

Hands-on with Warcbase The Workshop

Ian Milligan
Assistant Professor
@ianmilligan1



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Nick Ruest
Digital Assets Librarian
@ruebot





WARC

CATS : ALL YOUR BASE ARE BELONG
TO US.

The Web as a Primary Source

- **Web archives will fundamentally affect the way historians write history**
 - We will have easier access to information on a previously-unknown scale, as well as improved capability to parse it;
 - Yet historians need to reflect on the shape that Web-based primary sources will take, and **how we will be able to access them**

1990s

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



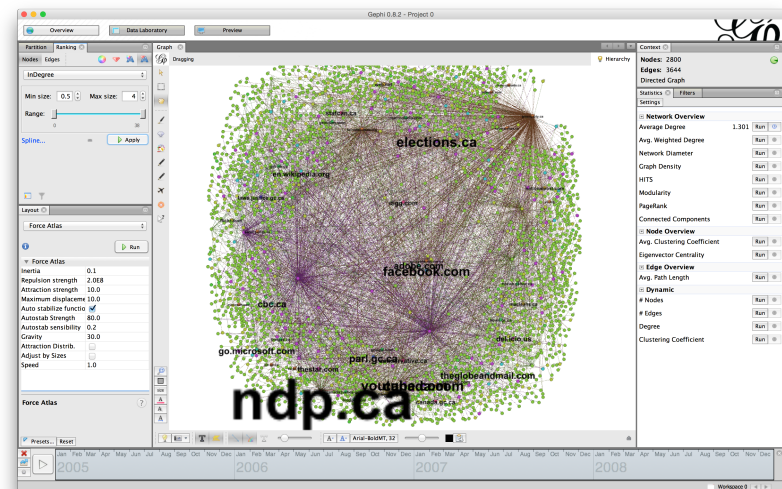
Warcbase

An open-source platform for managing web archives

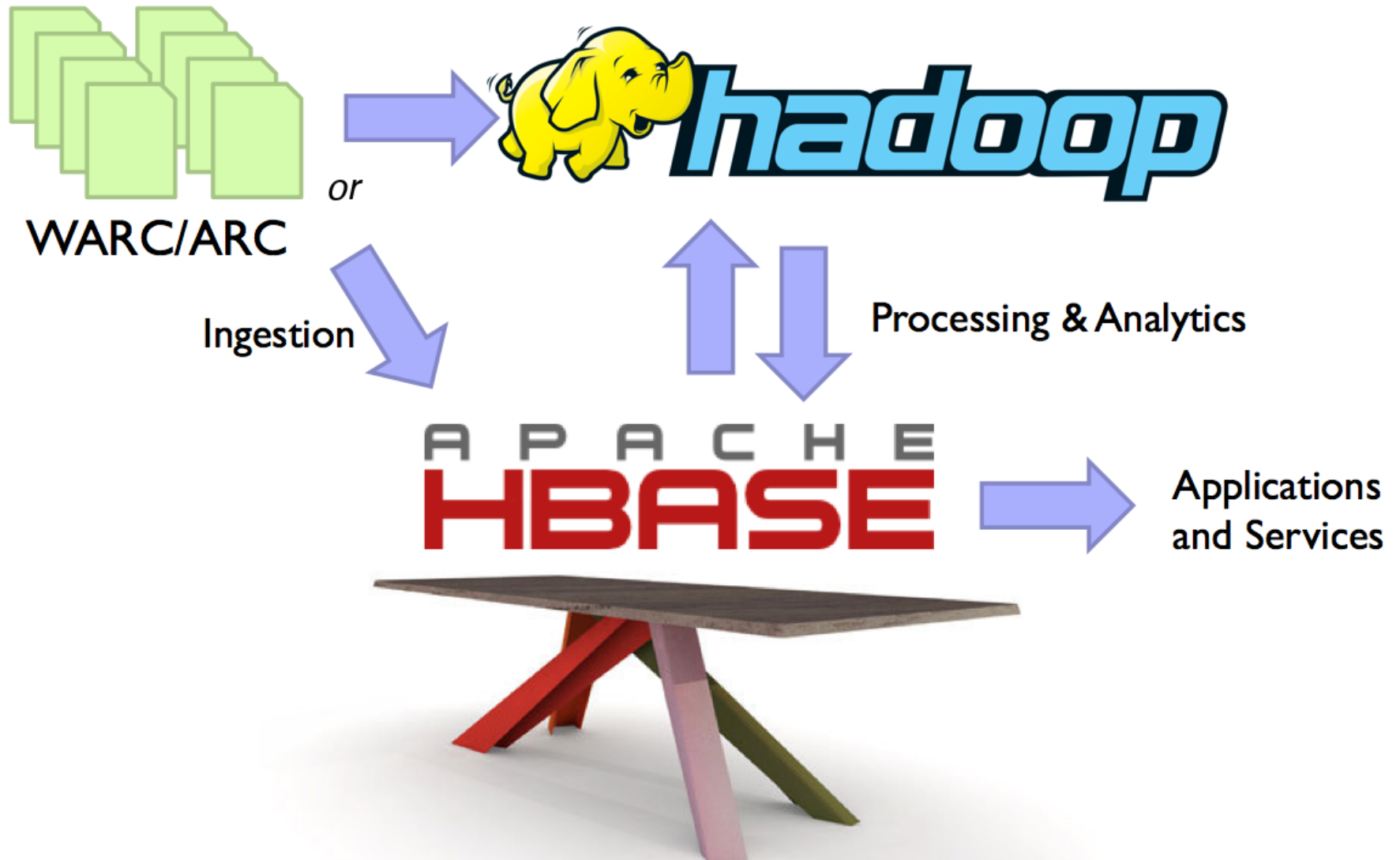
<http://warcbase.org>

Two main facets

- A flexible data store: your own Wayback Machine
- Scriptable analytics and data processing



