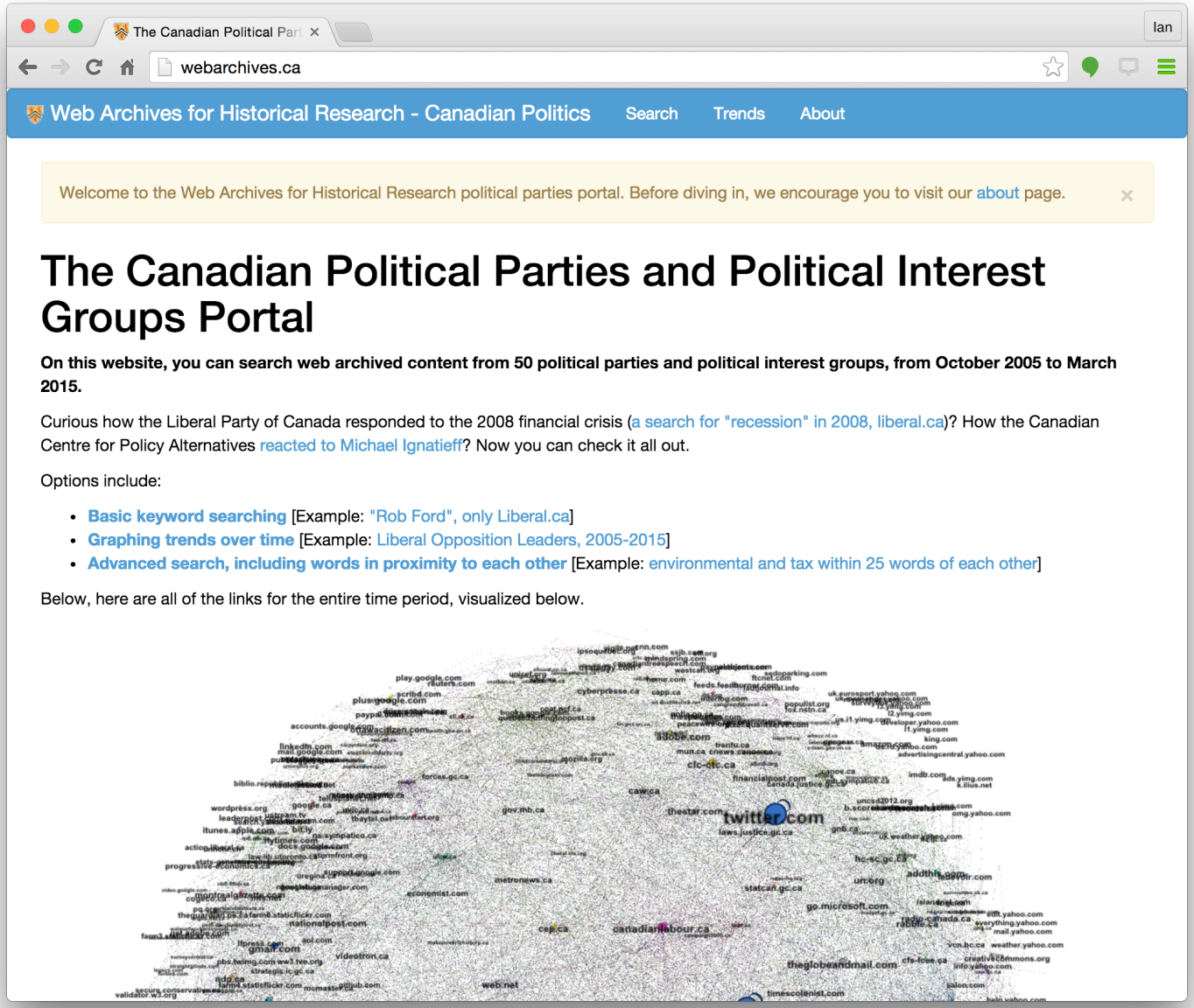
Pivotal Changes in Canadian Politics, 2005-2015

- 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!





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

Trend results 2005-2015 fo ×

webarchives.ca/graph

Web Archives for Historical Research - Canadian Politics

Search Trends About

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

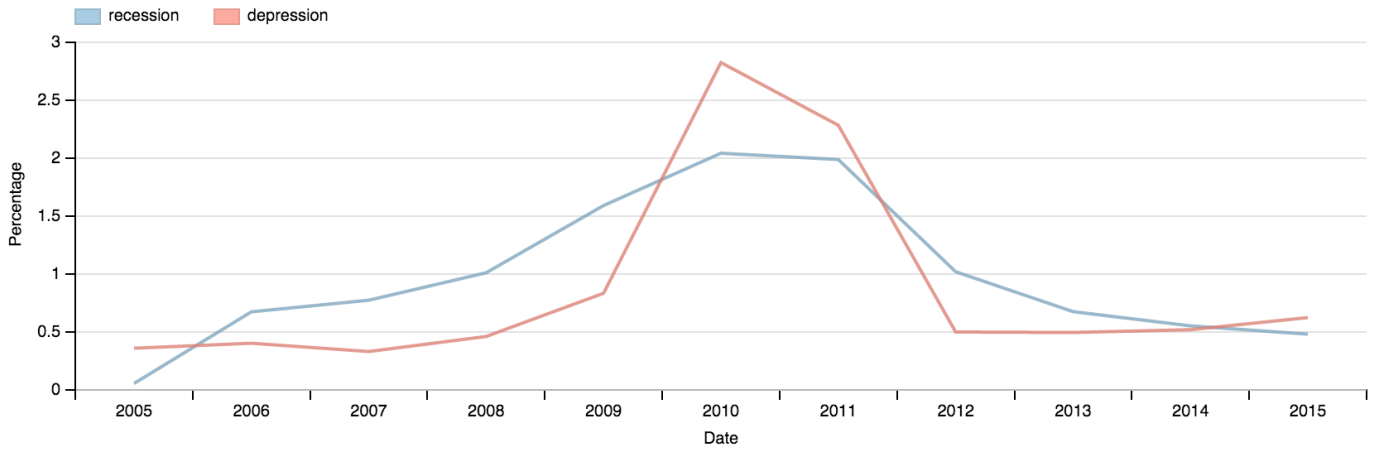
The graph shows the relative frequency of a term in each year - i.e. how often did "recession" appear in 2008 as a percentage of all keywords? Enter your search terms and click "Update Graph." To see actual results, click on the graph's line where it intersects a year.

You can also use this graph to run keyword searches on particular top-level domains. To see how often the word "global warming" appeared in the Conservative Party of Canada's domain versus the New Democratic Party of Canada, search with the query: "global warming" domain:conservative.ca, "global warming" domain:ndp.ca [view this example].

For a full list of domains, consult the about page.

Colours are assigned to the graph below depending on the order they are inputted (i.e. the Conservative Party of Canada will not always be blue).

recession,depression | 2005 | 2015 | Update Graph | Reset All



Some Sample Queries to Get Started With:

- "Public Transit": Liberals/NDP/Conservatives. We use quotation marks to search on the exact phrase.
- "Tar Sands": Amongst the three major parties, who uses and subsequently stops using this phrase for Canada's Oil Sands?
- "Climate Change": Similar to the above two queries, we can see this issue rise and possibly wane in public interest.

Feel free to take these examples and adapt them to your own purposes. If you find something interesting, please let us know!

Trend results 2005-2015 f

webarchives.ca/graph?query="public+transit"+domain%3Aconservative.ca%2C+"public+transit"+domain%3Andp.ca...

