

Exploring the Past with Warcbase

Ian Milligan
Assistant Professor
[@ianmilligan1](https://twitter.com/ianmilligan1)



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Jimmy Lin
Professor and David R. Cheriton Chair
[@lintool](https://twitter.com/lintool)



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

Jeremy Wiebe
PhD Candidate
[@jeremyw](https://twitter.com/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 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 of the Archive-It logo, the title, collector information, and details about the scope and subjects. Below this, there's a section titled "Narrow Your Results" with a search bar and buttons for "Sites" and "Search Page Text". At the bottom, it shows "Page 1 of 1 (54 Total)" and sorting options.

Explore > University of Toronto > Canadian Political Parties and Political Interest Groups

Canadian Political Parties and Political Interest Groups
Collected by: [University of Toronto](#)
Archived since: Oct, 2005
Description: Canadian Political Parties and Political Interest Groups, national Canadian political parties, and a number of special interest groups.
Subject: [Politics & Elections](#)
Collector: [University of Toronto](#)

Narrow Your Results

Enter search terms here

Sites Search Page Text

Page 1 of 1 (54 Total)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Two Case Studies



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

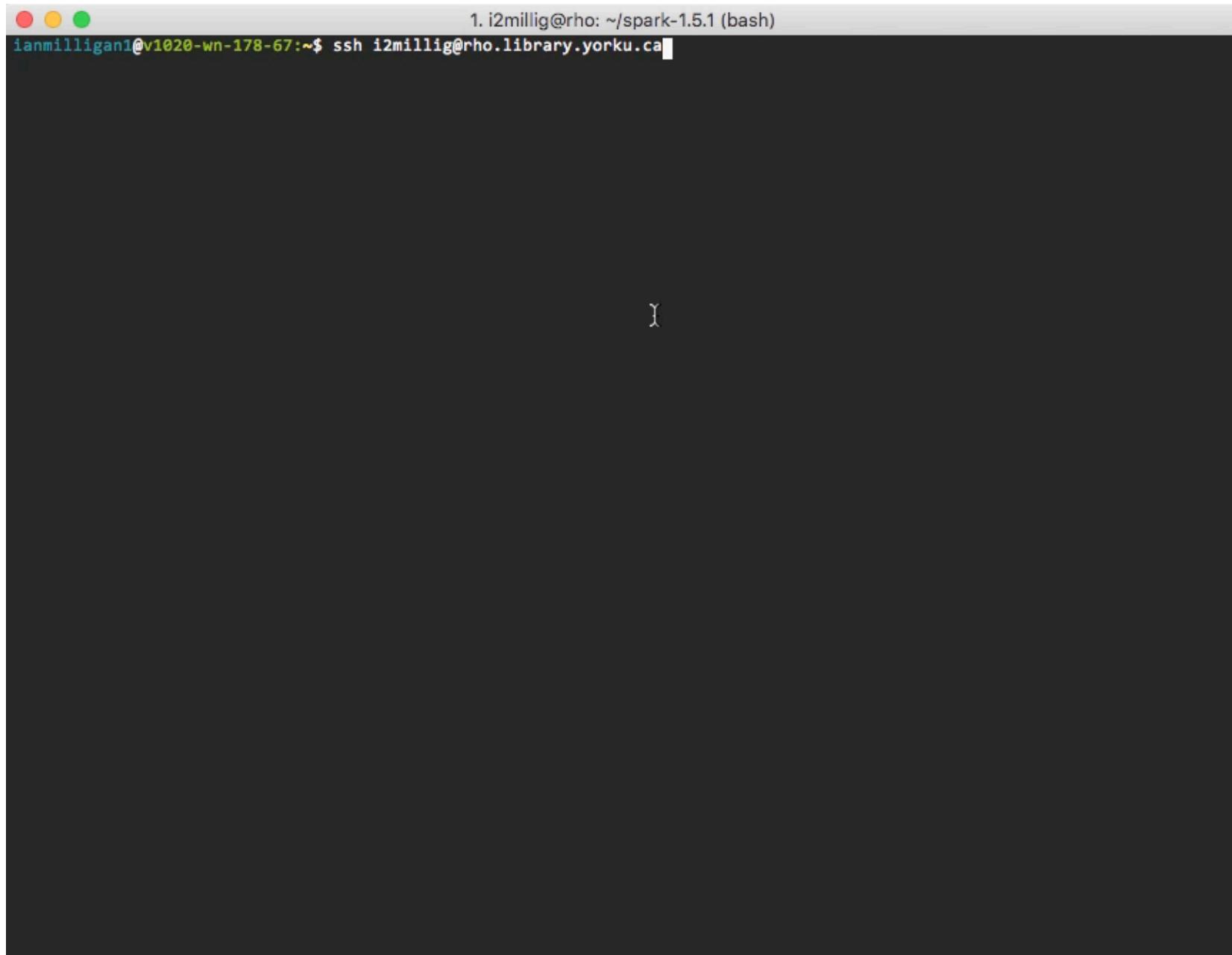
Using Warcbase to
Learn Cool Stuff about it!