Warcbase

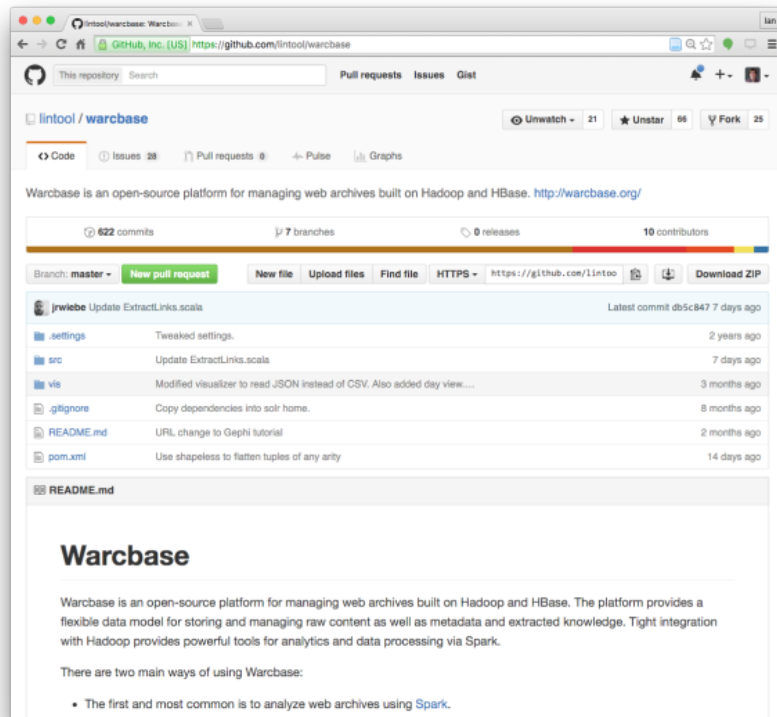


Warcbase

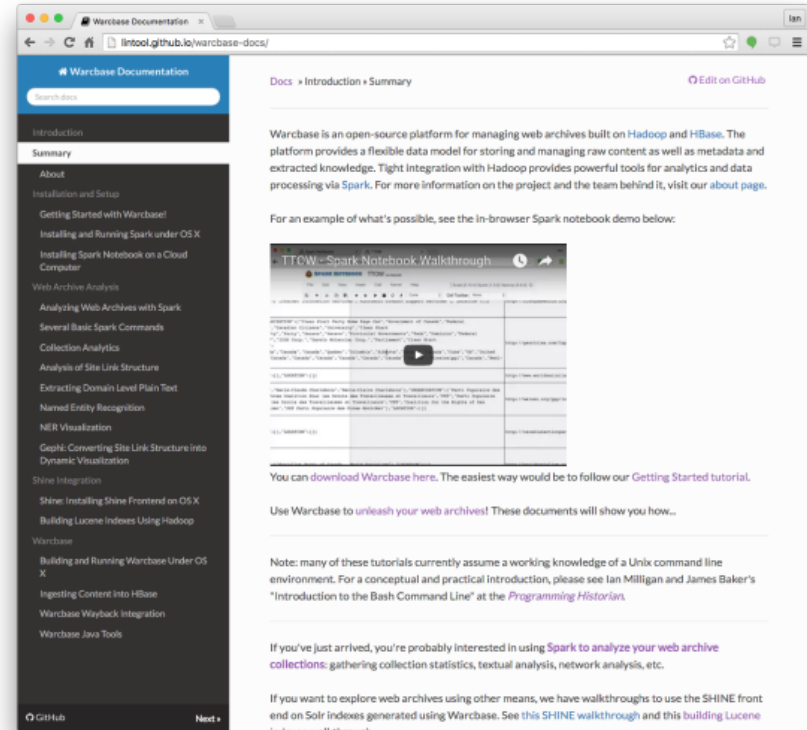
- Framework for distributed storage and distributed processing of very big data
- Scalable
 - From Raspberry Pi to Desktop Computer to Server to Cluster, **all with the same scripts & commands**
- **Potentially very powerful**
 - *Trantor*: 1.2PB of disk, 25 compute nodes (each w/ 128GB memory, 2×6-core Intel Xeon E5 v3 = 3.2TB memory and 300 current-generation Intel cores)



You can Warcbase Too (and will here!)