Web Archives for Historical Research - Canadian Politics

Search Trends About

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

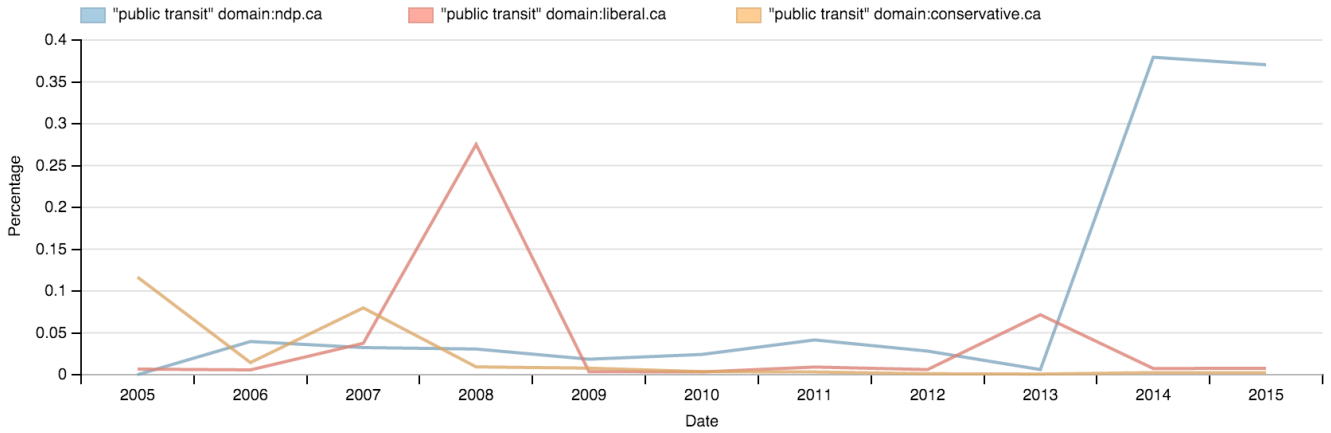
The graph shows the relative frequency of a term in each year - i.e. how often did "recession" appear in 2008 as a percentage of all keywords? Enter your search terms and click "Update Graph." To see actual results, click on the graph's line where it intersects a year.

You can also use this graph to run keyword searches on particular top-level domains. To see how often the word "global warming" appeared in the Conservative Party of Canada's domain versus the New Democratic Party of Canada, search with the query: "global warming" domain:conservative.ca, "global warming" domain:ndp.ca [view this example].

For a full list of domains, consult the about page.

Colours are assigned to the graph below depending on the order they are inputted (i.e. the Conservative Party of Canada will not always be blue).

"public transit" domain:conservative.ca, "p | 2005 | 2015 | Update Graph | Reset All



Some Sample Queries to Get Started With:

- "Public Transit": Liberals/NDP/Conservatives. We use quotation marks to search on the exact phrase.
- "Tar Sands": Amongst the three major parties, who uses and subsequently stops using this phrase for Canada's Oil Sands?
- "Climate Change": Similar to the above two queries, we can see this issue rise and possibly wane in public interest.

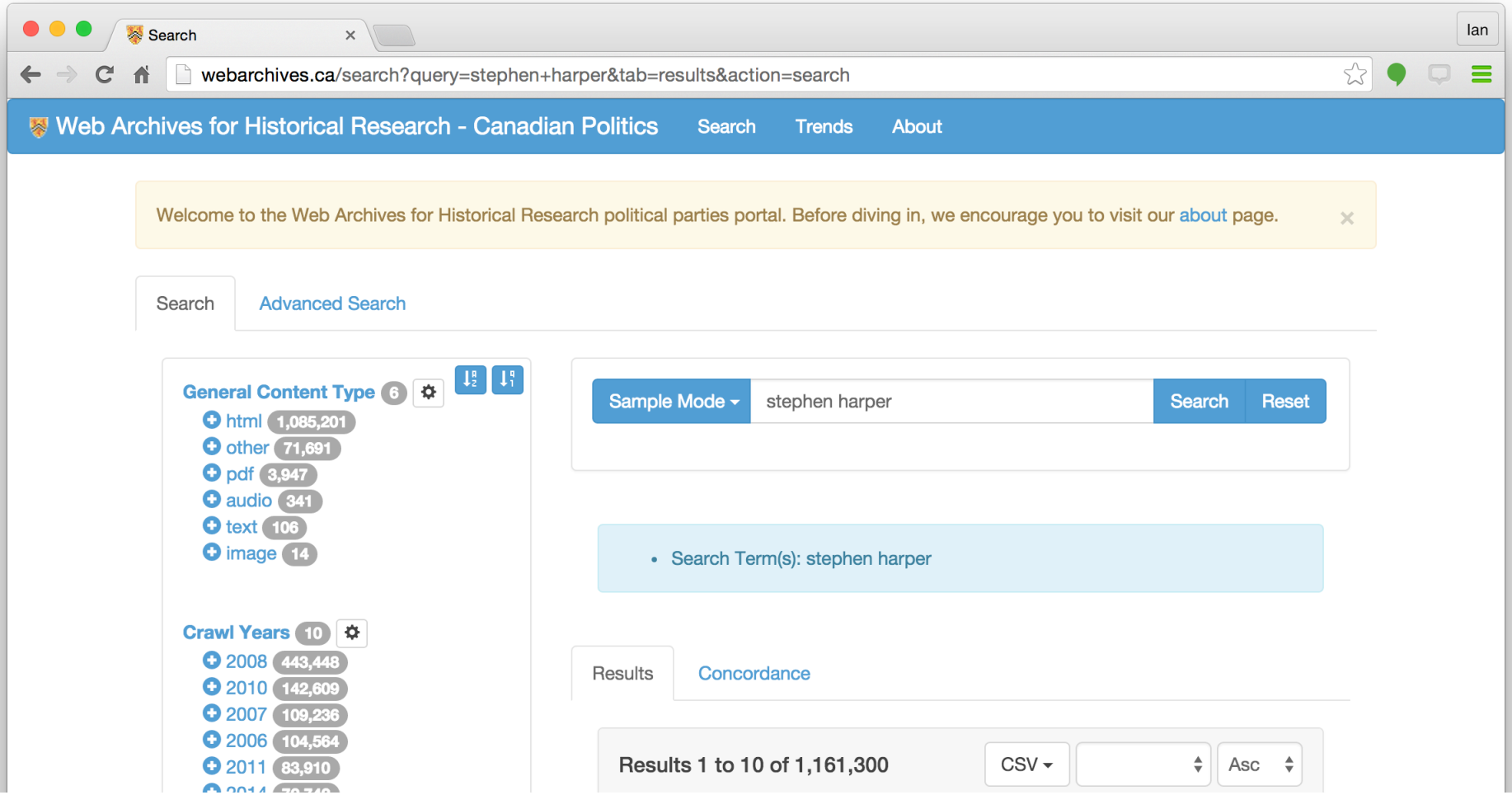
Feel free to take these examples and adapt them to your own purposes. If you find something interesting, please let us know!

Click on a point on the graph to show a random sample of the matching records from that year