Step One: Grabbing WARCs



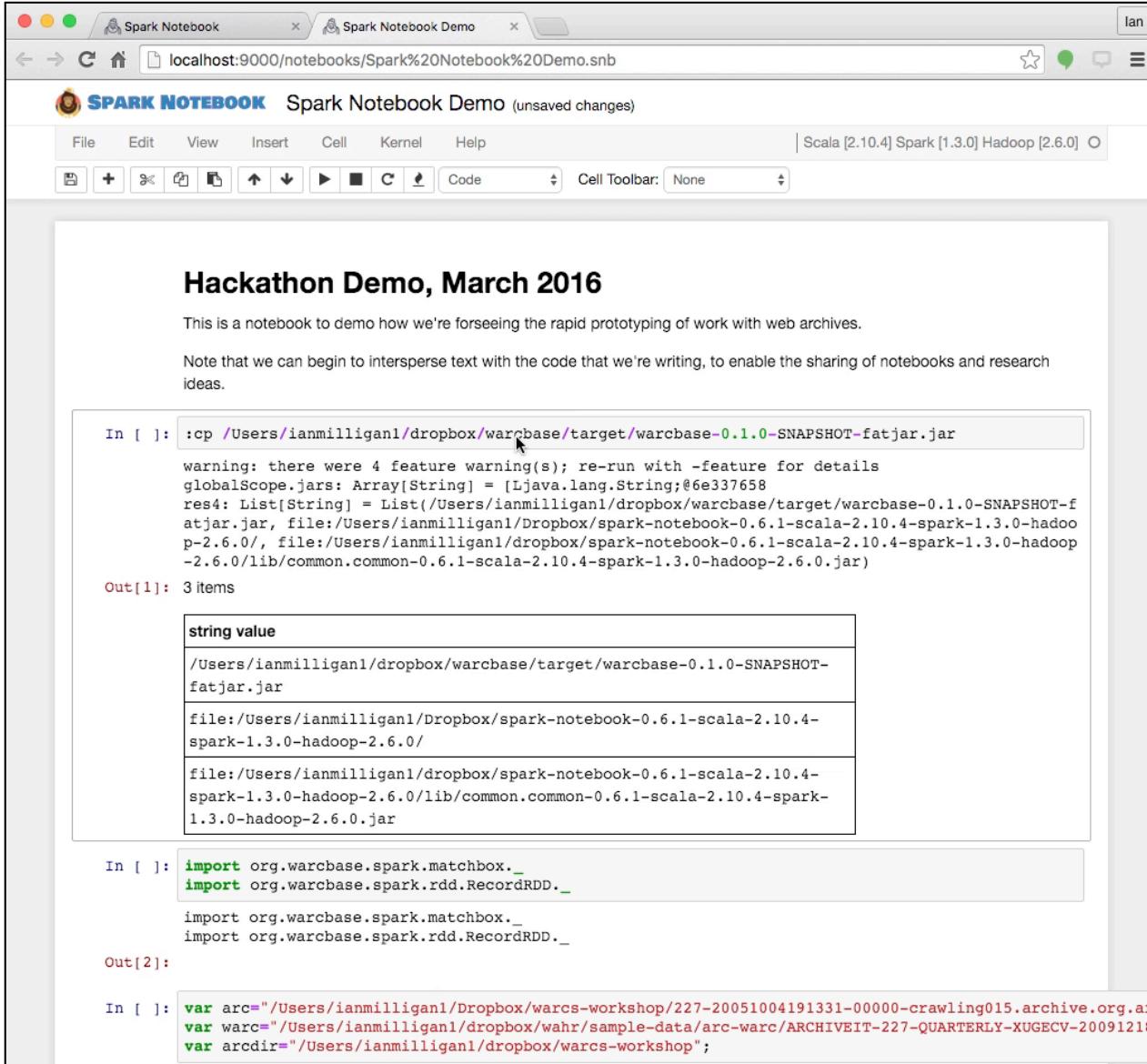
```
1. i2millig@rho: /mnt/vol1/data_sets/geocities/warc$ (ssh)
bash                                bash                                i2millig@rho: /mnt/vol1/data...
GEOCITIES-20091029114236-00191-ia400i10.us.archive.org.warc.gz
GEOCITIES-20091029115416-00171-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029123034-00172-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029130439-00173-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029134536-00174-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029140344-00192-ia400i10.us.archive.org.warc.gz
GEOCITIES-20091029141553-00193-ia400i10.us.archive.org.warc.gz
GEOCITIES-20091029141726-00175-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029144445-00176-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029152151-00177-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029160824-00178-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029164941-00179-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029165037-00194-ia400i10.us.archive.org.warc.gz
GEOCITIES-20091029170431-00195-ia400i10.us.archive.org.warc.gz
GEOCITIES-20091029171605-00180-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029174154-00181-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029180018-00182-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029182725-00183-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029185854-00184-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029193728-00185-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029194541-00196-ia400i10.us.archive.org.warc.gz
GEOCITIES-20091029195911-00197-ia400i10.us.archive.org.warc.gz
GEOCITIES-20091029202041-00186-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029212340-00198-ia400i10.us.archive.org.warc.gz
GEOCITIES-20091029222459-00199-ia400i10.us.archive.org.warc.gz
GEOCITIES-20091030021147-00197-ia400i03.us.archive.org.warc.gz
GEOCITIES-20091030022144-00198-ia400i03.us.archive.org.warc.gz
GEOCITIES-20091030022413-00171-ia400i04.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 typed: "ianmilligan1@v1020-wn-178-67:~\$ ssh i2millig@rho.library.yorku.ca". The terminal has a dark background and light-colored text. The title bar includes three colored dots (red, yellow, green) and the path "1. i2millig@rho: ~/spark-1.5.1 (bash)". The command line itself is white with black text.