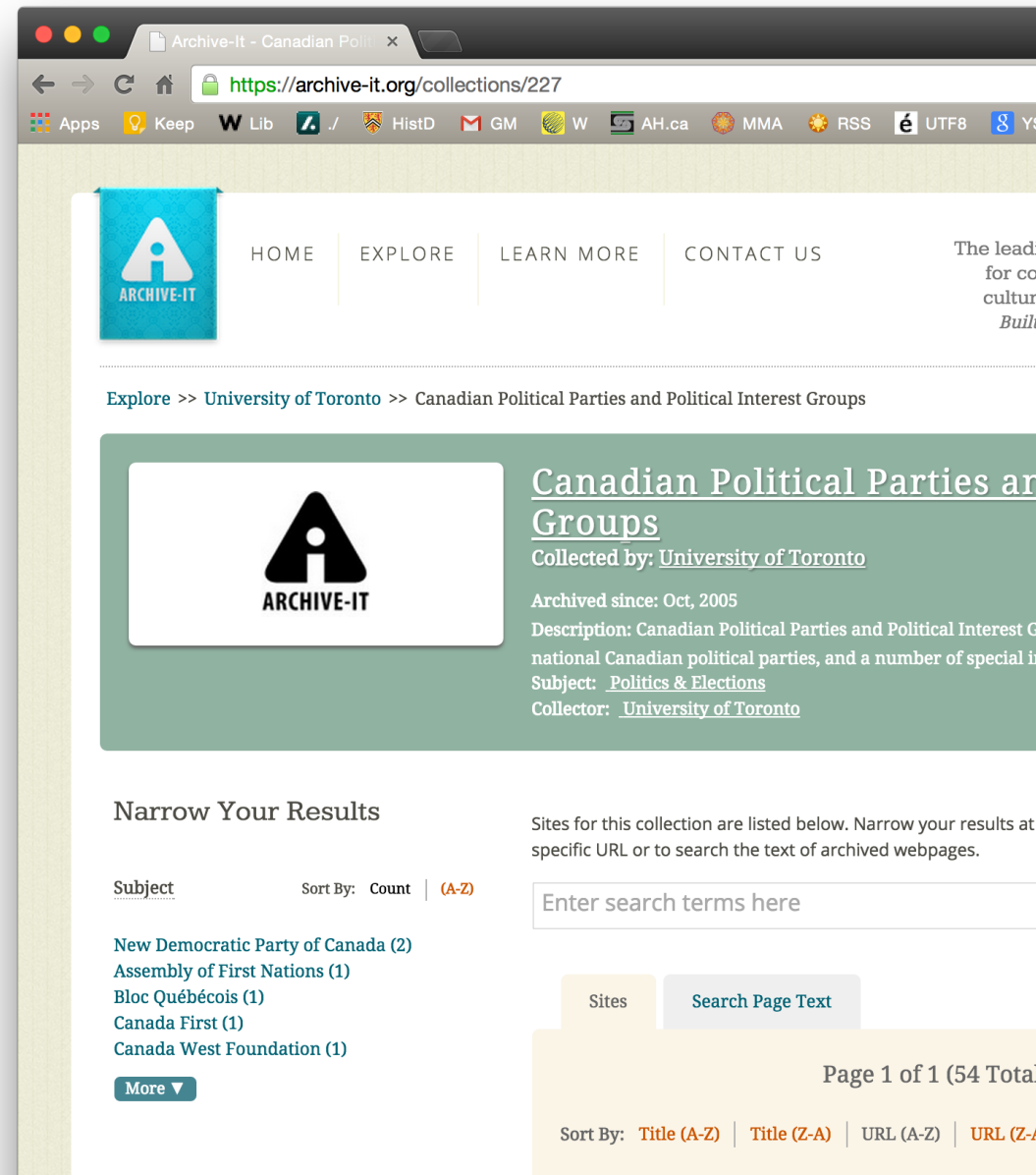
warcbase.org



docs.warcbase.org

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 features the Archive-It logo and navigation links: HOME, EXPLORE, LEARN MORE, and CONTACT US. The breadcrumb trail reads: Explore >> University of Toronto >> Canadian Political Parties and Political Interest Groups. The main heading is "Canadian Political Parties and Interest Groups", collected by the University of Toronto. It notes the collection was archived since October 2005 and describes it as containing national Canadian political parties and special interest groups. The subject is "Politics & Elections" and the collector is the "University of Toronto".

Narrow Your Results

Subject: Politics & Elections | Sort By: Count | (A-Z)

- New Democratic Party of Canada (2)
- Assembly of First Nations (1)
- Bloc Québécois (1)
- Canada First (1)
- Canada West Foundation (1)

[More ▼](#)

Sites for this collection are listed below. Narrow your results at a specific URL or to search the text of archived webpages.

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



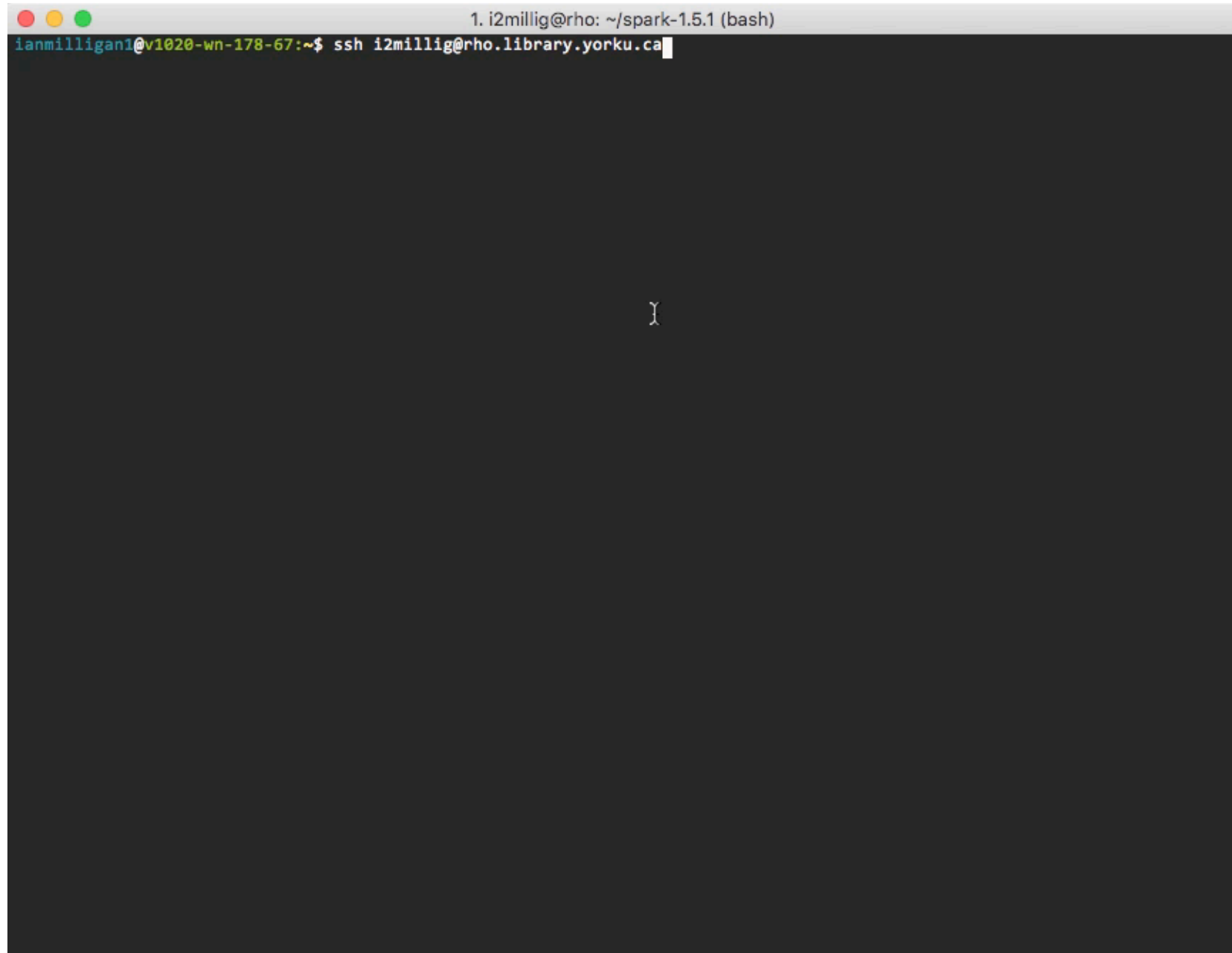
- **GeoCities**
- End-of-life crawl from 2009
- WARC files
- 4.1 TB, 186 million HTML documents