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

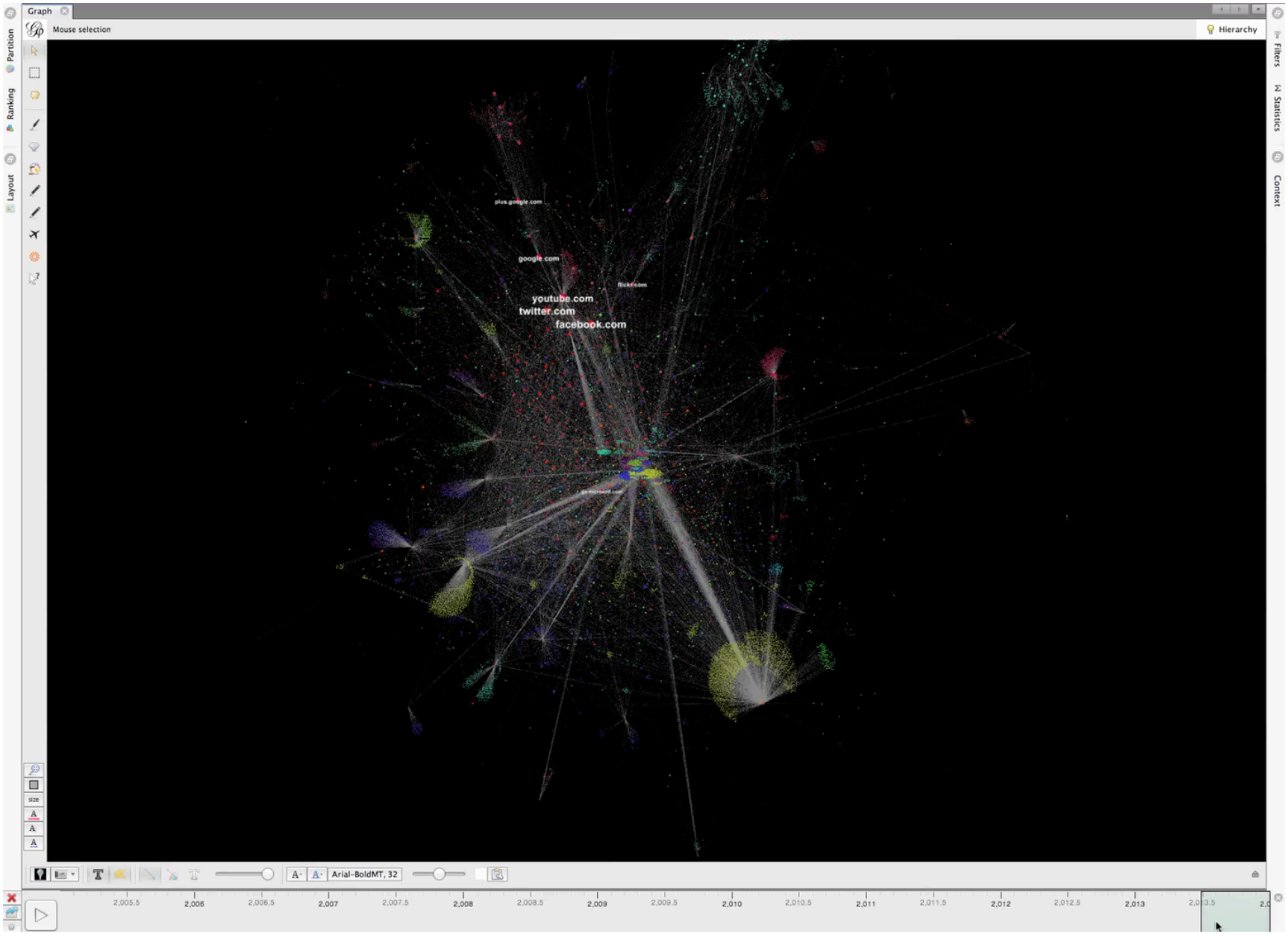
Good for public engagement - but limited for scholarship….



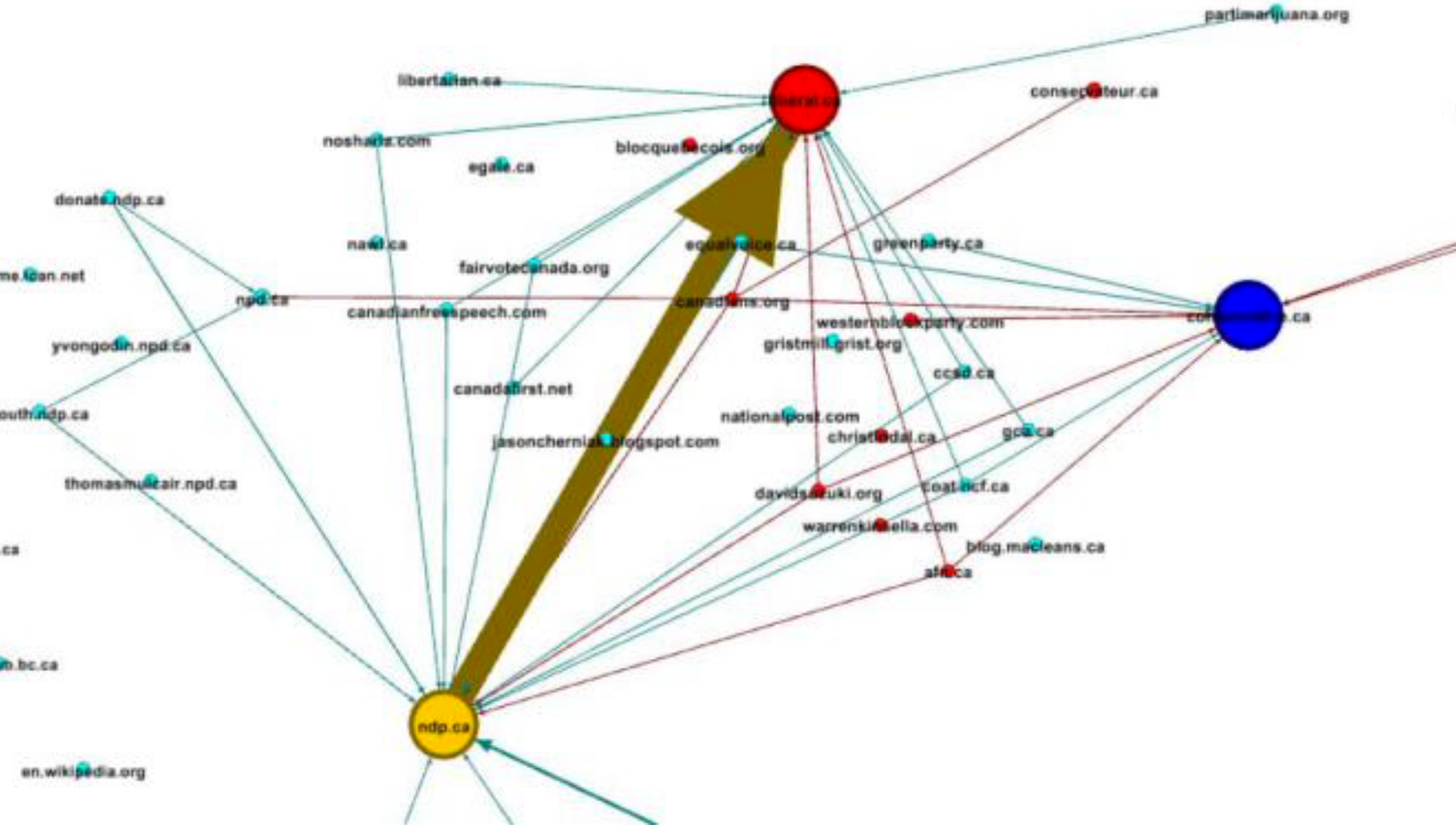
Do we want **metadata** or **content analysis**?

