

# 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

The screenshot shows a web browser window with the URL <https://archive-it.org/collections/227>. The page title is "Archive-It - Canadian Political Parties and Political Interest Groups". The header includes links for HOME, EXPLORE, LEARN MORE, and CONTACT US. A sidebar on the right says "The leading provider for cultural heritage institutions". The main content area displays a large "ARCHIVE-IT" logo and the title "Canadian Political Parties and Political Interest Groups" collected by "University of Toronto". It notes the collection was archived since Oct, 2005, with a description of Canadian political parties and interest groups. Below this, there's a section titled "Narrow Your Results" with a search bar and a list of subjects: New Democratic Party of Canada (2), Assembly of First Nations (1), Bloc Québécois (1), Canada First (1), and Canada West Foundation (1). At the bottom, there are buttons for "Sites" and "Search Page Text", and a footer note "Page 1 of 1 (54 Total)".

# Two Case Studies



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

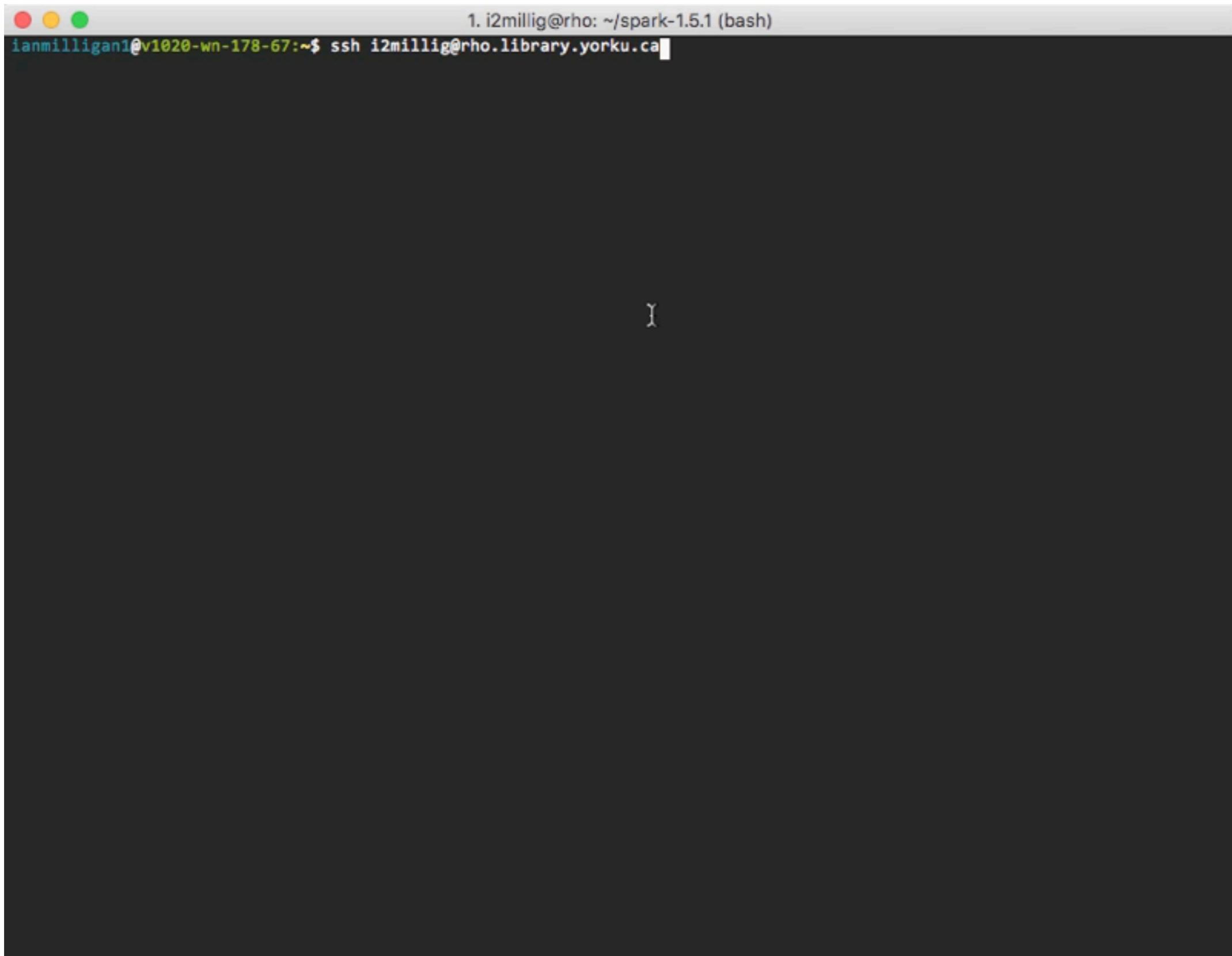
Using Warcbase to  
Learn Cool Stuff about it!