Step One: Grabbing WARCs



```
1. i2millig@rho: /mnt/vol1/data_sets/geocities/warcs (ssh)
bash bash i2millig@rho: /mnt/vol1/data...
GEOCITIES-20091029114236-00191-1a400110.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-20091029140344-00192-1a400110.us.archive.org.warc.gz
GEOCITIES-20091029141553-00193-1a400110.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-20091029165837-00194-1a400110.us.archive.org.warc.gz
GEOCITIES-20091029170431-00195-1a400110.us.archive.org.warc.gz
GEOCITIES-20091029171605-00180-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029174154-00181-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029180818-00182-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029182725-00183-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029185858-00184-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029193728-00185-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029194541-00196-1a400110.us.archive.org.warc.gz
GEOCITIES-20091029195911-00197-1a400110.us.archive.org.warc.gz
GEOCITIES-20091029202041-00186-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029221340-00198-1a400110.us.archive.org.warc.gz
GEOCITIES-20091029222459-00199-1a400110.us.archive.org.warc.gz
GEOCITIES-20091030021147-00197-1a400103.us.archive.org.warc.gz
GEOCITIES-20091030021444-00198-1a400103.us.archive.org.warc.gz
GEOCITIES-20091030022413-00171-1a400104.us.archive.org.warc.gz
i2millig@rho:/mnt/vol1/data_sets/geocities/warcs$ du -h
4.1T
i2millig@rho:/mnt/vol1/data_sets/geocities/warcs$
```

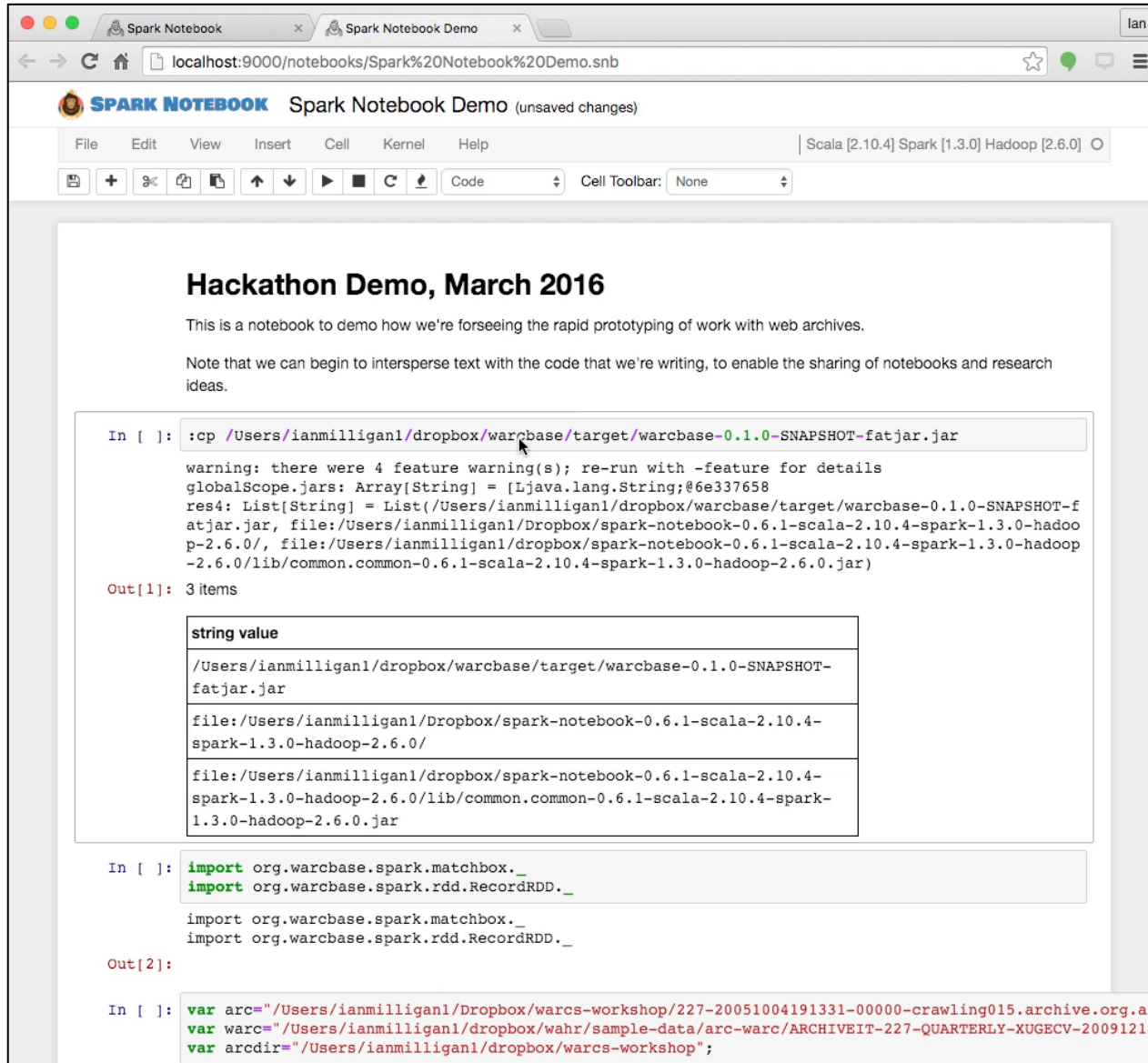
Step Two: Basic Shell Analysis



```
1. i2millig@rho: ~/spark-1.5.1 (bash)
ianmilligan1@v1020-wn-178-67:~$ ssh i2millig@rho.library.yorku.ca
```

The image shows a terminal window with a dark background. The title bar at the top reads "1. i2millig@rho: ~/spark-1.5.1 (bash)". The terminal content shows a prompt "ianmilligan1@v1020-wn-178-67:~\$" followed by the command "ssh i2millig@rho.library.yorku.ca" being entered. A cursor is visible at the end of the command line.