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

Metadata Extraction

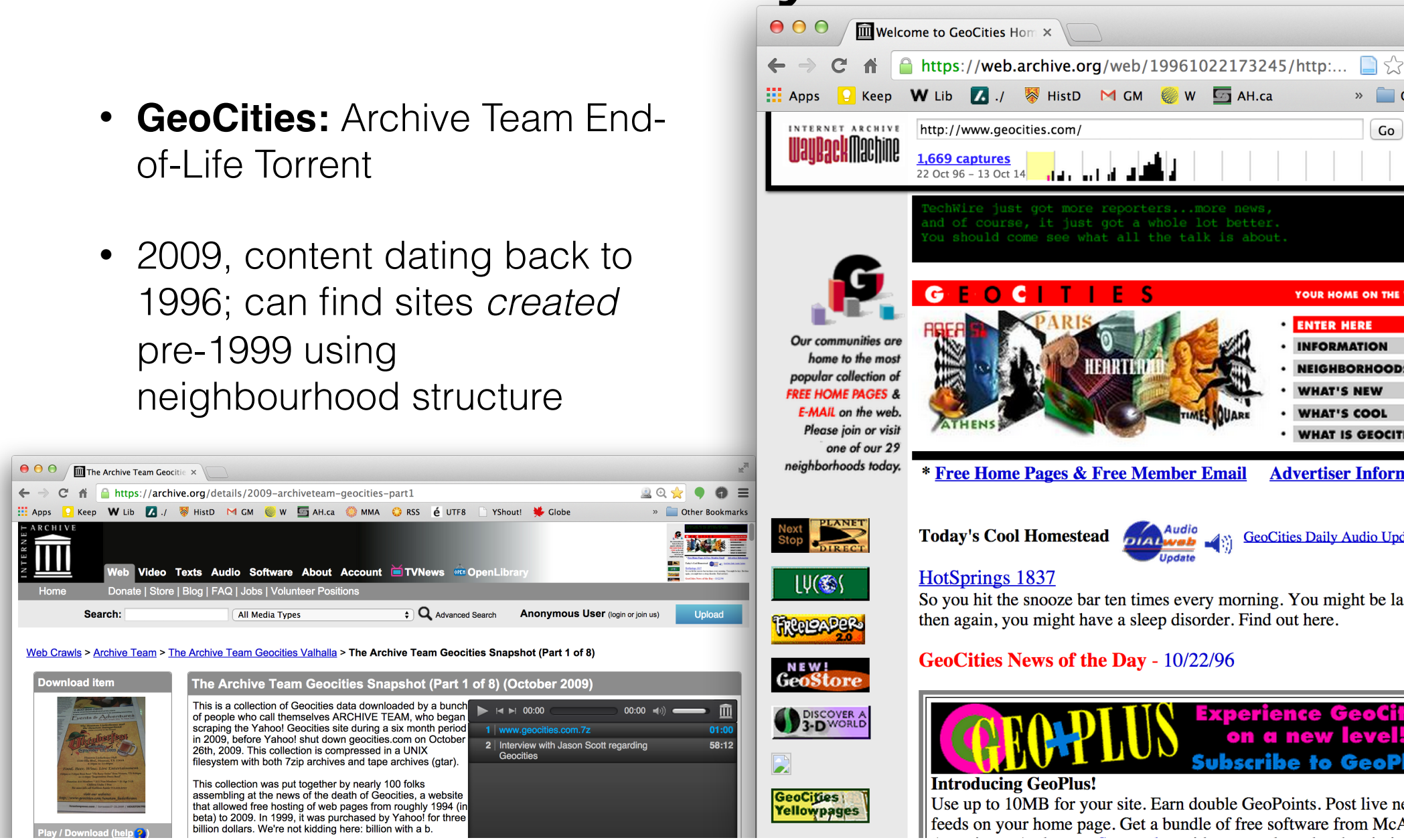


2005 Canadian Federal Election



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

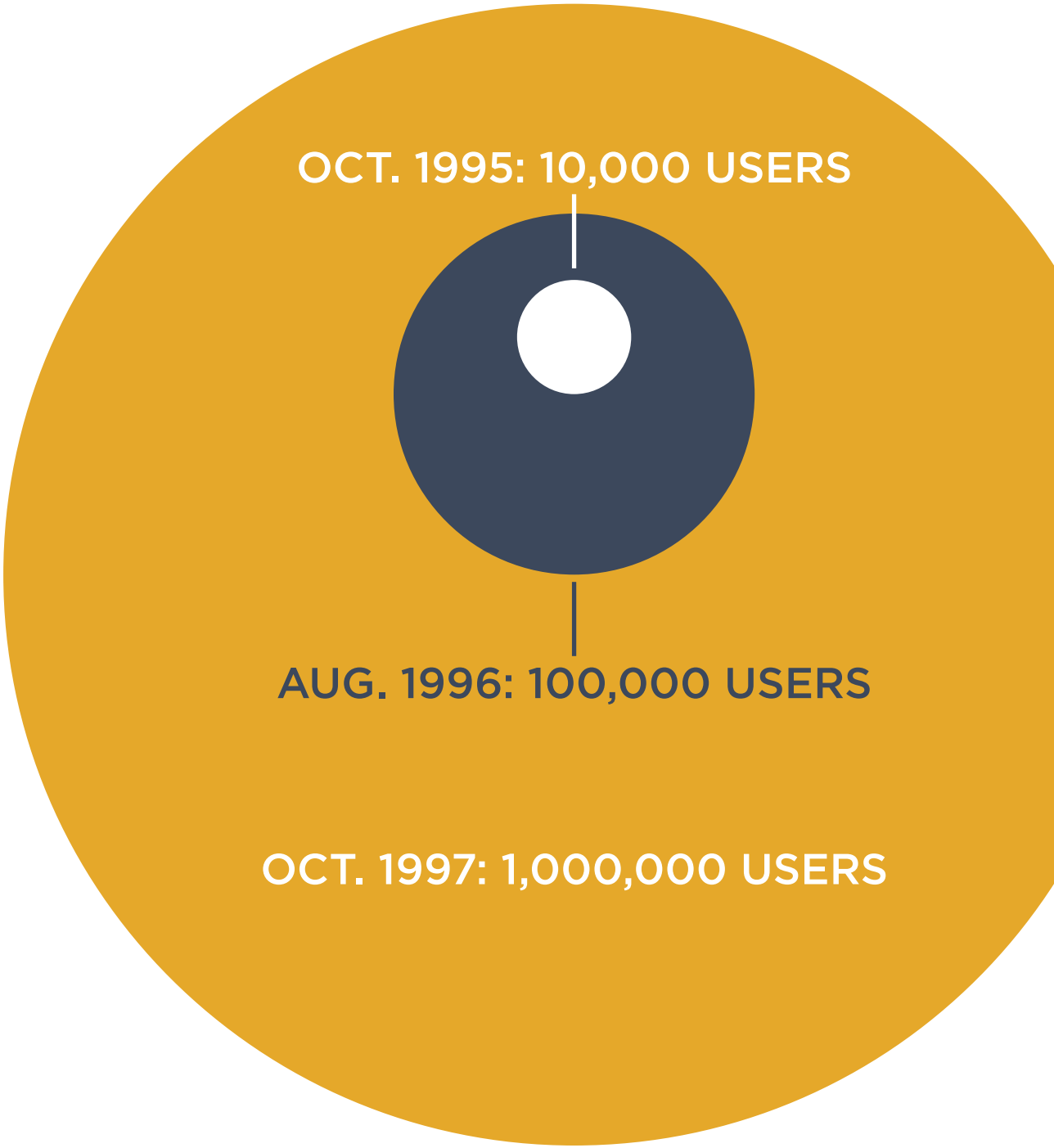


A substantive research question?

What was GeoCities?

Why does it matter?

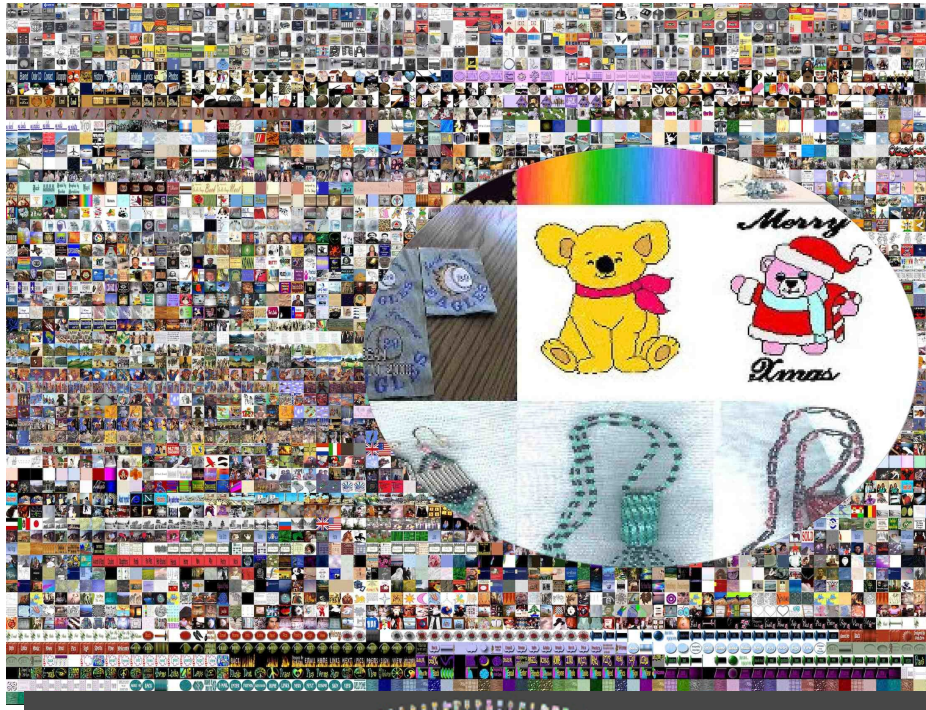
GEOCITIES USERS:

