Exploring the Past with Warcbase

Ian Milligan
Assistant Professor
@ianmilligan1

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Jimmy Lin
Professor and David R. Cheriton Chair
@lintool

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
David R. Cheriton School
of Computer Science

Jeremy Wiebe
PhD Candidate
@jeremyw

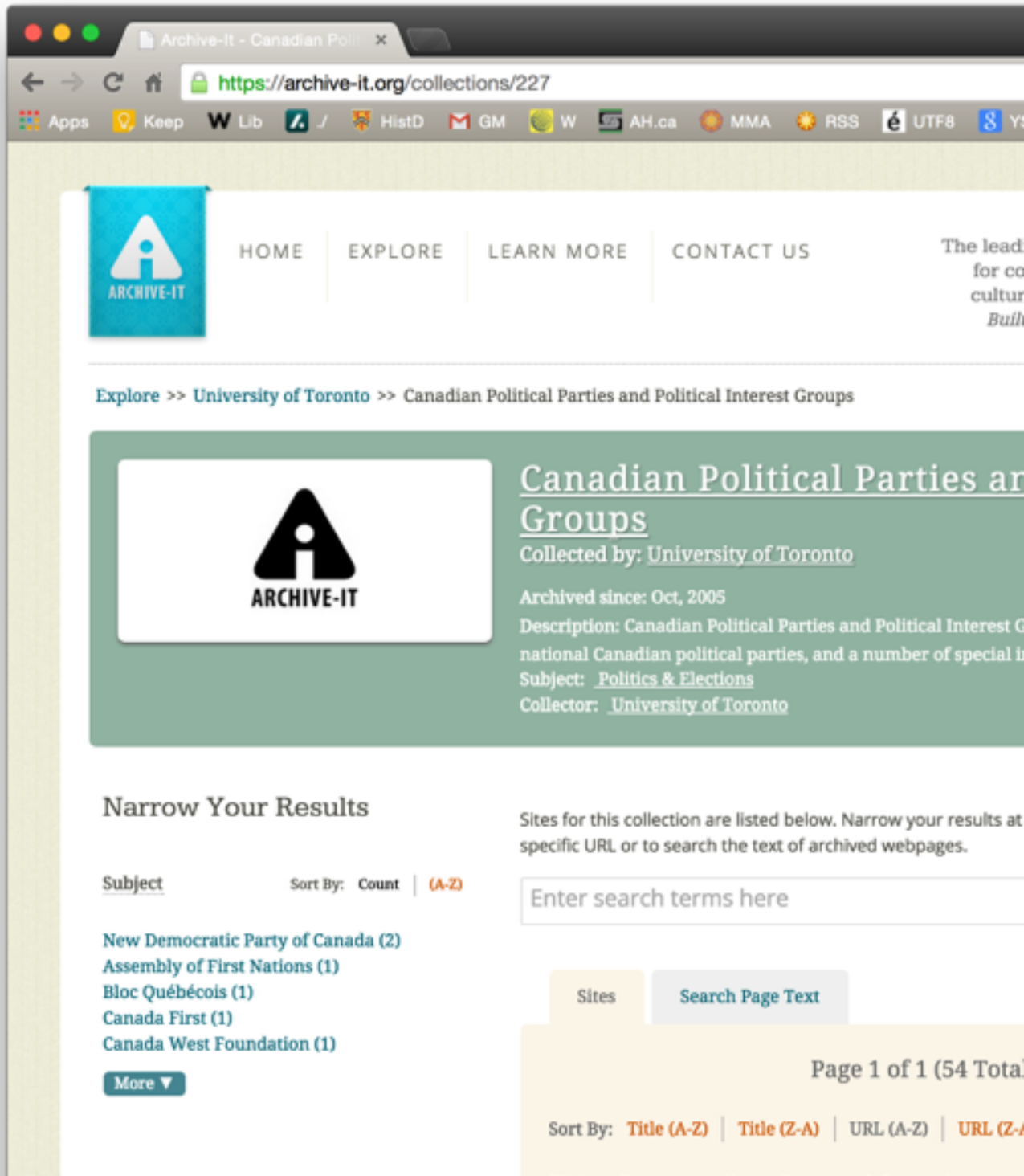
UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Two Case Studies

- **Archive-It Research Services**: “Canadian Political Parties and Political Interest Groups”