Step Two: Basic Analytics



The screenshot shows a Spark Notebook interface running in a web browser. The title bar reads "Spark Notebook" and "Spark Notebook Demo". The main window title is "SPARK NOTEBOOK Spark Notebook Demo (unsaved changes)". The toolbar includes File, Edit, View, Insert, Cell, Kernel, Help, and a Scala [2.10.4] Spark [1.3.0] Hadoop [2.6.0] dropdown. Below the toolbar is a toolbar with various icons for file operations like copy, paste, and search.

Hackathon Demo, March 2016

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 []: :cp /Users/ianmilligan1/dropbox/warcbase/target/warcbase-0.1.0-SNAPSHOT-fatjar.jar

warning: there were 4 feature warning(s); re-run with -feature for details
globalScope.jars: Array[String] = [Ljava.lang.String;@6e337658
res4: List[String] = List(/Users/ianmilligan1/dropbox/warcbase/target/warcbase-0.1.0-SNAPSHOT-fatjar.jar, file:/Users/ianmilligan1/Dropbox/spark-notebook-0.6.1-scala-2.10.4-spark-1.3.0-hadoop-2.6.0/, file:/Users/ianmilligan1/dropbox/spark-notebook-0.6.1-scala-2.10.4-spark-1.3.0-hadoop-2.6.0/lib/common.common-0.6.1-scala-2.10.4-spark-1.3.0-hadoop-2.6.0.jar)

Out[1]: 3 items

string value
/Users/ianmilligan1/dropbox/warcbase/target/warcbase-0.1.0-SNAPSHOT-fatjar.jar
file:/Users/ianmilligan1/Dropbox/spark-notebook-0.6.1-scala-2.10.4-spark-1.3.0-hadoop-2.6.0/
file:/Users/ianmilligan1/dropbox/spark-notebook-0.6.1-scala-2.10.4-spark-1.3.0-hadoop-2.6.0/lib/common.common-0.6.1-scala-2.10.4-spark-1.3.0-hadoop-2.6.0.jar

In []: import org.warcbase.spark.matchbox._
import org.warcbase.spark.rdd.RecordRDD._

import org.warcbase.spark.matchbox._
import org.warcbase.spark.rdd.RecordRDD._

Out[2]:

In []: var arc="/Users/ianmilligan1/Dropbox/warcs-workshop/227-20051004191331-00000-crawling015.archive.org.ar
var warc="/Users/ianmilligan1/dropbox/wahr/sample-data/arc-warc/ARCHIVEIT-227-QUARTERLY-XUGEBCV-20091218
var armdir="/Users/ianmilligan1/dropbox/warcs-workshop";

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.getCreateDate, 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

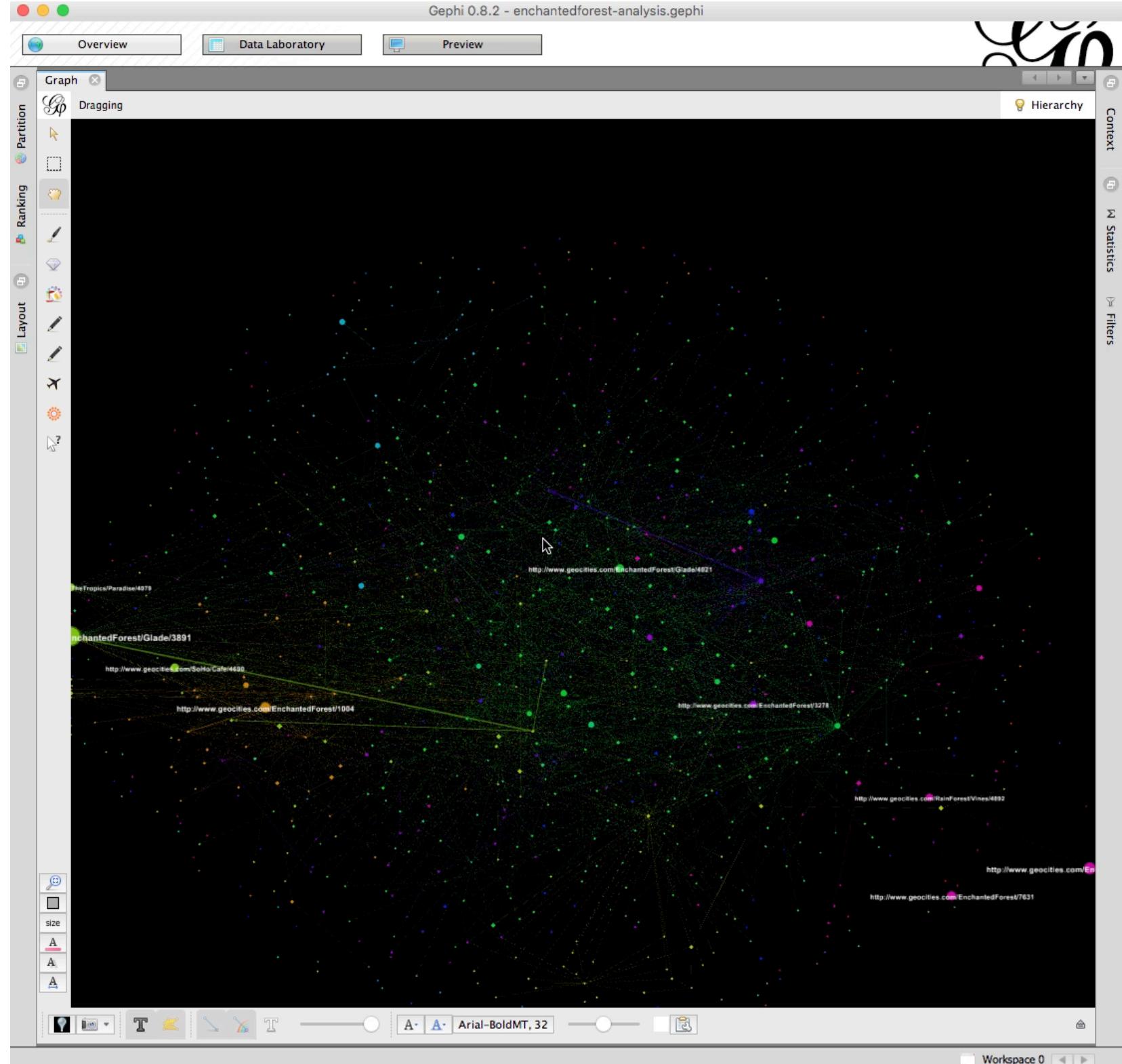
Step Three: Filtering a Corpus