Step Two: Basic Analytics



The screenshot shows a web browser window with a Spark Notebook interface. The notebook is titled "Spark Notebook Demo" and contains the following content:

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-XUGECV-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.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")
```

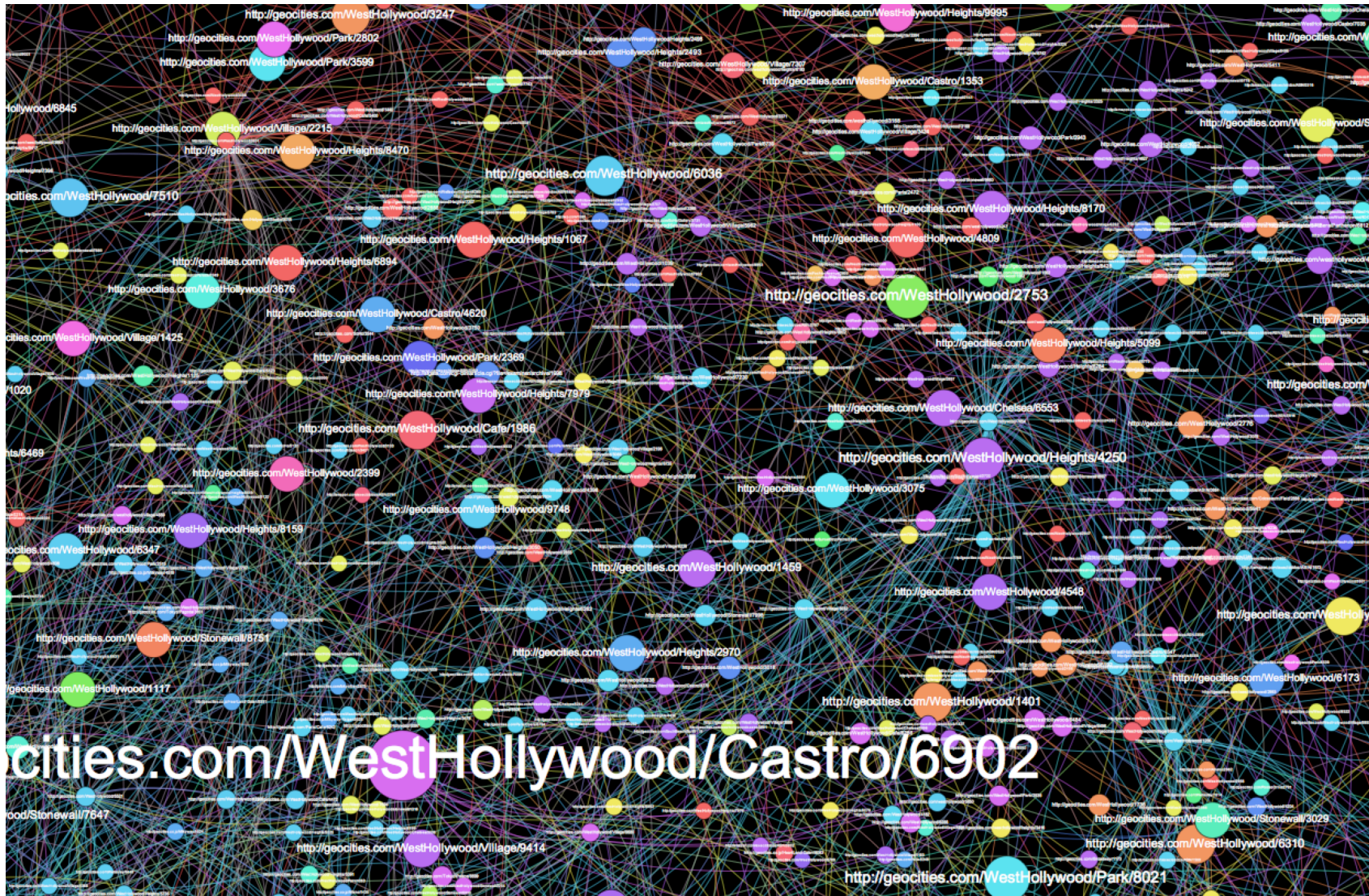
A Link Graph

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)
```