- 2005 - 2015

- WARC files

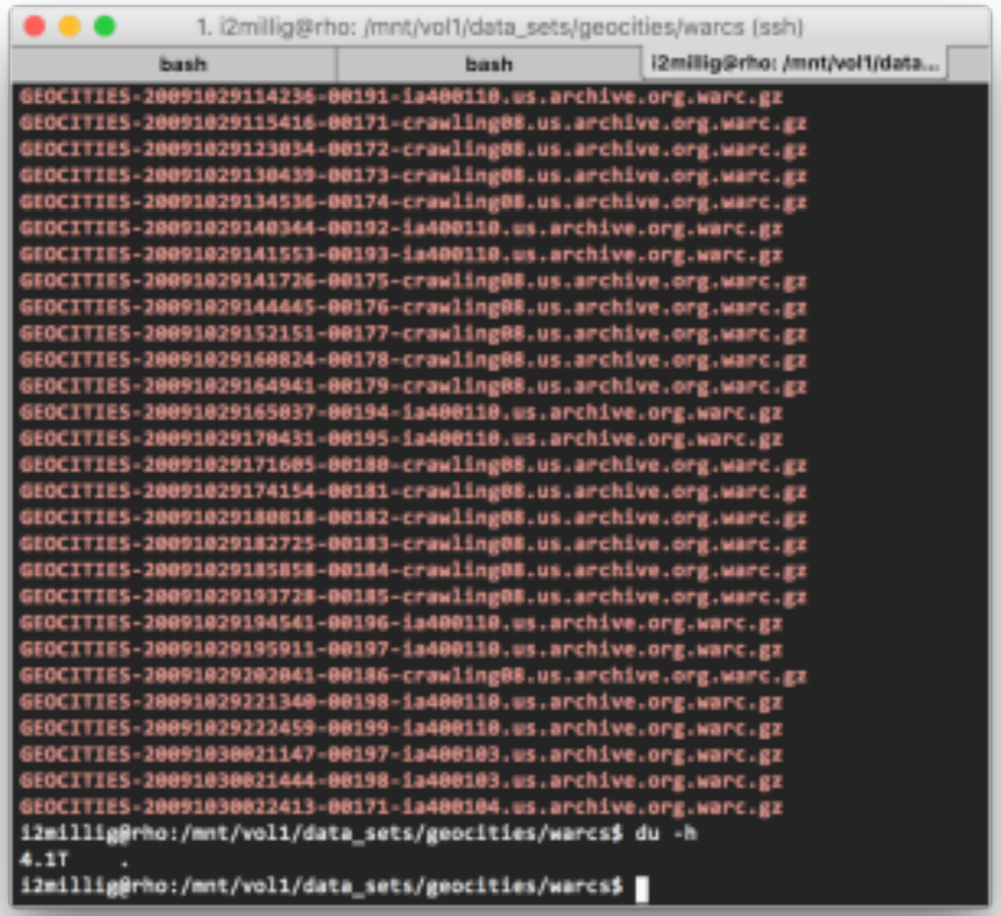


Two Case Studies

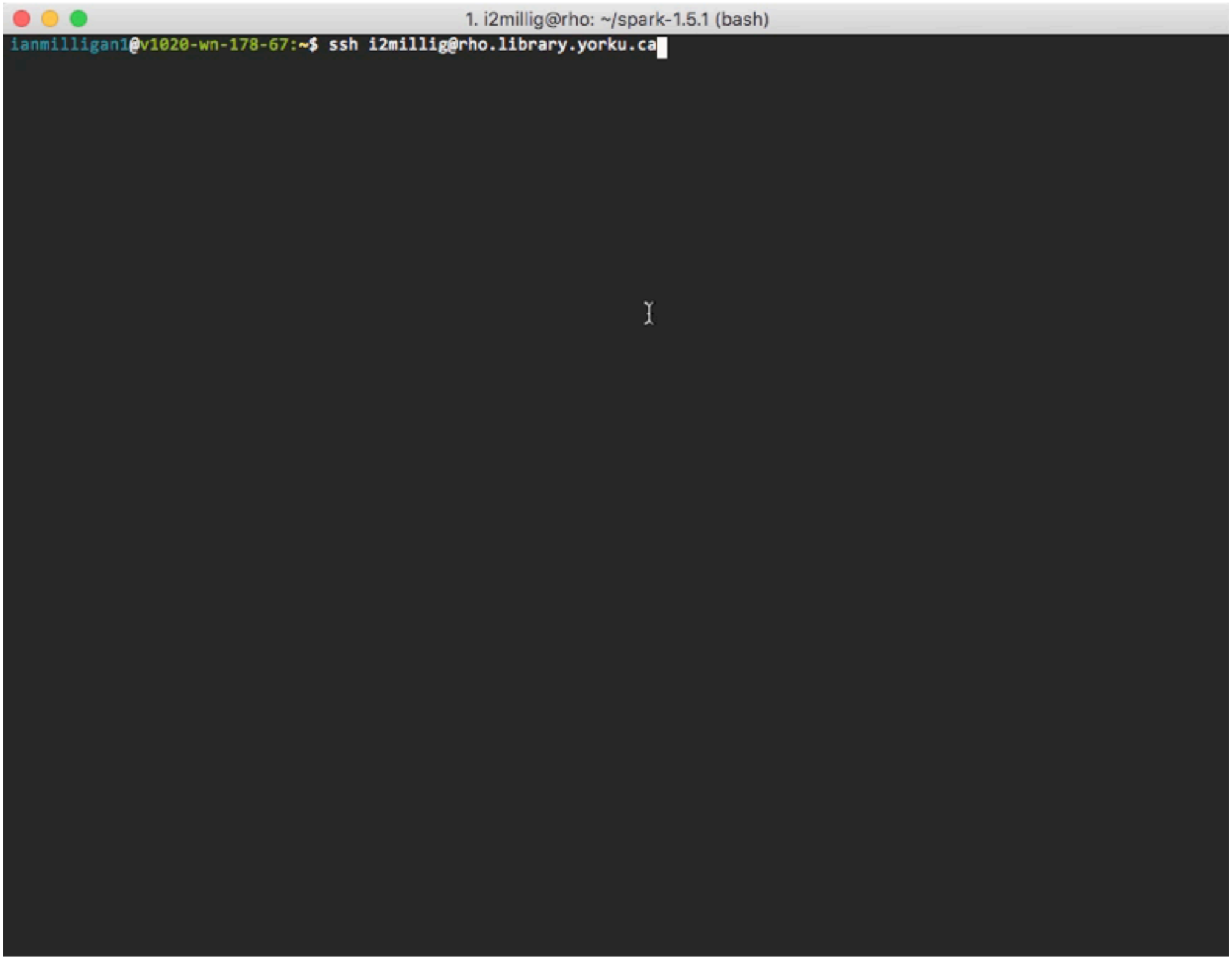
- **GeoCities**
- End-of-life crawl from 2009
- WARC files
- 4.1 TB, 186 million HTML documents