```
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

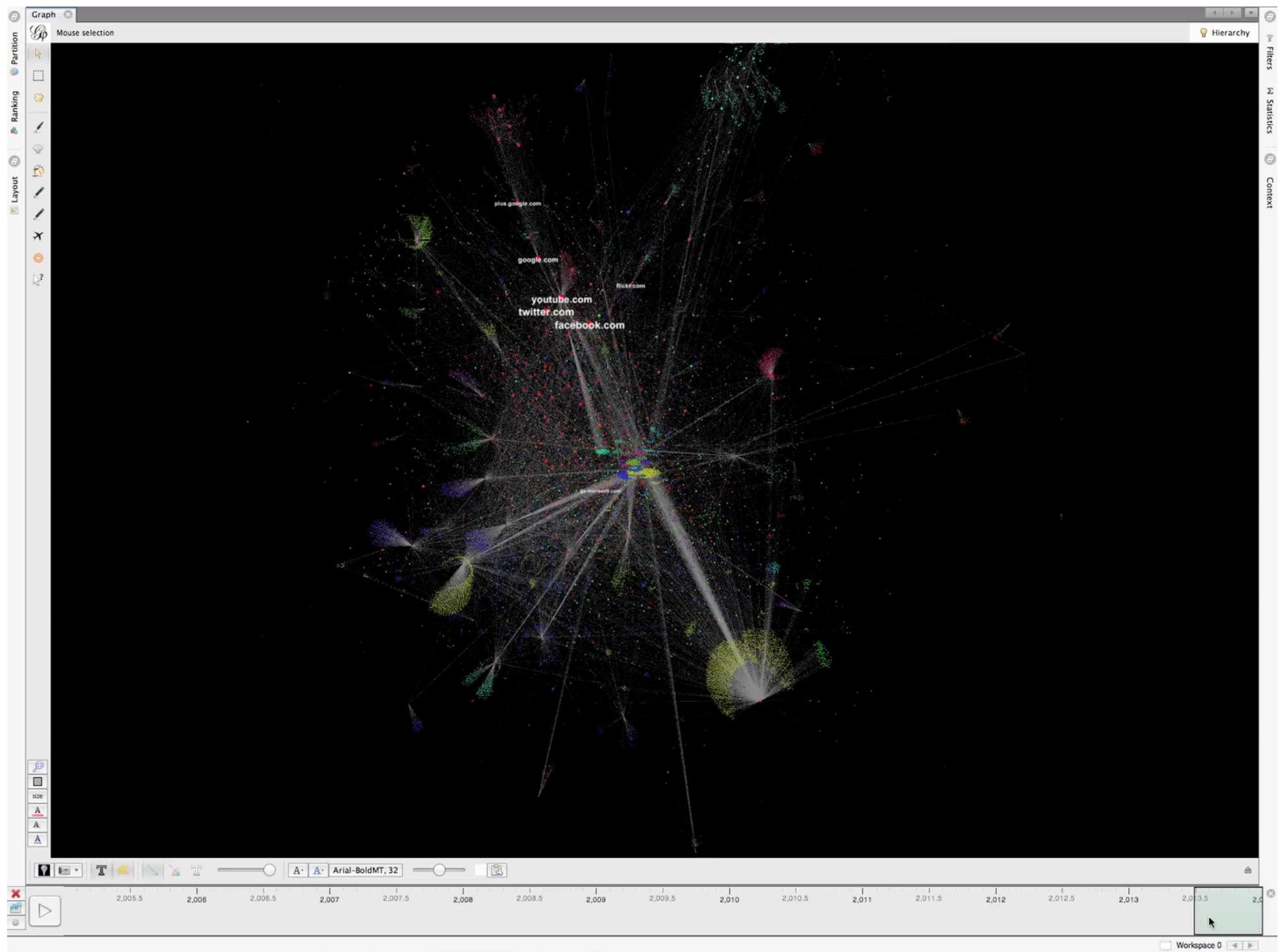
Filtering