# Step One: Grabbing WARCs



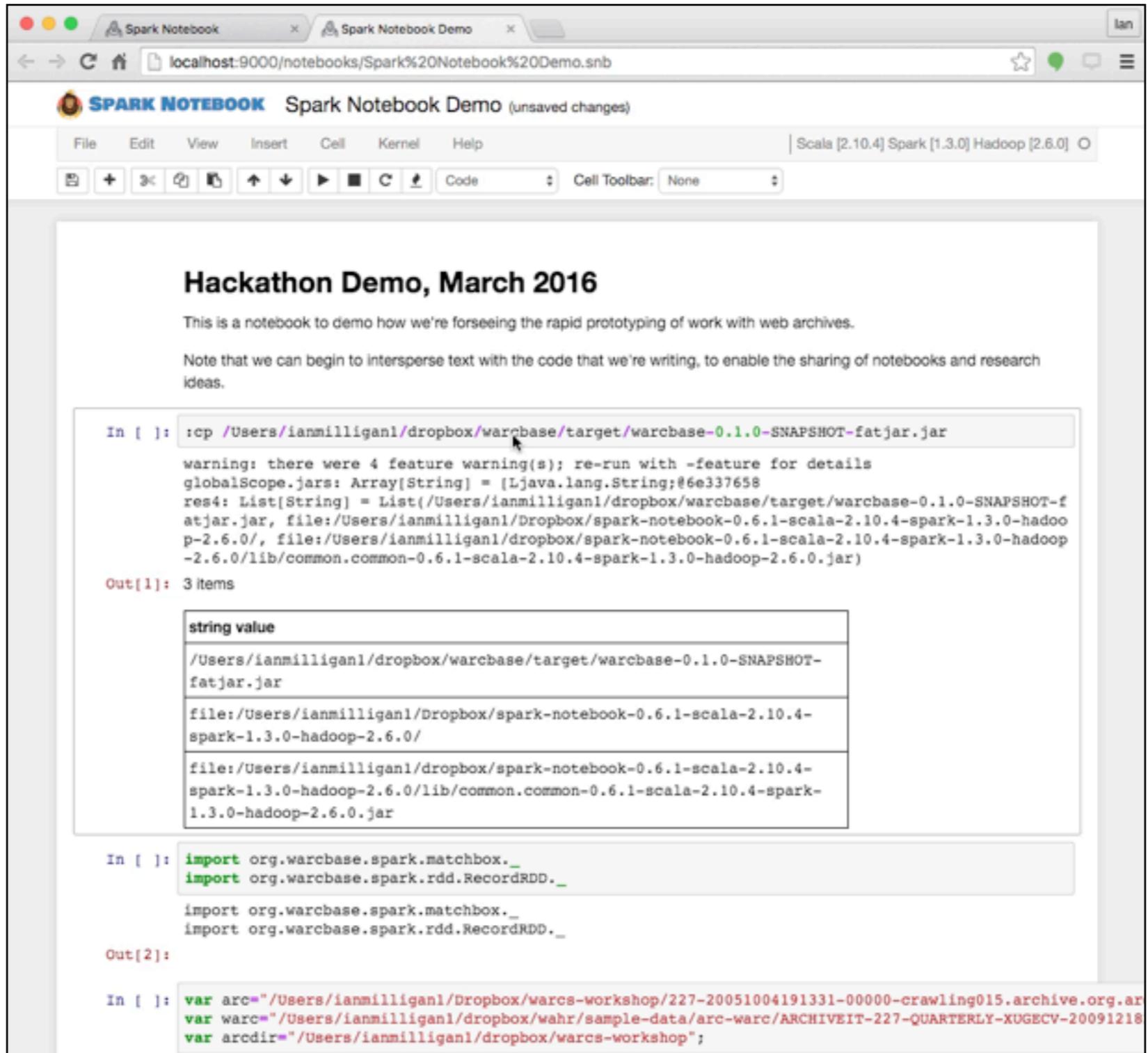
```
i2millig@rho: /mnt/vol1/data_sets/geocities/warc (ssh)
bash                                bash          i2millig@rho: /mnt/vol1/data...
GEOCITIES-20091029114236-00191-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029115416-00171-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029123034-00172-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029130439-00173-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029134536-00174-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029140344-00192-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029141553-00193-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029141726-00175-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029144445-00176-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029152153-00177-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029160824-00178-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029164941-00179-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029165837-00194-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029170431-00195-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029171605-00180-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029174154-00181-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029180818-00182-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029182725-00183-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029185858-00184-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029193728-00185-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029194541-00196-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029195911-00197-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029202841-00186-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029221340-00198-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029222459-00199-ia400110.us.archive.org.warc.gz
GEOCITIES-20091030021147-00197-ia400103.us.archive.org.warc.gz
GEOCITIES-20091030021444-00198-ia400103.us.archive.org.warc.gz
GEOCITIES-20091030022413-00171-ia400104.us.archive.org.warc.gz
i2millig@rho:/mnt/vol1/data_sets/geocities/warc$ du -h
4.1T .
i2millig@rho:/mnt/vol1/data_sets/geocities/warc$
```