Using Warcbase to Learn Cool Stuff about it!

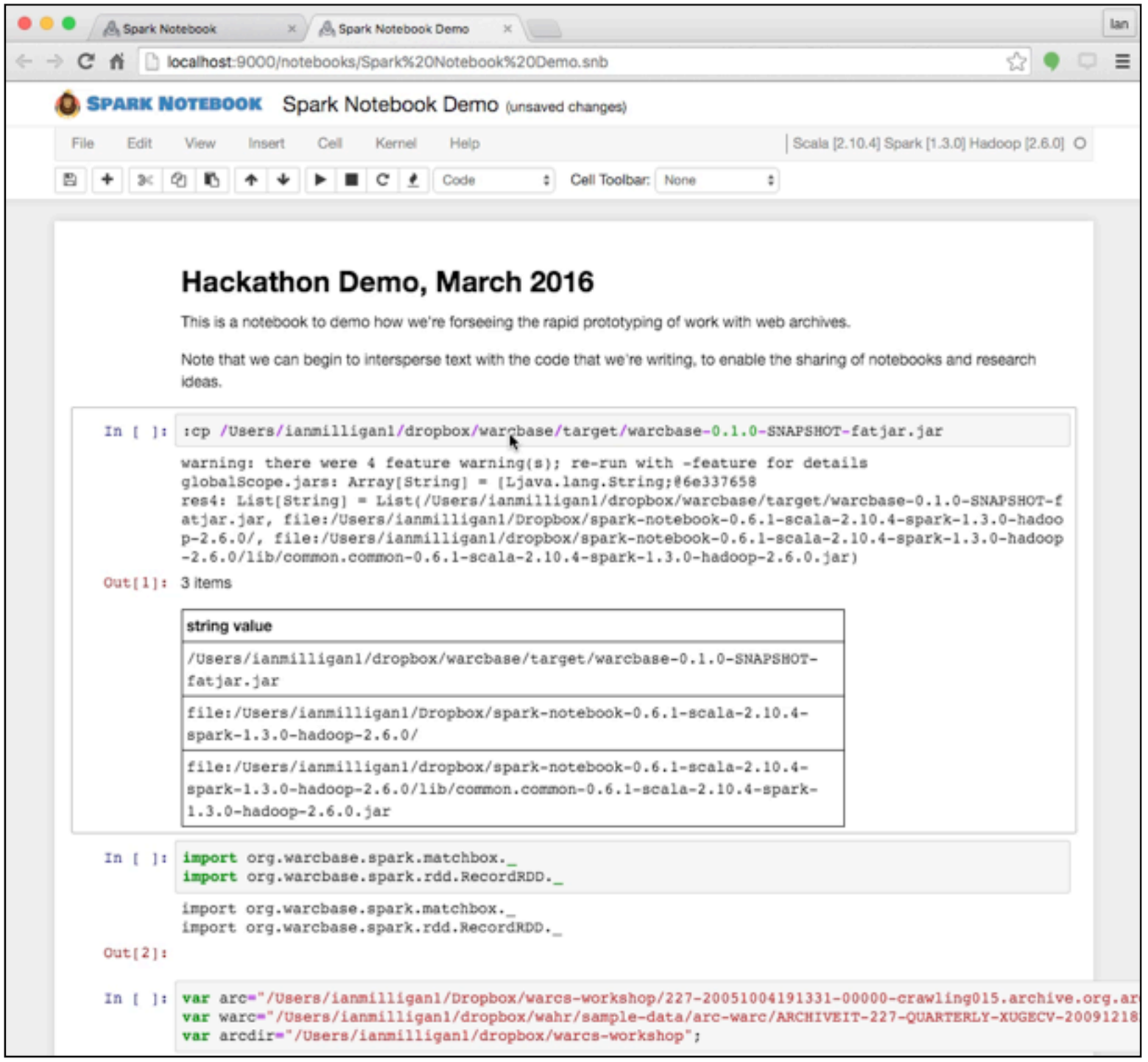
Step One: Grabbing WARCs



Step Two: Basic Shell Analysis



Step Two: Basic Analytics



Step Three: Filtering a Corpus

```
1  import org.warcbase.spark.matchbox.{ExtractTopLevelDomain,
       ExtractLinks, RecordLoader}
2  import org.warcbase.spark.rdd.RecordRDD._
3
4  RecordLoader.loadArc("/mnt/vol1/data_sets/geocities/warcs/*", sc)
5  .keepValidPages()
6  .map(r => (r.getCrawldate, ExtractLinks(r.getUrl, r.
       getContentString)))
7  .flatMap(r => r._2.map(f => (r._1, ExtractTopLevelDomain(f._1).
       replaceAll("^\\s*www\\.", ""), ExtractTopLevelDomain(f._2).
       replaceAll("^\\s*www\\.", ""))))
8  .filter(r => r._2 != "" && r._3 != "")
9  .countItems()
10 .filter(r => r._2 > 5)
11 .saveAsTextFile("/mnt/vol1/data_sets/geocities/geocities.
       sitelinks")
```

A Link Graph

Step Three: Filtering a Corpus

```
1  ((20090903,http://geocities.com/saganaki2000/ADSLGR/adslgr.htm,
   http://www.adslgr.com),15337)
2  ((20091026,http://geocities.com/saganaki2000/ADSLGR/adslgr.htm,
   http://www.adslgr.com),15337)
3  ((20091027,http://geocities.com/spankbank69hard/,http://pg.photos
   .yahoo.com/ph/spankbank69hard/my_photos/),9807)
4  ((20090903,http://geocities.com/spankbank69hard/index.html,http:/
   /pg.photos.yahoo.com/ph/spankbank69hard/my_photos/),9807)
5  ((20091027,http://geocities.com/CollegePark/Locker/8187/,http://
   www.comercialuruapan.com),8056)
6  ((20090903,http://geocities.com/CollegePark/Locker/8187/,http://
   www.comercialuruapan.com),8056)
```