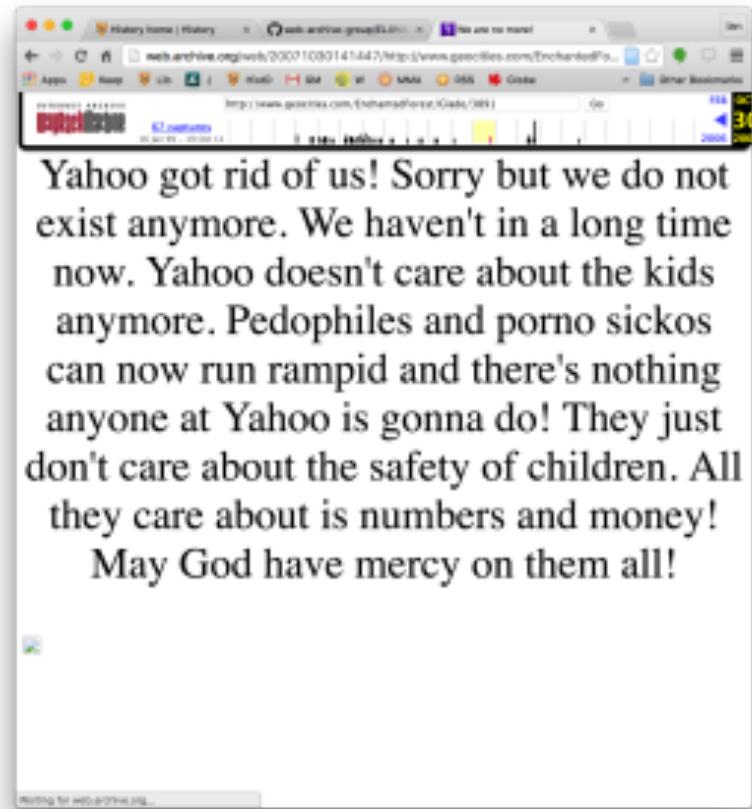
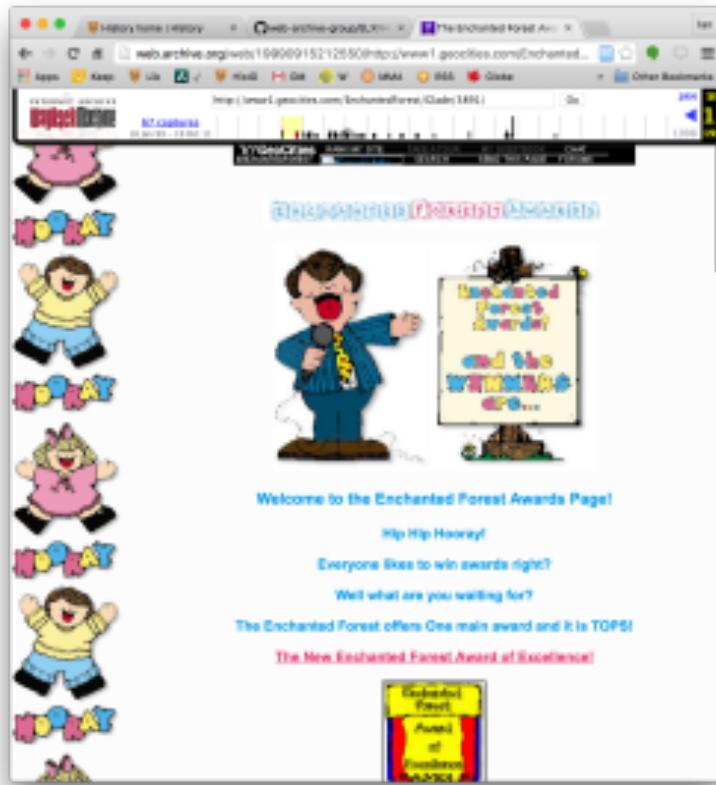


Finding cool sites!