# Step Two: Basic Shell Analysis



A screenshot of a terminal window titled "1. i2millig@rho: ~/spark-1.5.1 (bash)". The window shows a command being entered: "ianmilligan1@v1020-wn-178-67:~\$ ssh i2millig@rho.library.yorku.ca". The terminal has a dark background with light-colored text. The cursor is visible at the end of the command line.

# Step Two: Basic Analytics



The screenshot shows a Spark Notebook interface running in a web browser. The title bar reads "Spark Notebook Demo" and the address bar shows "localhost:9000/notebooks/Spark%20Notebook%20Demo.snb". The notebook title is "SPARK NOTEBOOK Spark Notebook Demo (unsaved changes)". The toolbar includes File, Edit, View, Insert, Cell, Kernel, Help, and Scala [2.10.4] Spark [1.3.0] Hadoop [2.6.0]. Below the toolbar is a toolbar with various icons for cell operations like copy, paste, run, etc. The main content area displays a section titled "Hackathon Demo, March 2016" with the following text:

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.

Below this, there are three code cells:

# Step Three: Filtering a Corpus

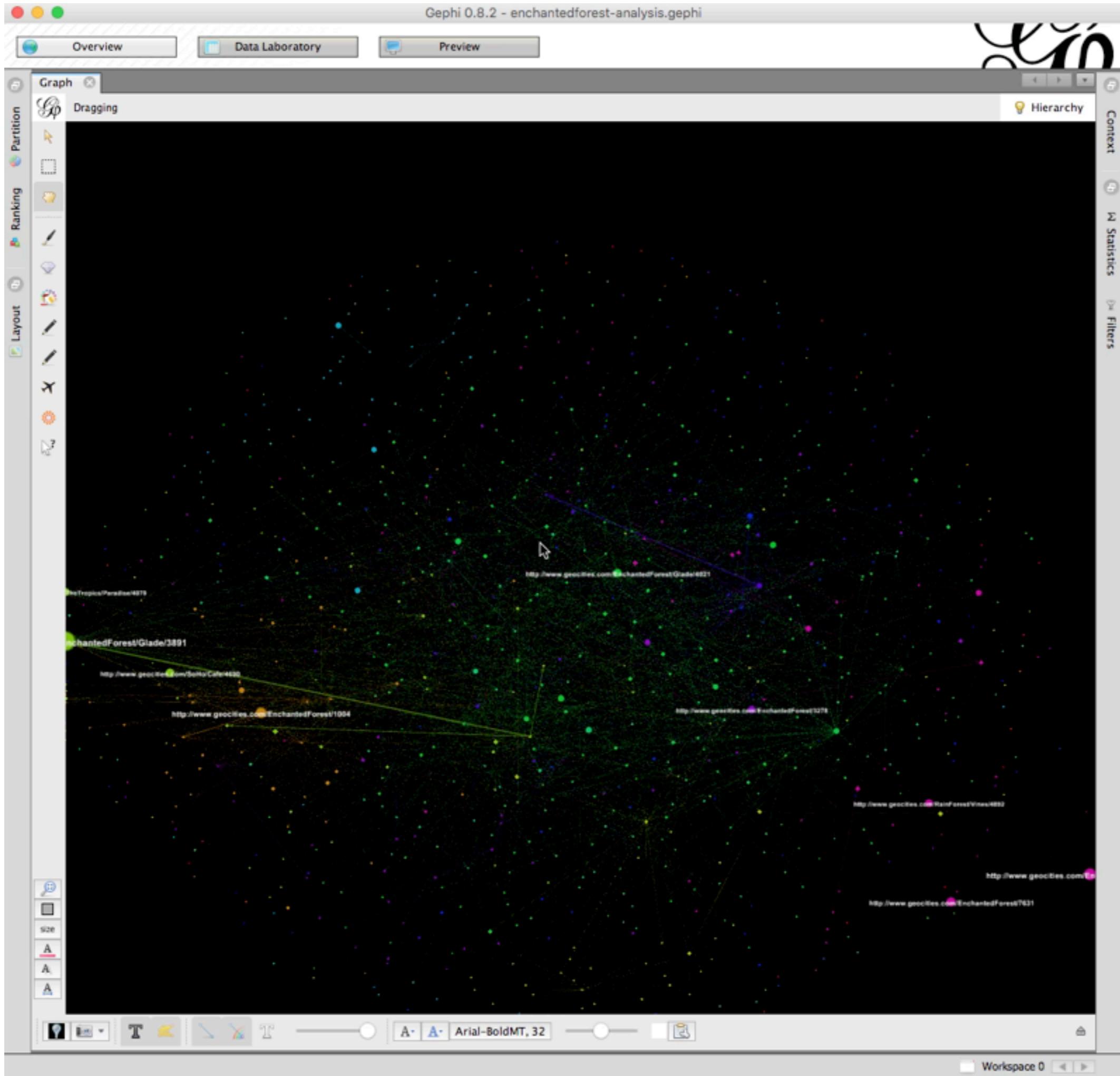
```
1 import org.warcbase.spark.matchbox.{ExtractTopLevelDomain,  
    ExtractLinks, RecordLoader}  
2 import org.warcbase.spark.rdd.RecordRDD._  
3  
4 RecordLoader.loadArc("/mnt/vol1/data_sets/geocities/warcs/*", sc)  
5 .keepValidPages()  
6 .map(r => (r.getCrawldate, ExtractLinks(r.getUrl, r.  
    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

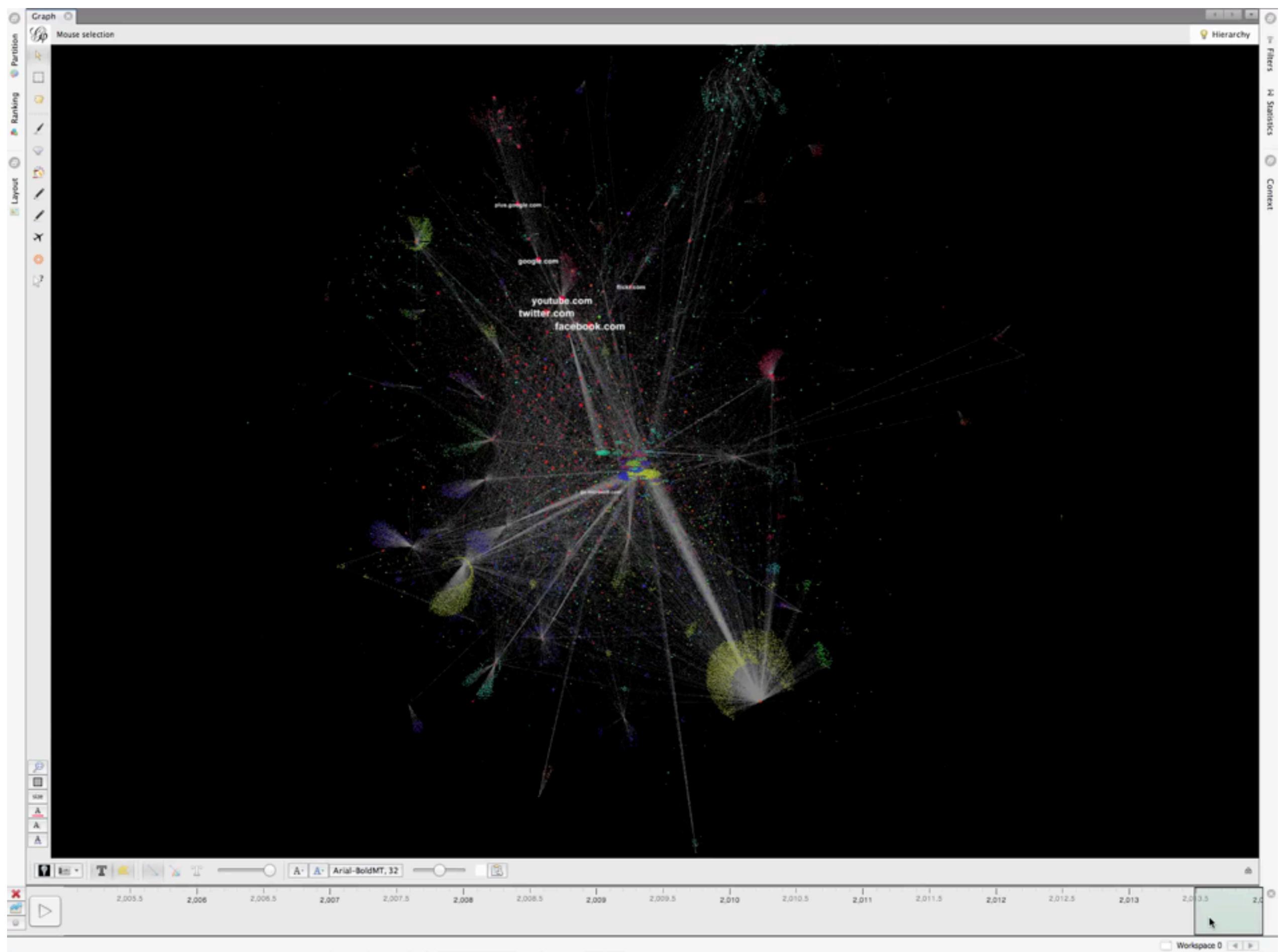
```
1 ((20090903,http://geocities.com/saganaki2000/ADSLGR/adslgr.htm,  
http://www.adslgr.com),15337)  
2 ((20091026,http://geocities.com/saganaki2000/ADSLGR/adslgr.htm,  
http://www.adslgr.com),15337)  
3 ((20091027,http://geocities.com/spankbank69hard/,http://pg.photos  
.yahoo.com/ph/spankbank69hard/my_photos/),9807)  
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