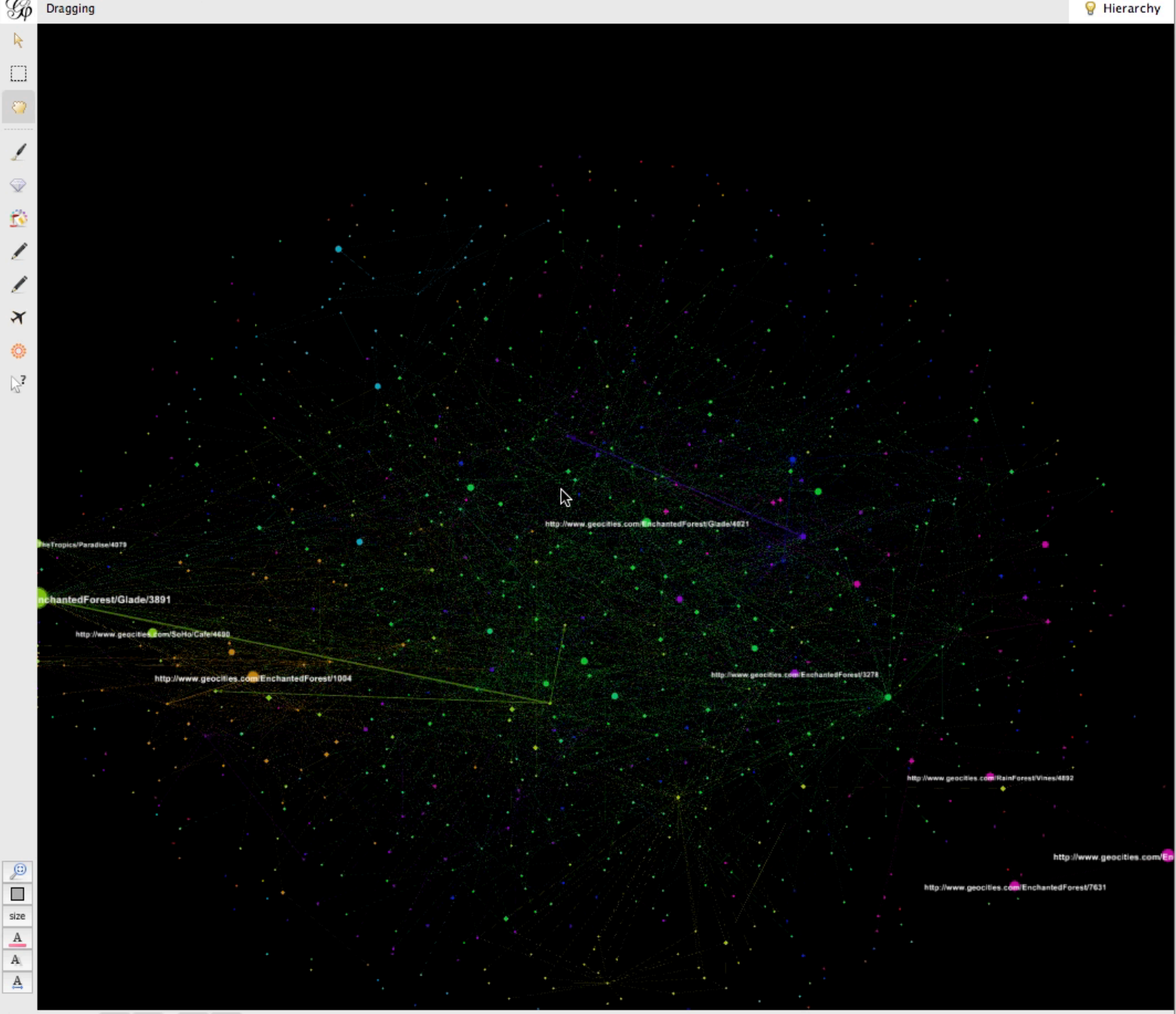
Results

Filtering





- Partition
- Ranking
- Layout

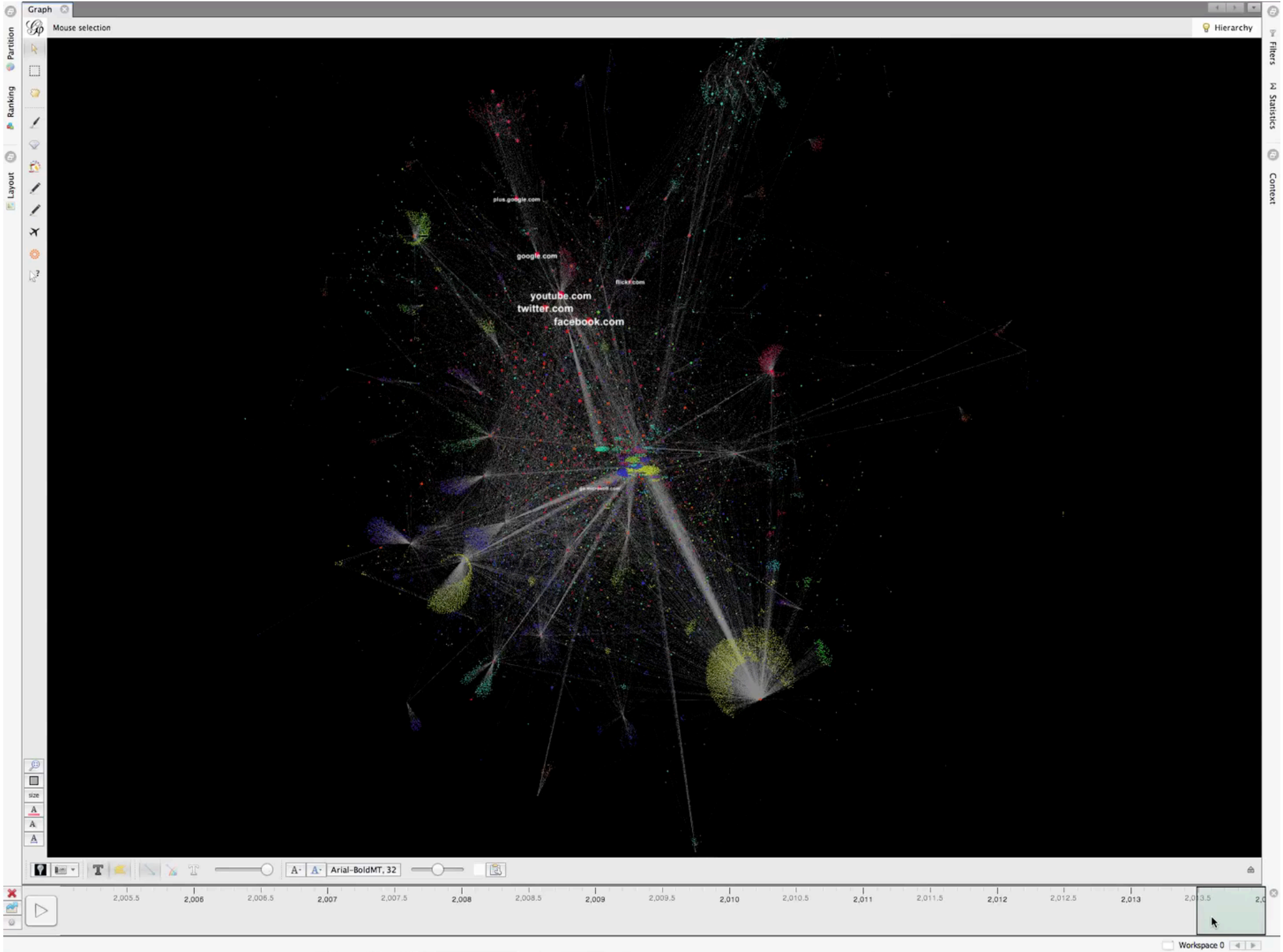


- Context
- Statistics
- Filters

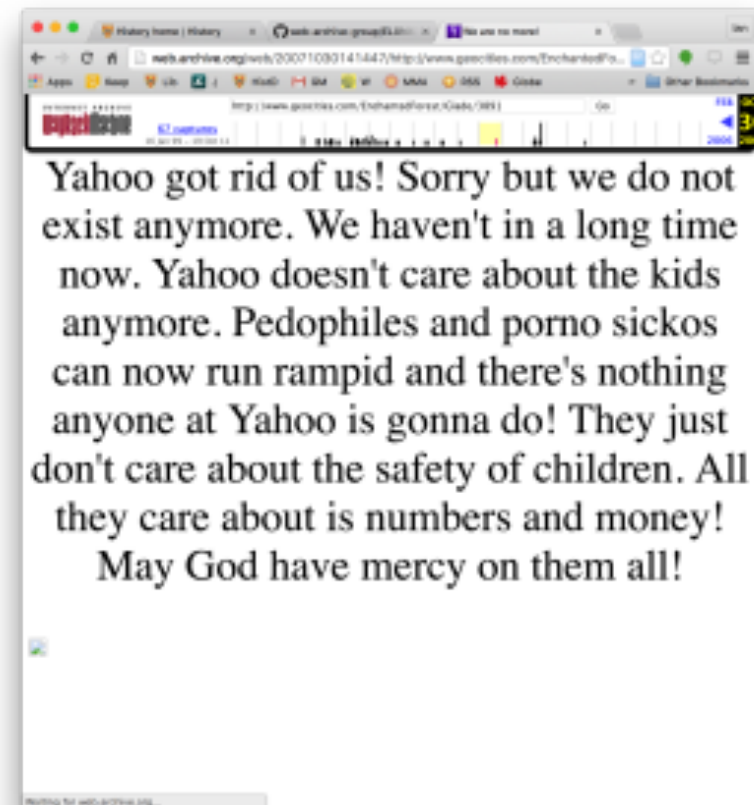