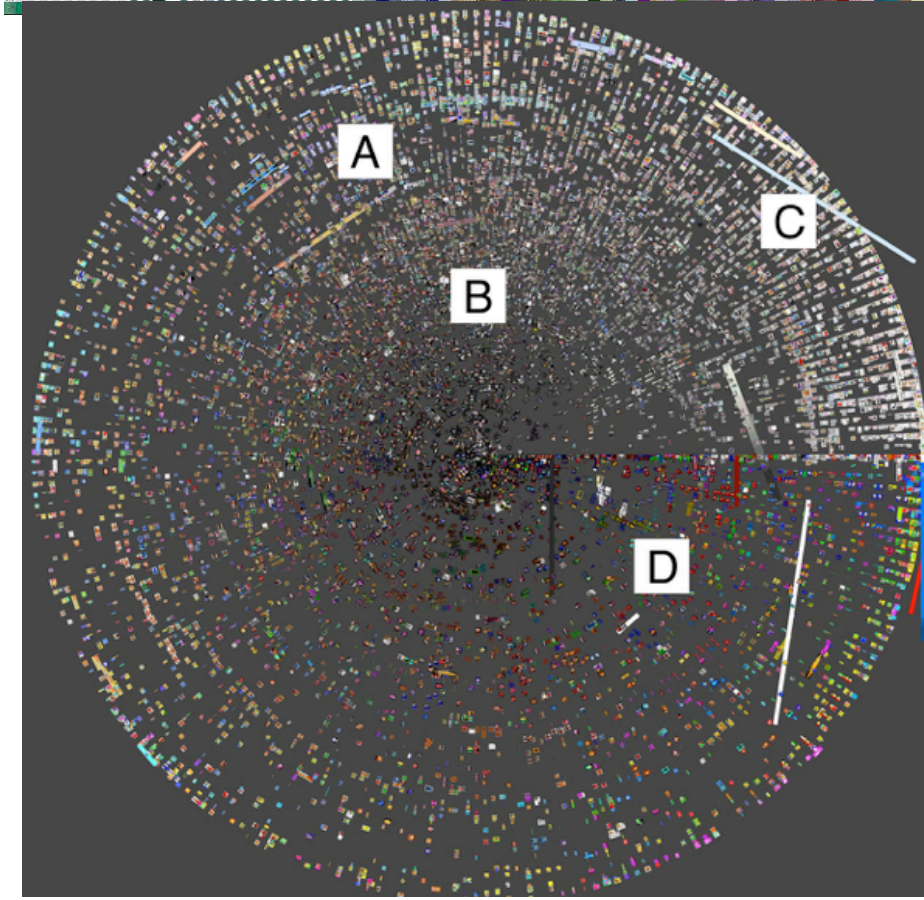


Topic Modelling Community to Test Coherence

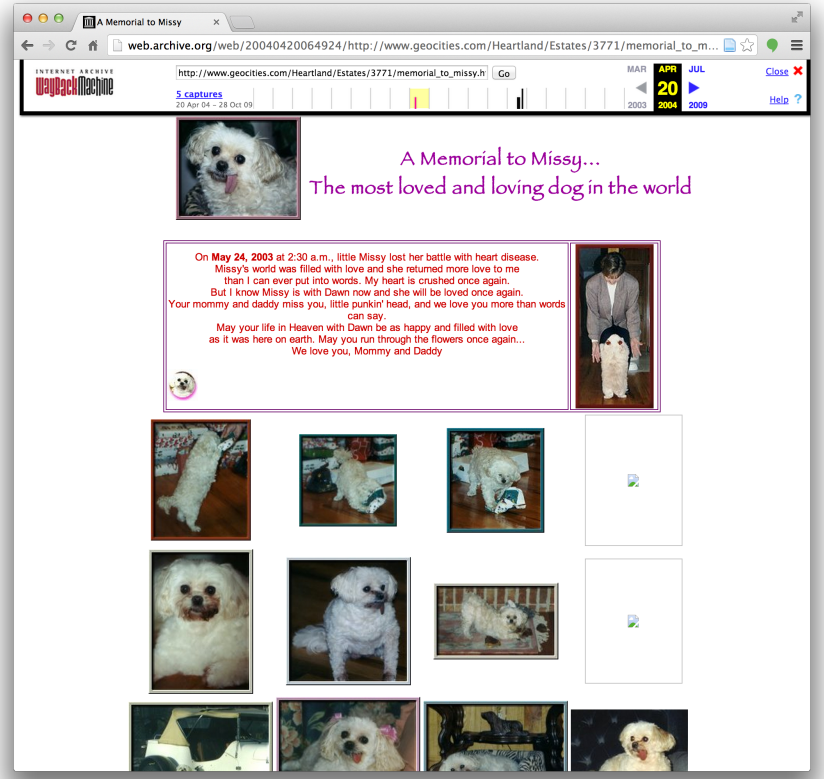
Selected Neighbourhoods	Top Two Topics
Athens *"… based on education, teaching, reading, writing and philosophy".*	people things time person sense life man work world h soul make nature body case made point part parts goddess witch healing incense witchcraft lov shaman witches sun spirit protection light circle earth
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 jq battalion show st jonny horse battery armored lt artillery camp sailor army field col pingu wa
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 yea 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 *"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 g frasier niles martin daphne roz don back ll door room scene ve dad turns takes crane good
Pentagon *Military men and women.*	war people president government american world state united general military public soviet political clinton an fort war civil island iran world adams army british hist 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 femin transgendered marriage man children transsexual

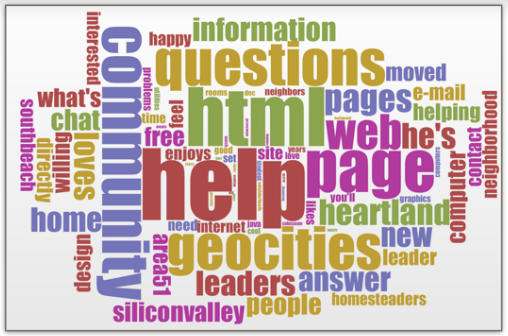
Looking at millions of user-contributed & generated images





And the stories of significant users and meaningful experiences.