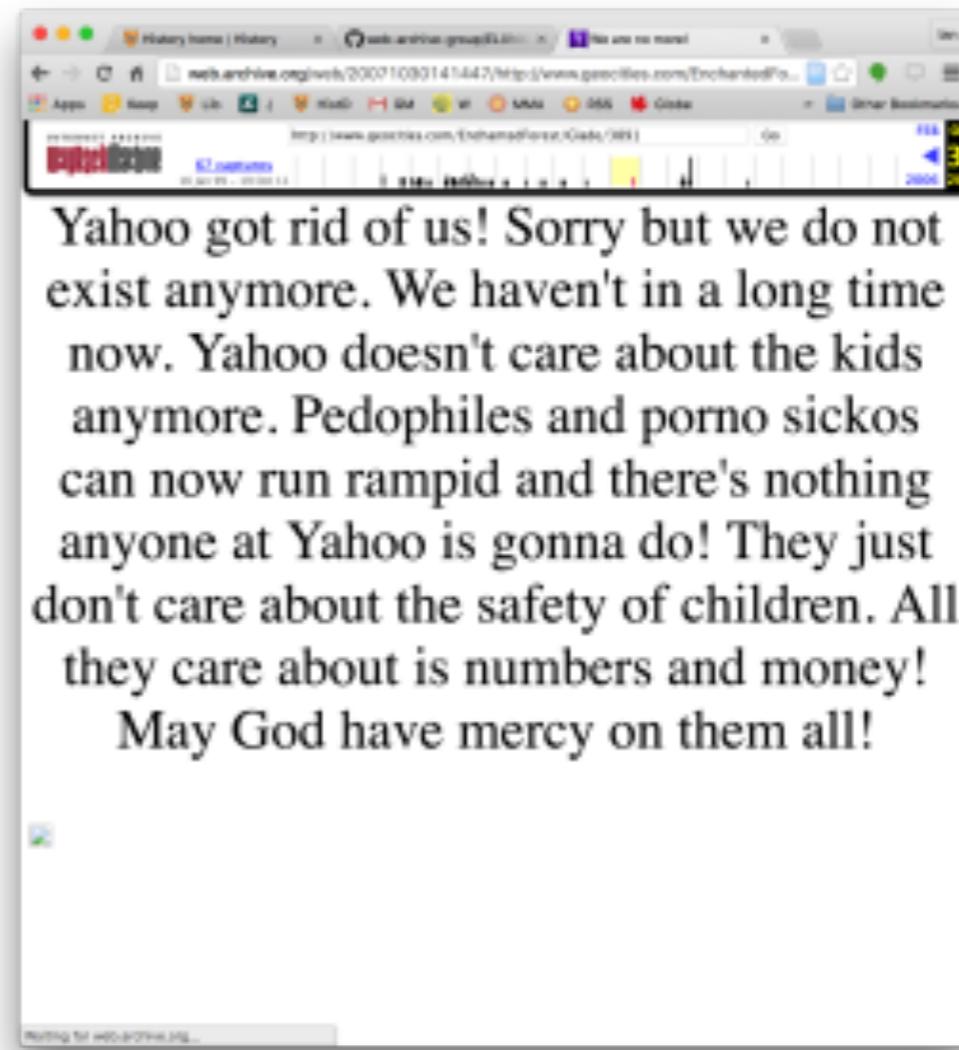


# Finding cool sites!

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



# Step Four: Finding