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



Step Four: Finding Significant Sites w/ PageRank



Step Five: Text Analysis

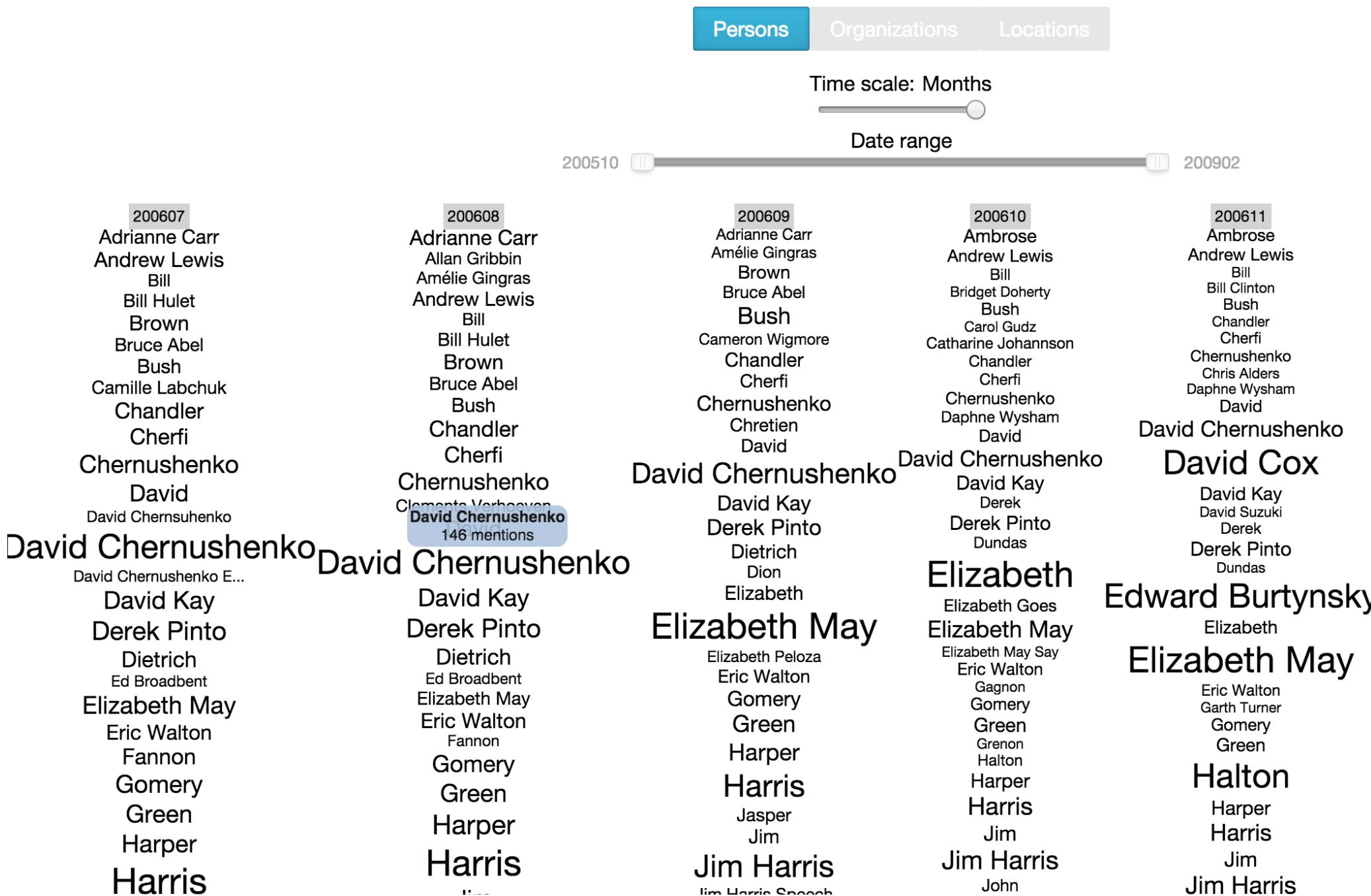
Different Ways to Filter

- 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](#)