- size
- A
- A
- A

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

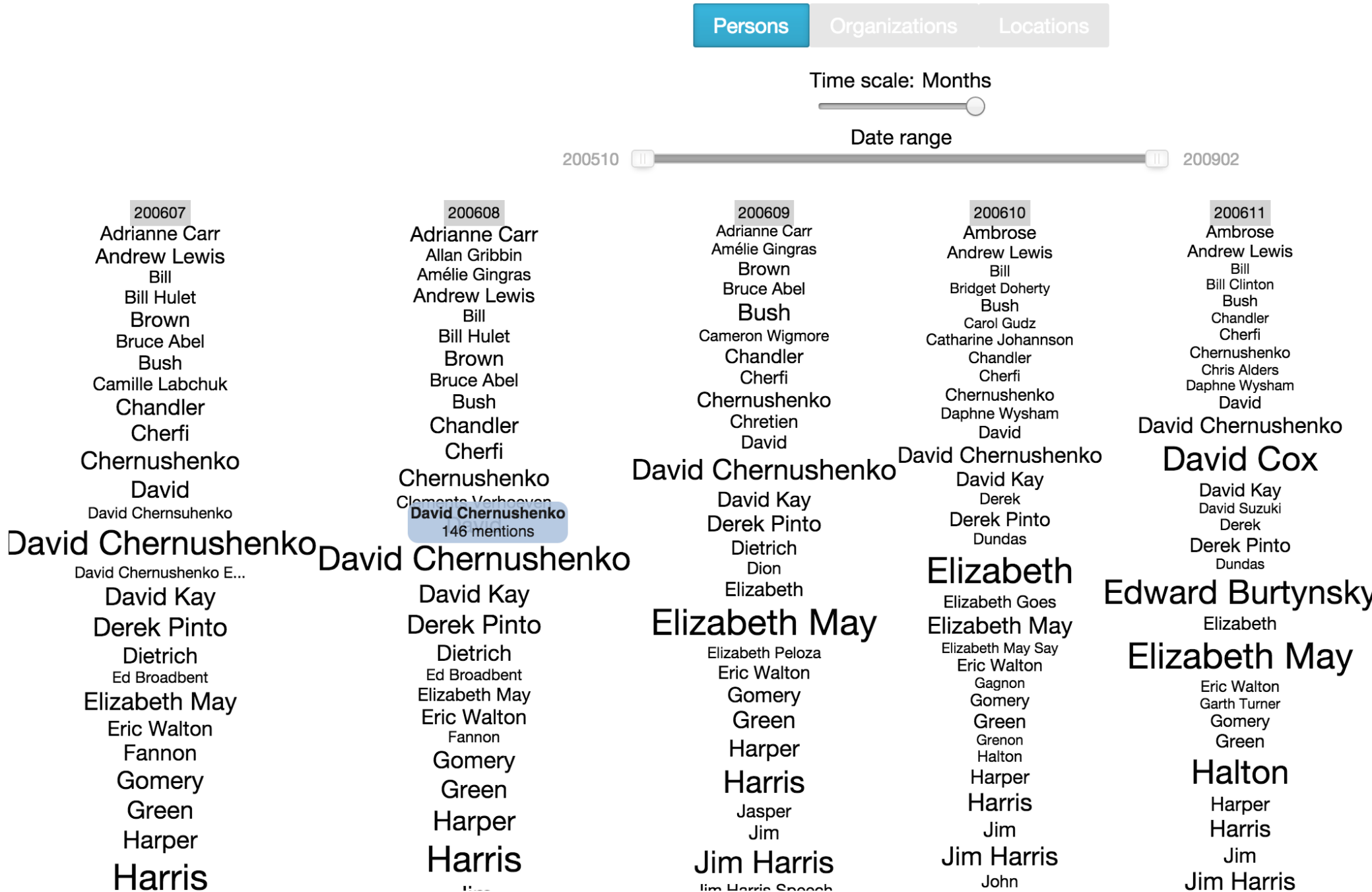
Different Ways to Filter

- 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](#)