Significant Sites w/ PageRank



# Step Five: Text Analysis

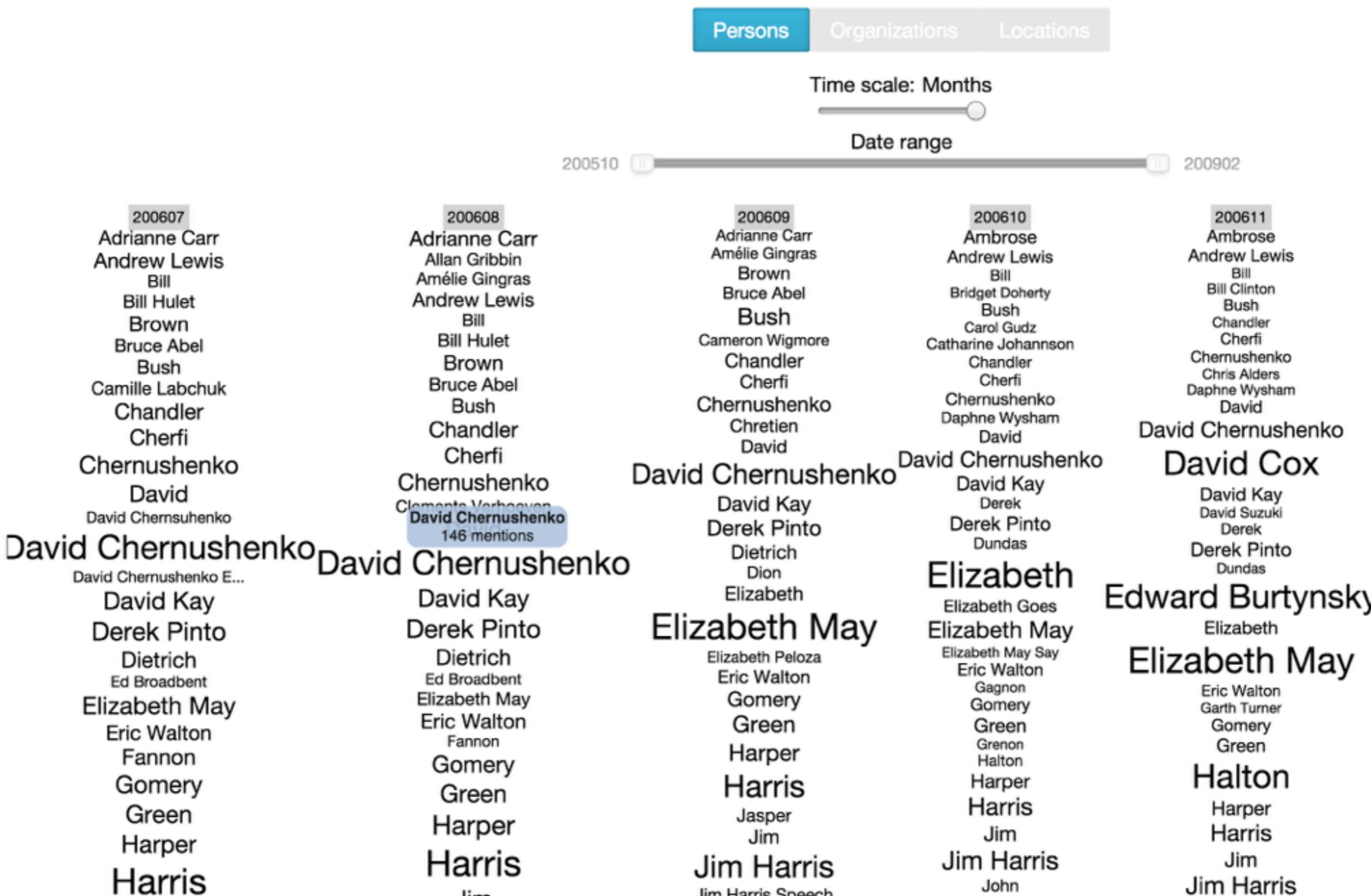
# Different Ways to Filter

- 