Jillian Muo
FLORENCE NIGHTINGALE

Catherine Johnson

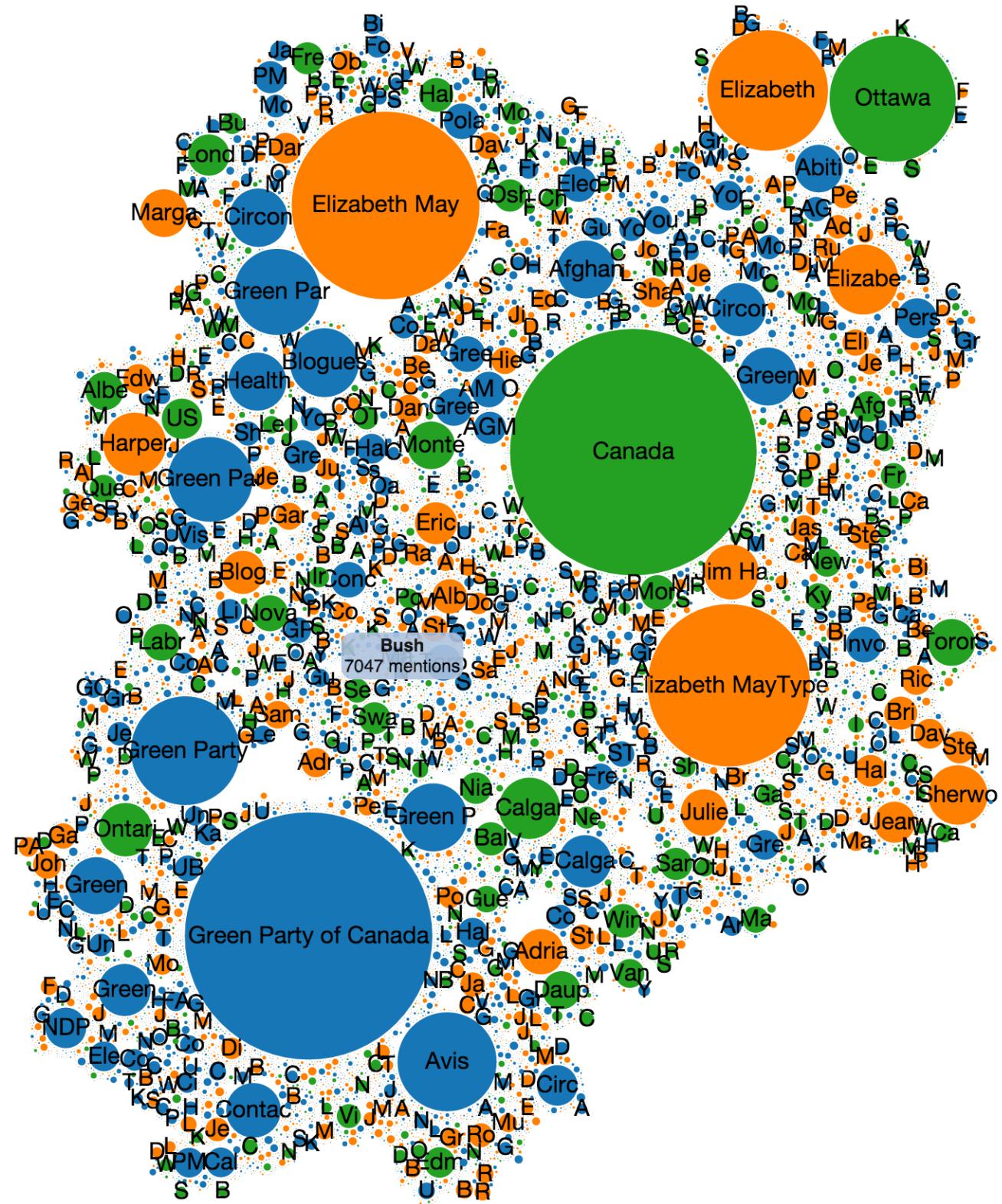
Julian Minn

William King

Julian Minn

FLORENCE NIGHTINGALE

Catherine Johansson



Or generate Solr indexes