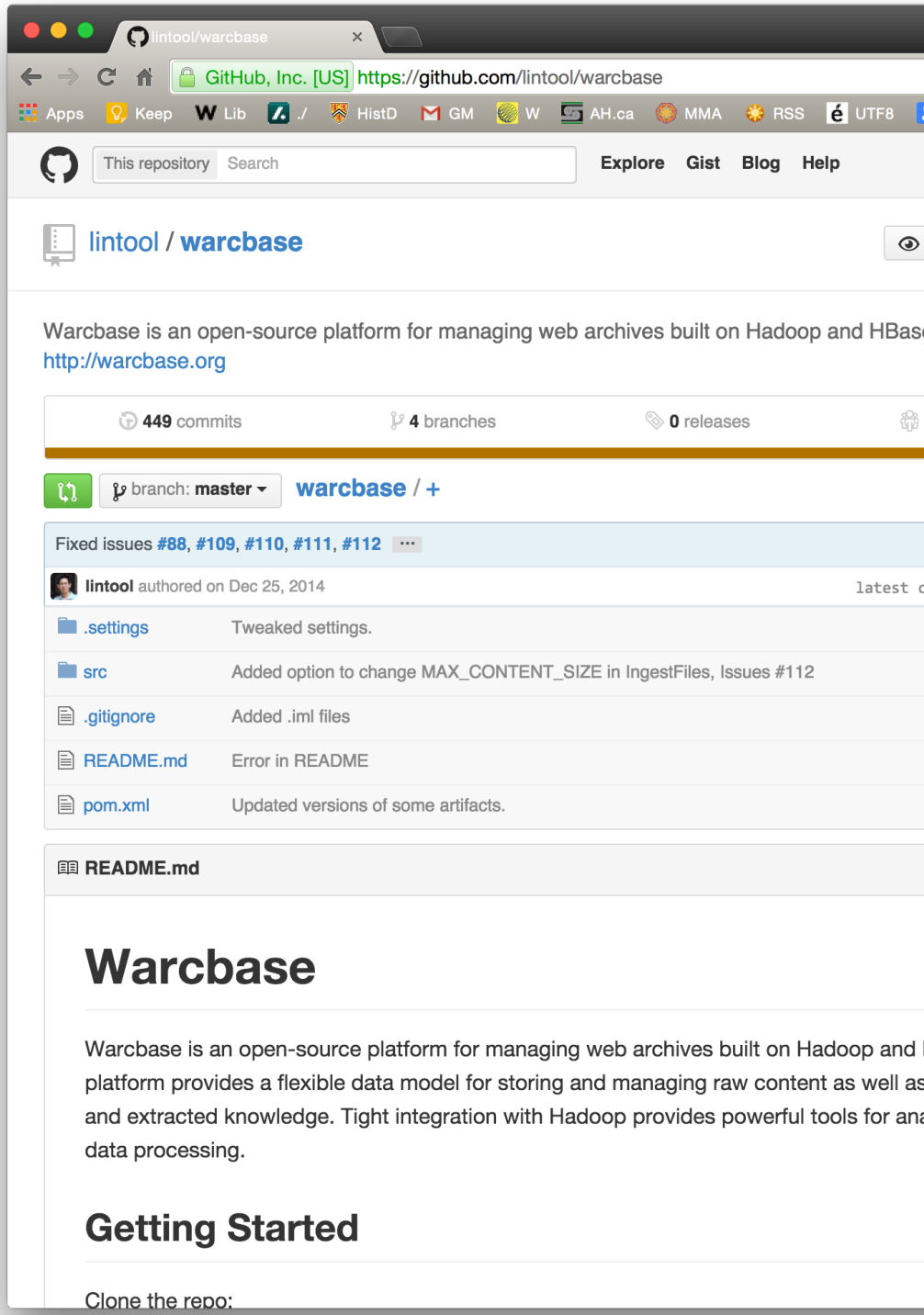




Shared Problems

- 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?

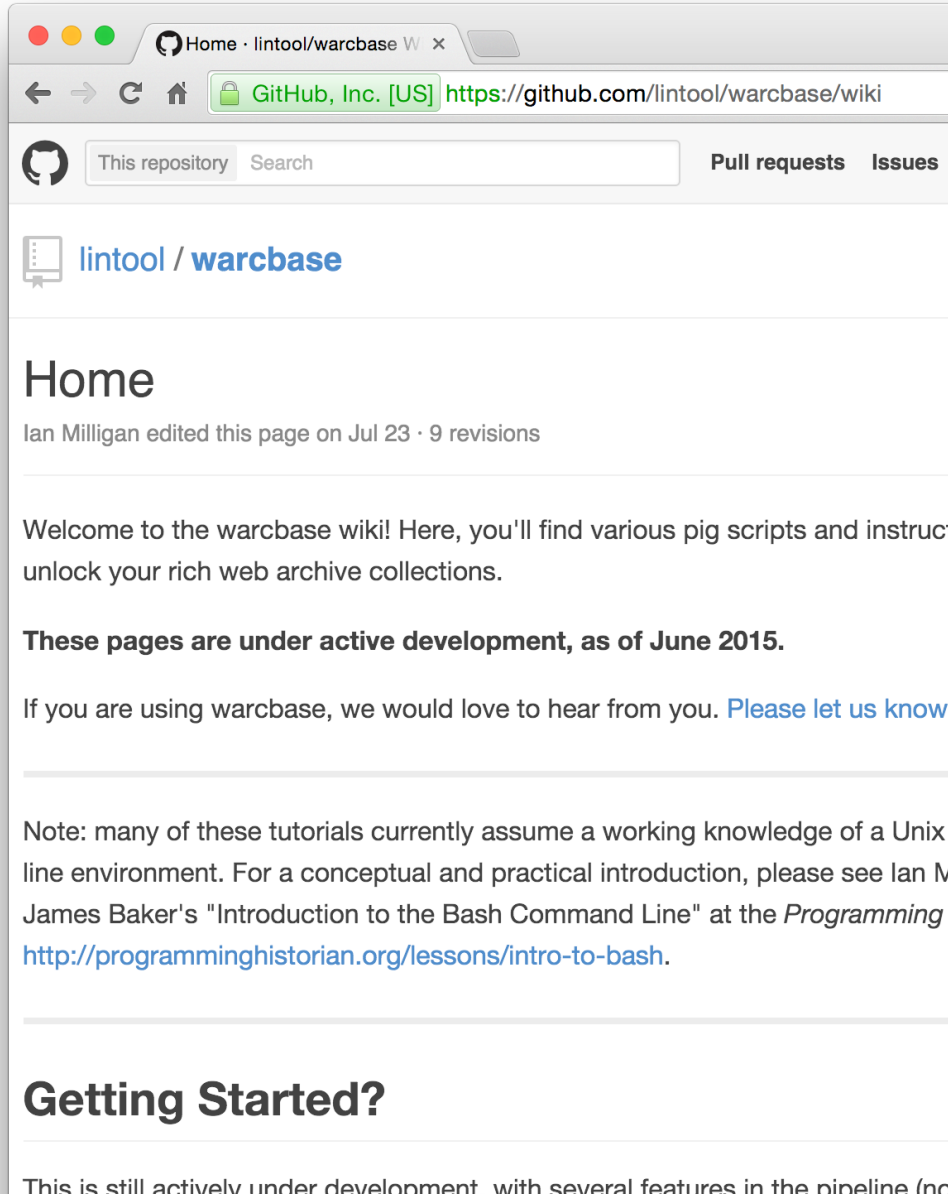
End-user tools and co-operation with CS, librarian, archivists colleagues is key.



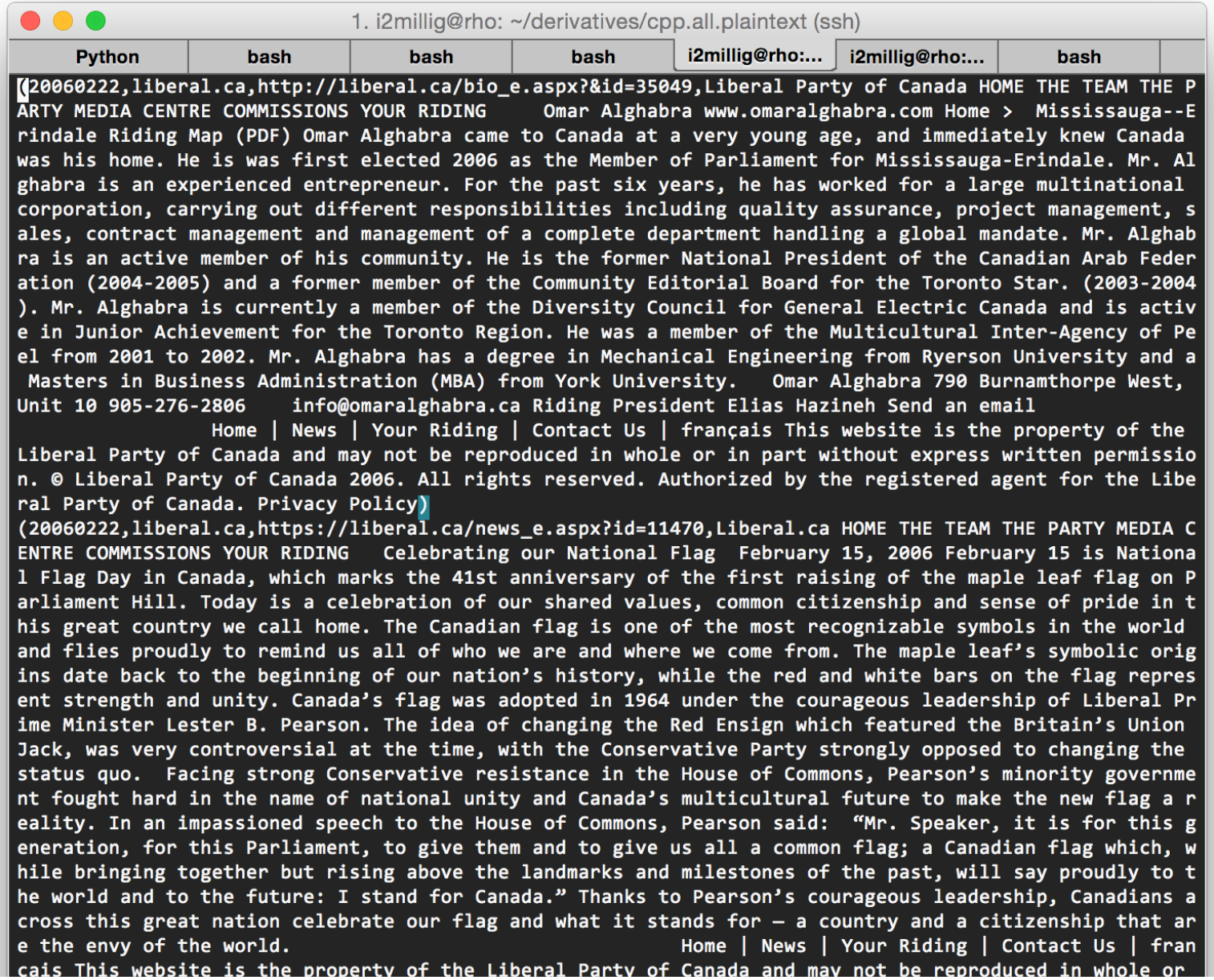
A platform for all kinds of questions?