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!

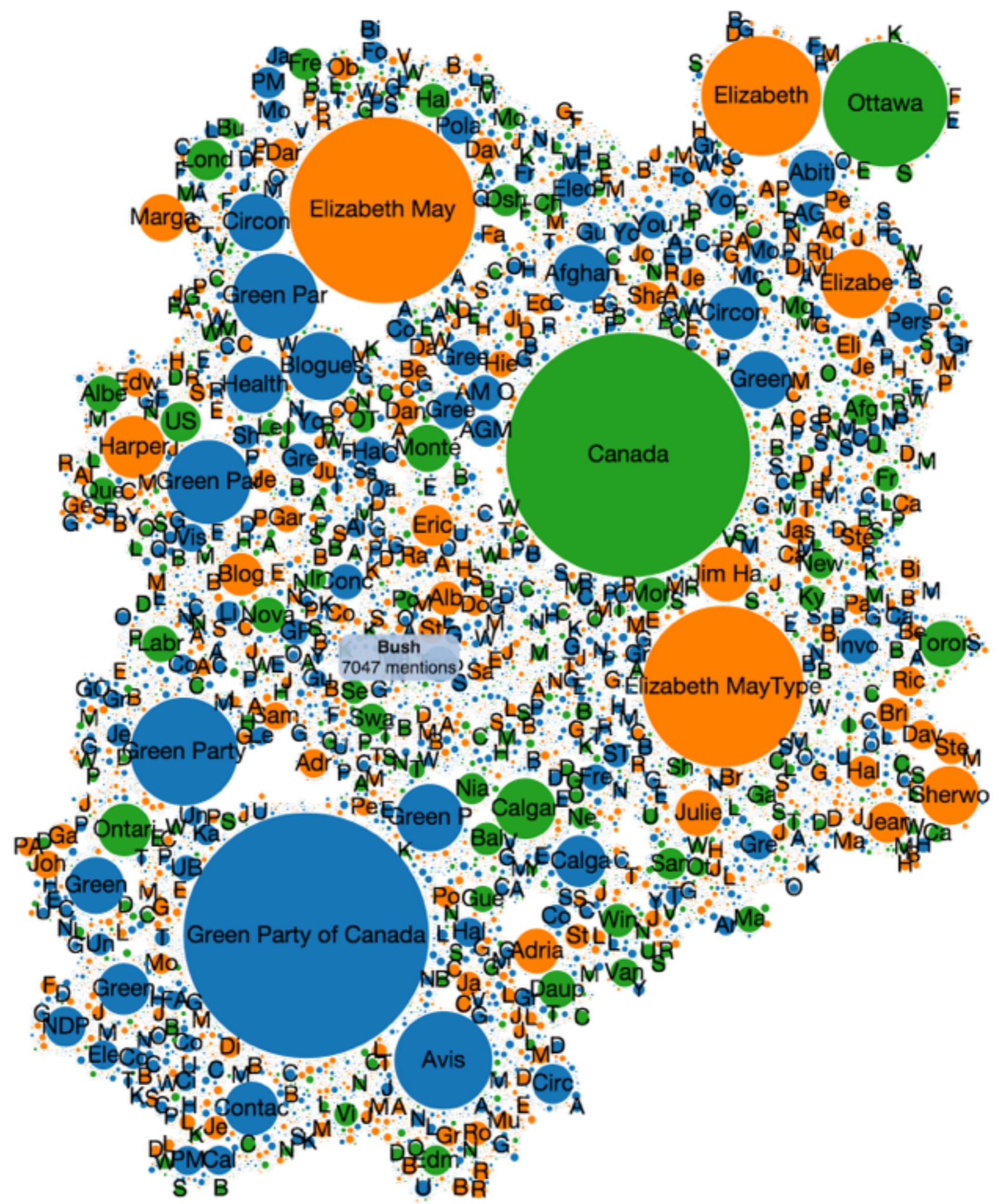


# Named Entity Visualization

Data source: [greenparty.csv](#)