Manley
22 mentions



Scott Reid

Jean-François Pinel
Suzanne Laberge

Frank deJong
Andrew Carkner
Sauvageau

Martin C. Barry

Megan Dietrich
Anne McLellan

Thomas Homer-Dixon

Jose Etcheverry
Mulroney

Lori Gadzala
Jack Layton
Ed Broadbent

Fiona Roe
Edward
Mark MacGillivray
George Read

Ralph Klein
Amélie Gingras
Tom Clarke

Harper

Green

Ronald Wright
Tom Manley
Brown

Bruce Abel

Jim Harris
Cherfi
Jim Harris
Speech

Sharon Labeluk
Steve White
Bill Clinton
May

George Bush
Manley
22 mentions

Harris

Bush

Adrienne Carr
Jasper
Lydia Dotto

FLORENCE NIGHTINGALE

Allan Gribbin
Camille Labeluk
David Drake

Steve White
Bill Clinton

Walton
Schiller
Manley

Martin
Steve

Harris

David
Kay

McIntyre

Chandler
Mackenzie

Jim Harris
Dion
Chernushenko

Burr
Dion
Chernushenko

Becky Smit
David Kay
Elizabeth Pelozo
Chris Lackner
William Gavor
Cochrane

Ralph Torrie
Al Gore
Wigmore
ibea

David
Dietrich
John
Chernushenko

Derek Pinto
Stephane Dion
Dion
Chernushenko
May
Dion
Chernushenko
May
Dion
Chernushenko

Bromwich

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

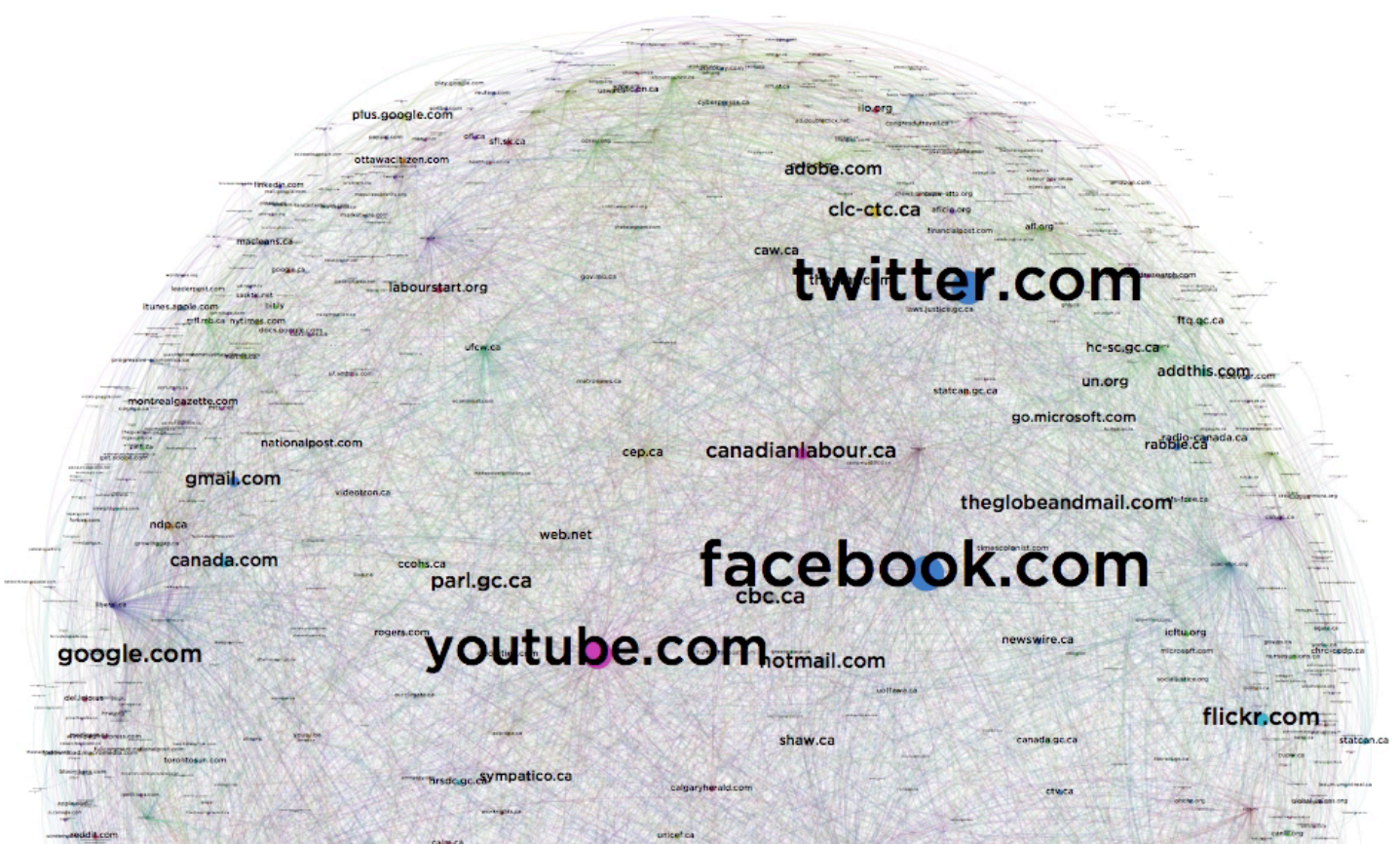
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

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
C A N A D A



UNIVERSITY OF
WATERLOO

**[https://github.com/web-
archive-group/
warcbase_workshop_vagrant](https://github.com/web-archive-group/warcbase_workshop_vagrant)**