Warcbase

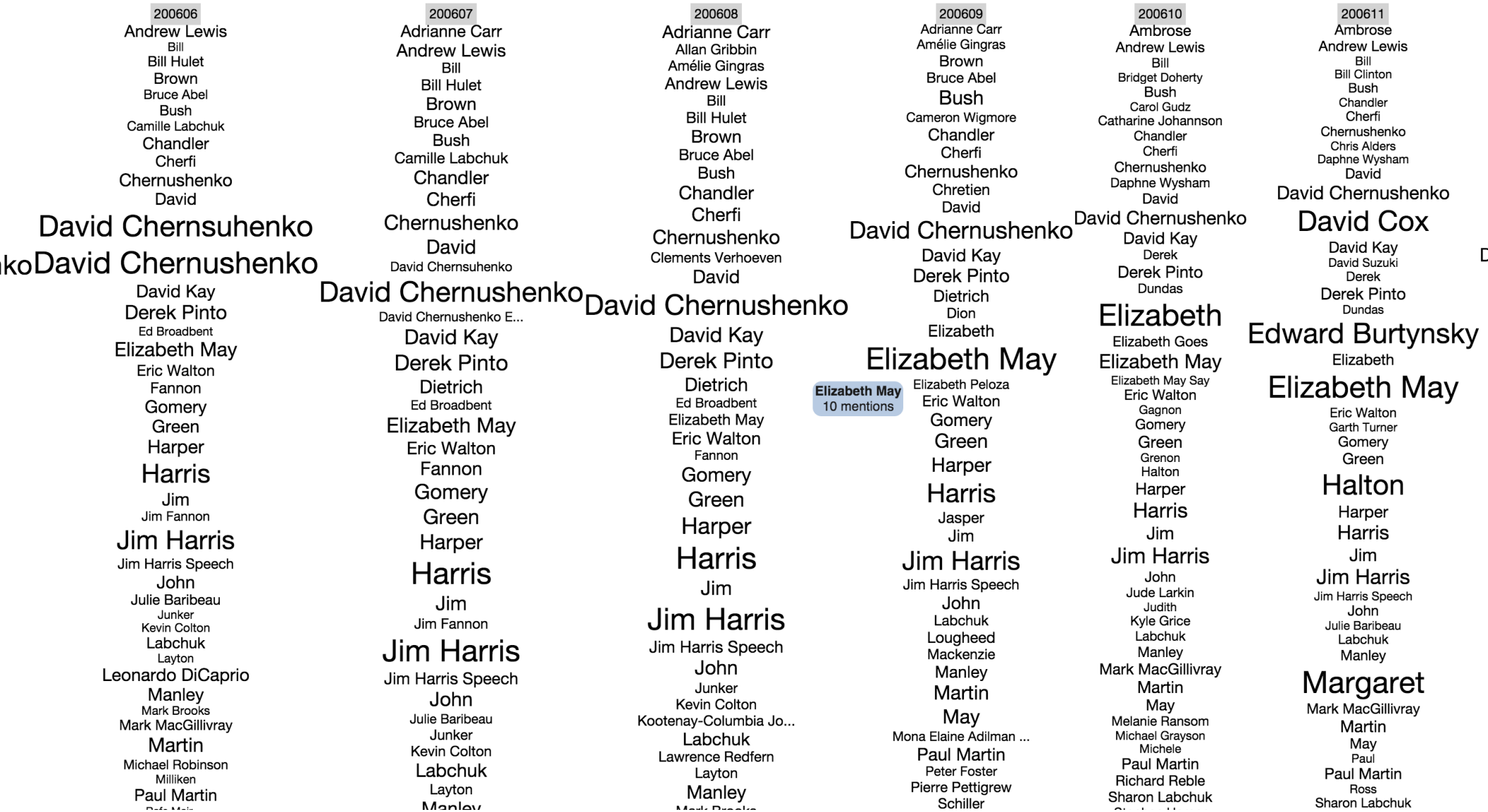
- Jimmy Lin (main developer), Ian Milligan, Jeremy Wiebe, Alice Zhou, all University of Waterloo
- Developing code, walkthroughs, and instructions for historians to be able to take a directory of WARCs and…