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

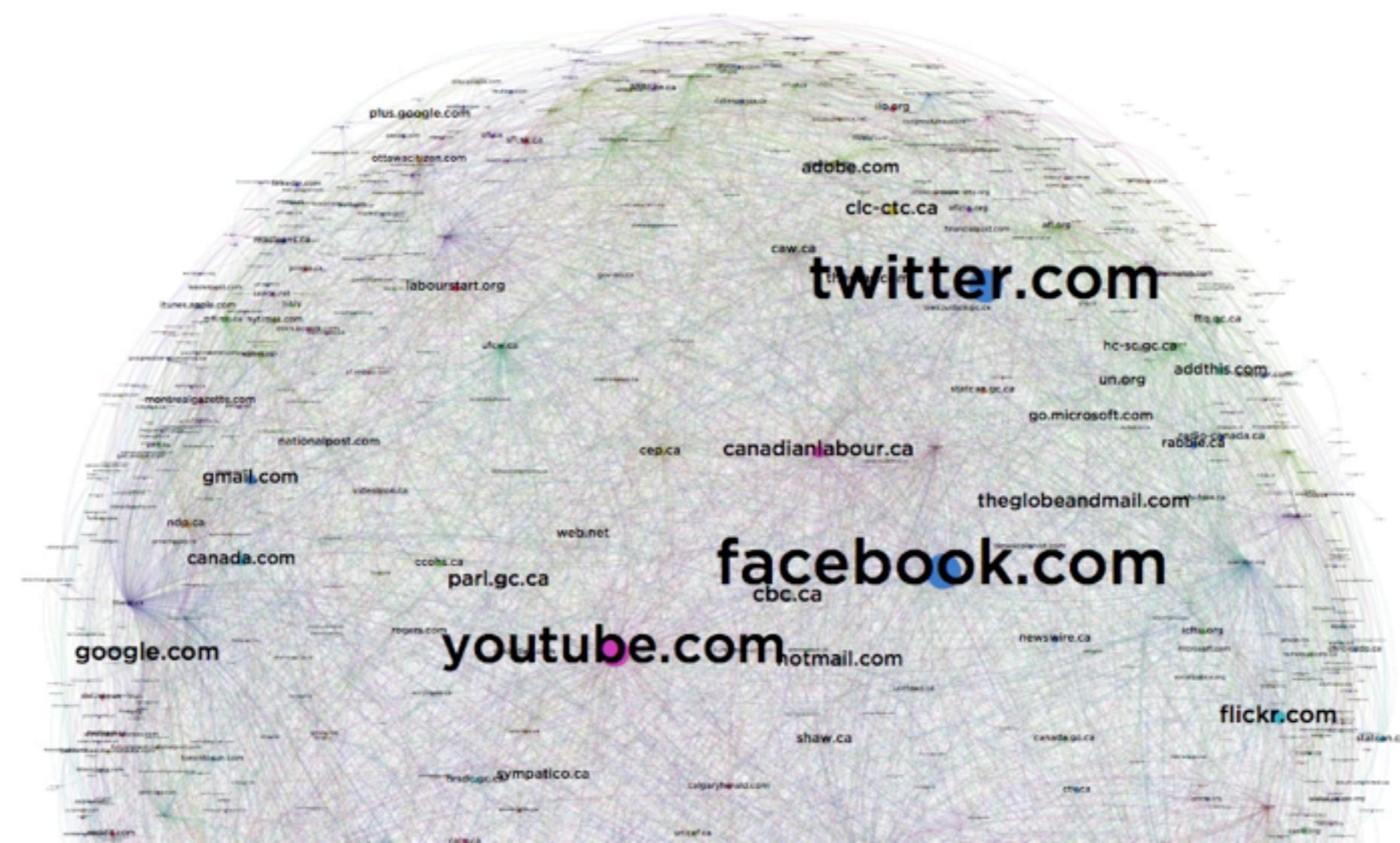
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

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



compute • calcul  
CANADA



UNIVERSITY OF  
**WATERLOO**

# Thanks very much!

## Stay tuned for the technical explanation!