Results

Gephi 0.8.2 - enchantedforest-analysis.gephi

Overview | Data Laboratory | Preview

Graph

Dragging

Hierarchy

Partition | Ranking | Layout

Context | Statistics | Filters

http://www.geocities.com/EnchantedForest/Glade/4021

http://www.geocities.com/EnchantedForest/Glade/3891

http://www.geocities.com/EnchantedForest/1004

http://www.geocities.com/EnchantedForest/1278

http://www.geocities.com/RainForest/Vines/4892

http://www.geocities.com/EnchantedForest/7631

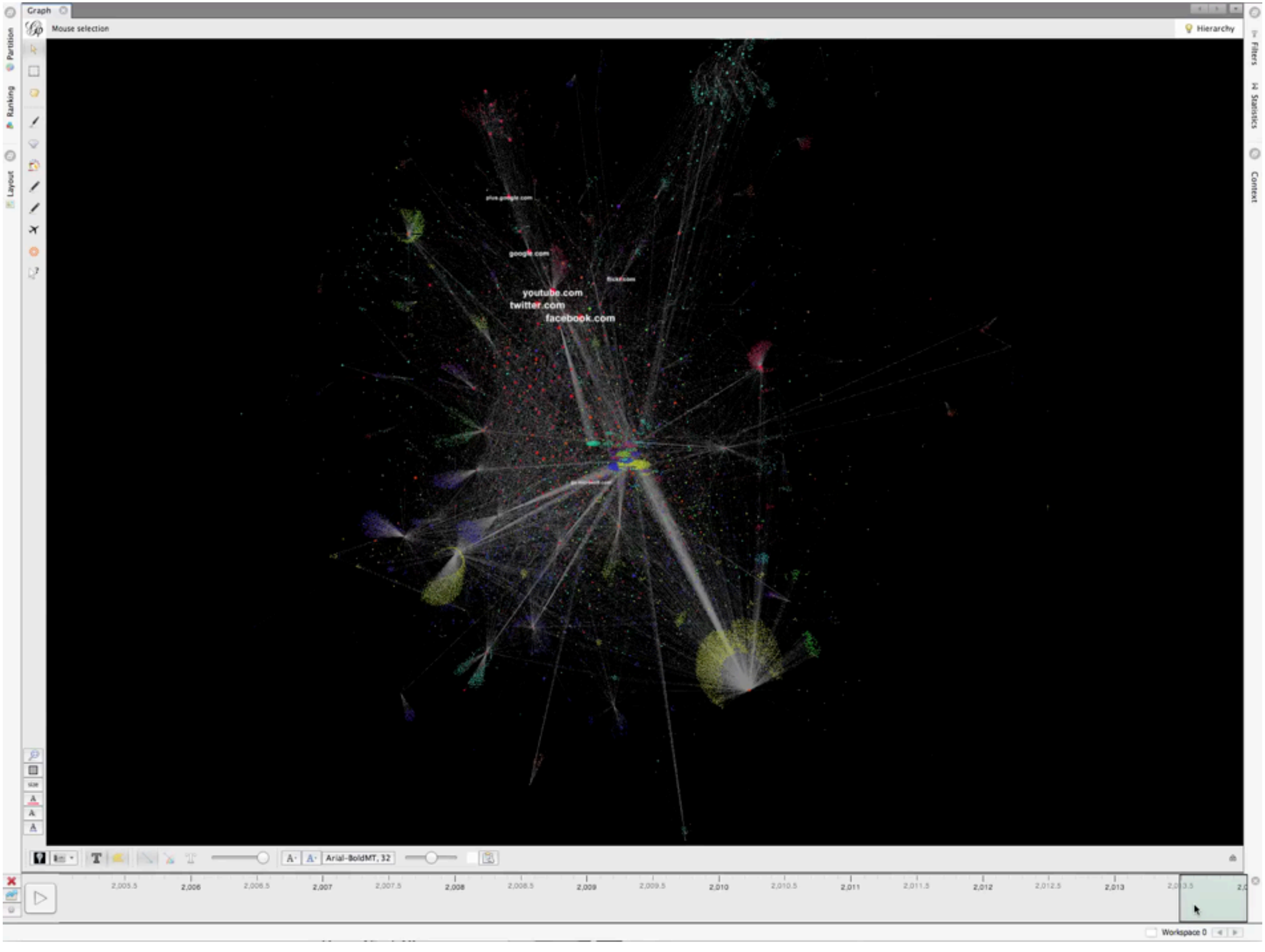
Size

Arial-BoldMT, 32

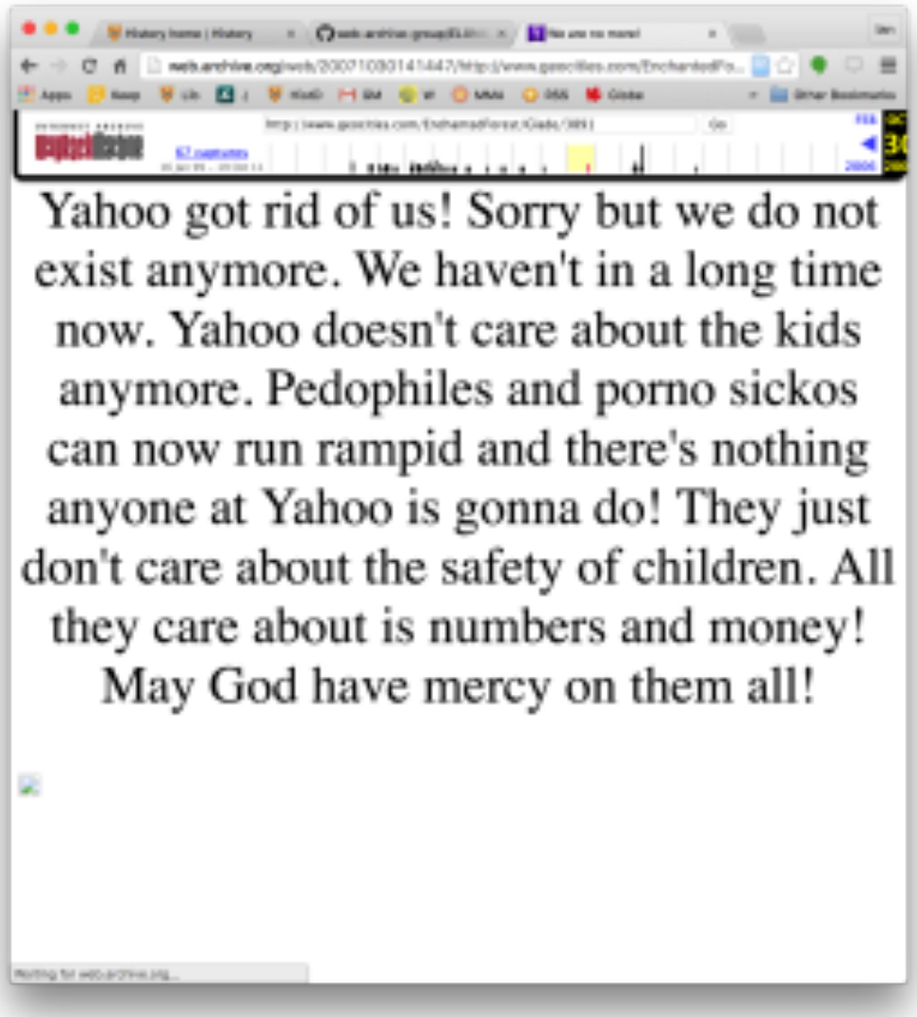
Workspace 0

Finding cool sites!