Extract all Plain Text

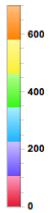


Extract Entities



Extract Entities

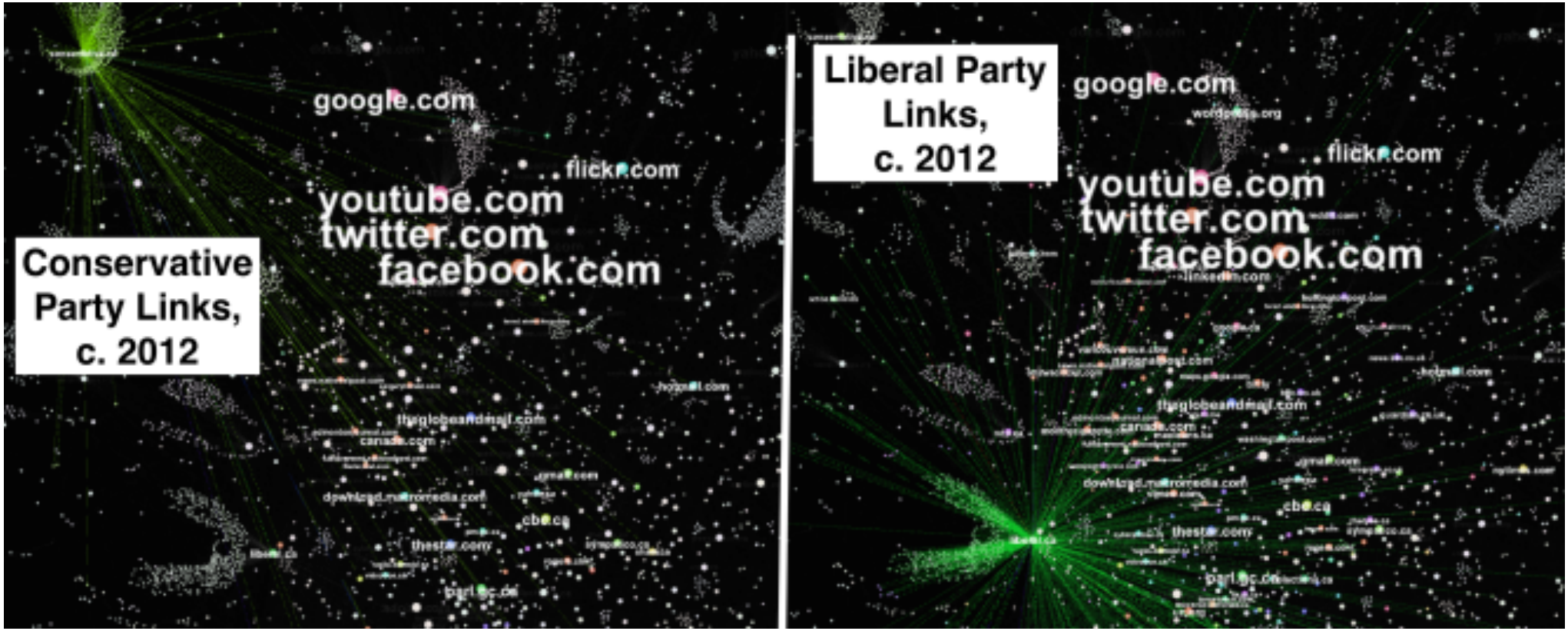
Extract Entities



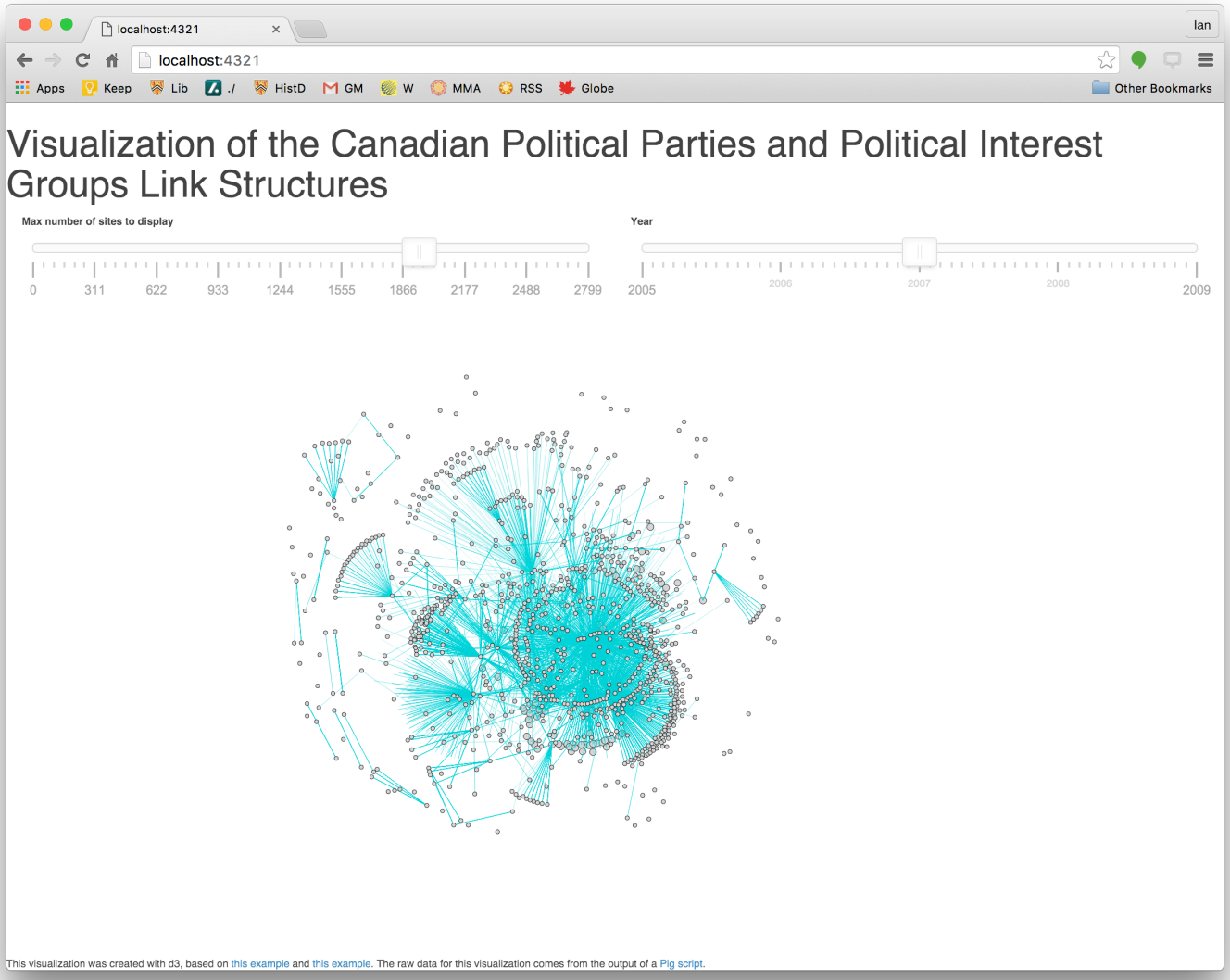

```
In[95]:= int = SemanticInterpretation[#] & /@ processedfreq[[All, 1]]
```

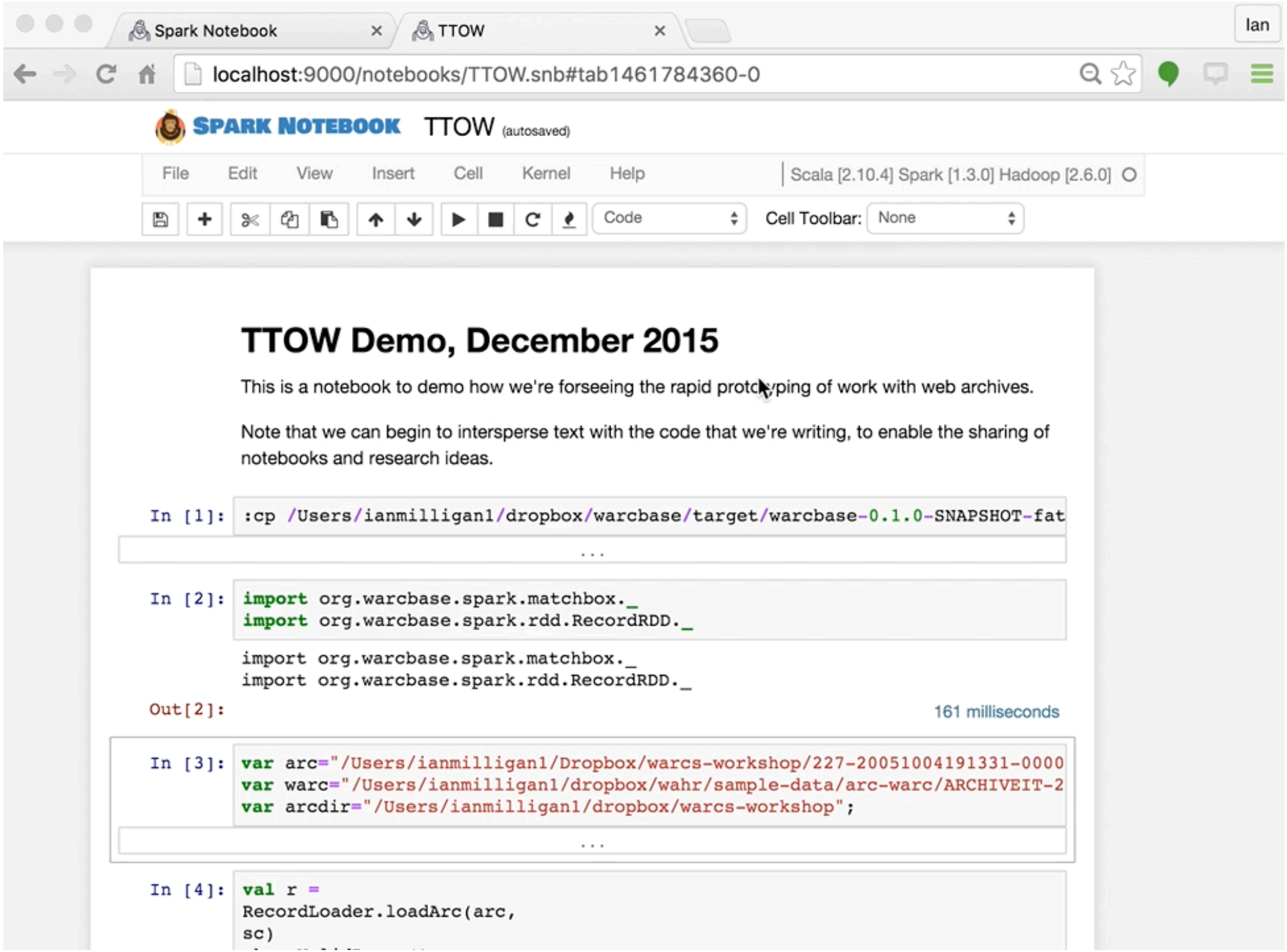
Out[95]= { Canada , Calgary , Colombia , Montreal , $Failed , Ontario, Canada , Afghanistan ,
Ottawa , Manitoba, Canada , British Columbia, Canada , Toronto , Nova Scotia, Canada ,
Saskatchewan, Canada , Quebec, Canada , Alberta, Canada , Winnipeg , Washington ,
Canada , Nunavut, Canada , White Horse , 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
Ian
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]
Code
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

Let's figure it out together!

"Archives Unleashed" Hackathon

March 3 - 5 2016, University of Toronto Library

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