using Warchbase 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

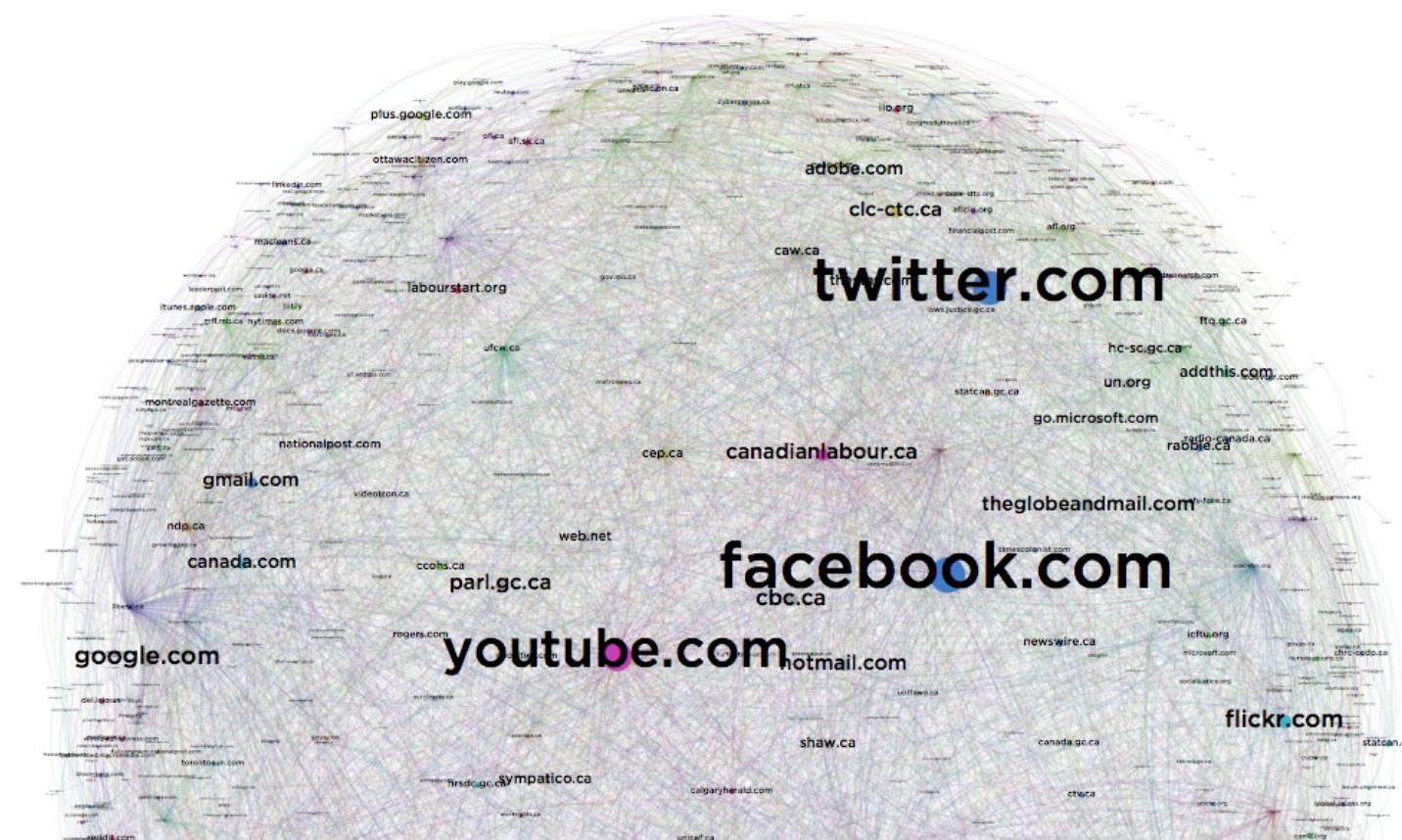
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

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!**