Label	PageRank	In-Degree	Out-Degree	Degree
http://www.geocities.com/EnchantedForest/Glade/3891	0.008	145	1	146
http://www.geocities.com/EnchantedForest/Glade/9378	0.005	6	13	19
http://www.geocities.com/EnchantedForest/1004	0.005	63	26	89
http://www.geocities.com/EnchantedForest/7631	0.004	6	3	9
http://www.geocities.com/SoHo/Cafe/4690	0.004	241	0	241
http://www.geocities.com/EnchantedForest/Glade/4021	0.004	151	0	151
http://www.geocities.com/TheTropics/Paradise/4079	0.003	248	0	248
http://www.geocities.com/RainForest/Vines/4892	0.003	5	6	11
http://www.geocities.com/EnchantedForest/3278	0.003	106	0	106
http://www.geocities.com/EnchantedForest/3696	0.003	70	0	70
http://www.geocities.com/EnchantedForest/Dell/5914	0.003	180	1	181
http://www.geocities.com/EnchantedForest/1469	0.003	16	49	65
http://www.geocities.com/EnchantedForest/Tower/9644	0.003	19	42	61
http://www.geocities.com/EnchantedForest/Dell/9501	0.003	79	362	441
http://www.geocities.com/EnchantedForest/Glade/8851	0.003	17	0	17
http://www.geocities.com/Heartland/Meadows/6263	0.003	9	0	9
http://www.geocities.com/Heartland/6188	0.003	56	0	56
http://www.geocities.com/EnchantedForest/4213	0.003	158	0	158
http://www.geocities.com/Athens/Acropolis/1465	0.003	20	0	20
http://www.geocities.com/EnchantedForest/8012	0.003	42	197	239
http://www.geocities.com/EnchantedForest/3810	0.003	98	147	245
http://www.geocities.com/EnchantedForest/Glade/3899	0.002	14	11	25
http://www.geocities.com/EnchantedForest/3015	0.002	64	0	64
http://www.geocities.com/EnchantedForest/Tower/8143	0.002	50	40	90
http://www.geocities.com/EnchantedForest/Meadow/1426	0.002	41	185	226


Graph
Mouse selection
Hierarchy
Partition
Ranking
Layout
Filters
Statistics
Context
plus.google.com
google.com
youtube.com
twitter.com
facebook.com
Arial-BoldMT, 32
2,005.5
2,006
2,006.5
2,007
2,007.5
2,008
2,008.5
2,009
2,009.5
2,010
2,010.5
2,011
2,011.5
2,012
2,012.5
2,013
Workspace 0

Step Four: Finding Significant Sites w/ PageRank



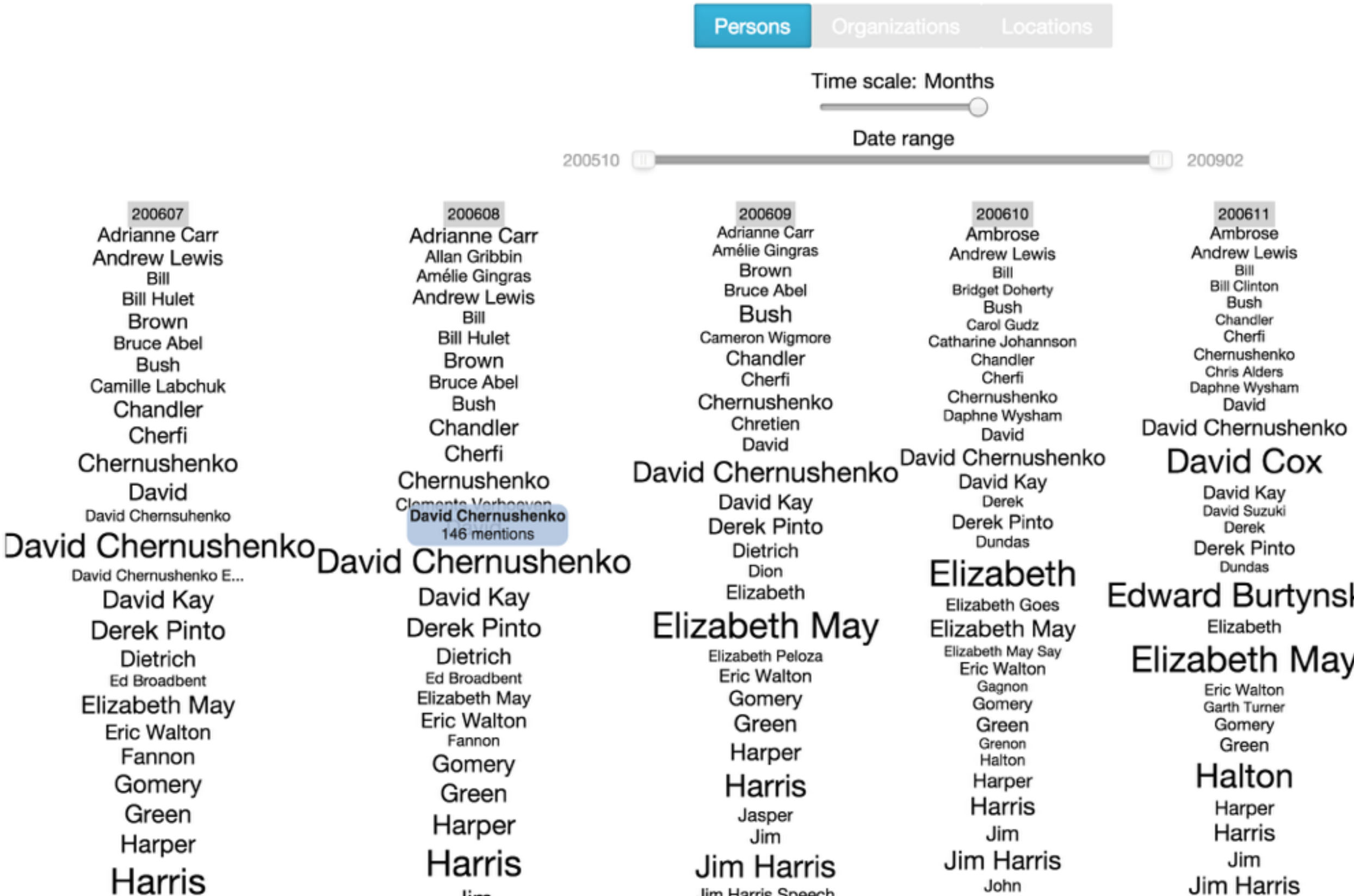
Step Five: Text Analysis

Different Ways to Filter

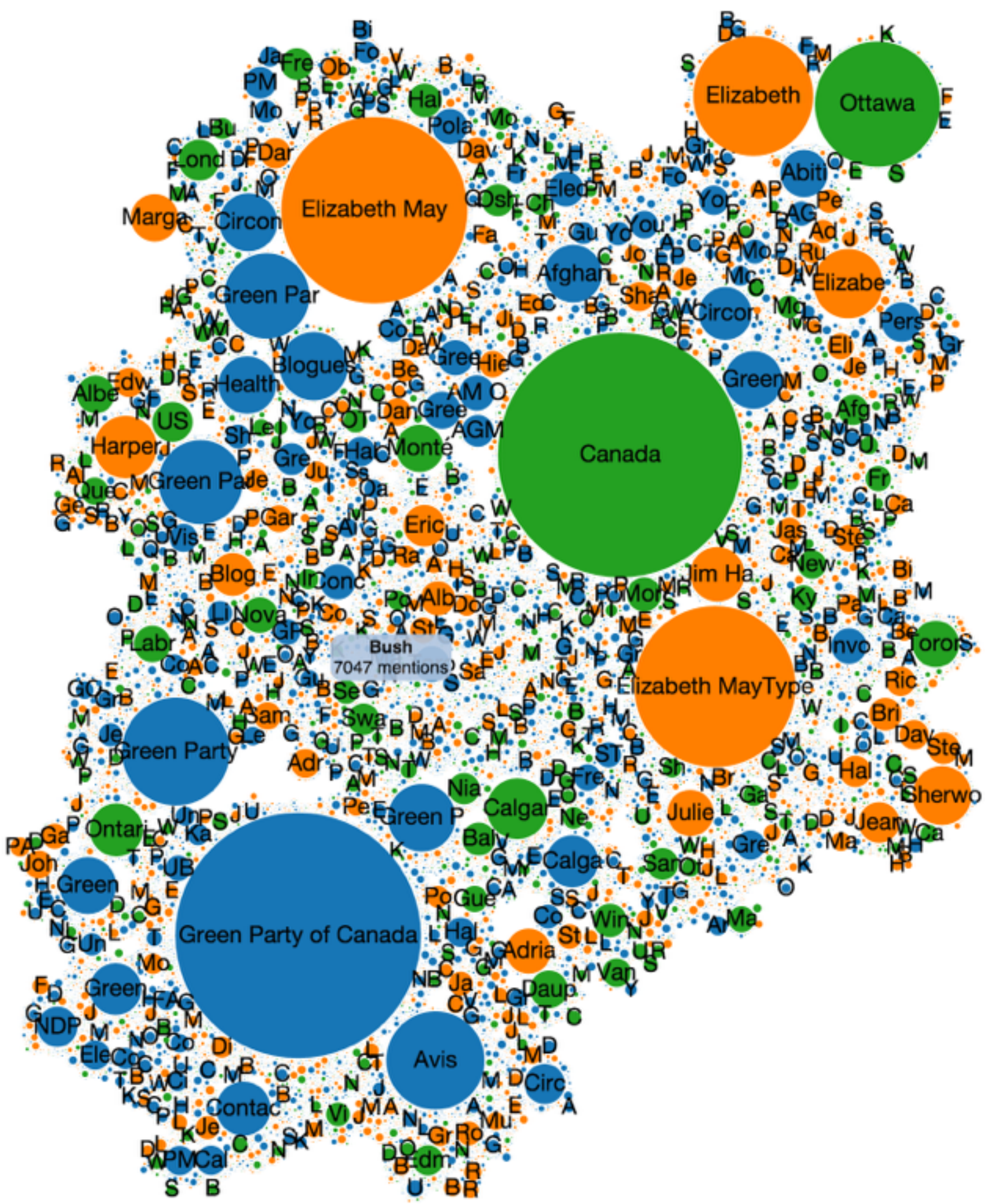
- Get everything
- Filter by domain (i.e. all pages in “greenparty.ca”)
- Filter by URL pattern (i.e. all pages in “greenparty.ca/vegetables/*”)
- Filter out boilerplate (i.e. advertisements, navigational elements, templates, etc.)
- Filter by date (i.e. all pages on July 4th, 2015)
- Filter by languages (i.e. only French language pages from greenparty.ca)
- Or any of the above!

Named Entity Visualization

Data source: greenparty.csv




Manley
22 mentions


Elizabeth
Ottawa
Elizabeth May
Green Par
Blogue
Health
US
Harper
Green Pa
Canada
Green Party
Green Party of Canada
Elizabeth MayType
Avis
Bush
7047 mentions

Or generate Solr indexes using Warcbase too!

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

The Canadian Political Parties and Political Interest Groups Portal

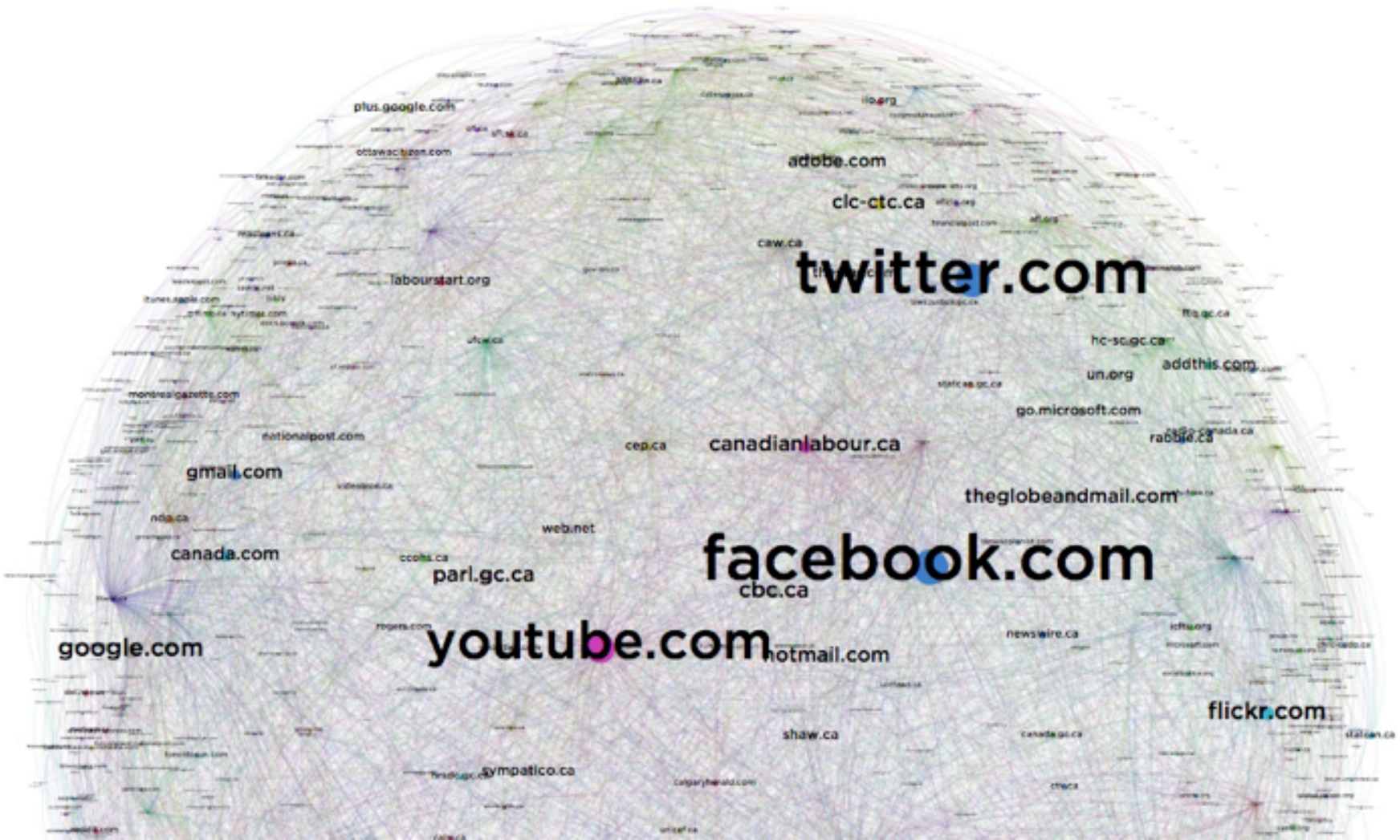
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

Curious how the Liberal Party of Canada responded to the 2008 financial crisis (a search for "recession" in 2008, liberal.ca)? How the Canadian Centre for Policy Alternatives reacted to Michael Ignatieff? Now you can check it all out.

Options include:

- **Basic keyword searching** [Example: "Rob Ford", only Liberal.ca]
- **Graphing trends over time** [Example: Liberal Opposition Leaders, 2005-2015]
- **Advanced search, including words in proximity to each other** [Example: environmental and tax within 25 words of each other]

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



Step Five: $$$$

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada
Ontario
compute • calcul CANADA
UNIVERSITY OF WATERLOO

Thanks very much!

Stay tuned for the technical explanation!