Exploring and Discovering Archive-It Collections with Warcbase

Ian Milligan (Assistant Professor, History)
Jimmy Lin (Professor, Computer Science)
Jeremy Wiebe (PhD Candidate, History)
Alice Zhou (BSE Student, Computer Science)

UNIVERSITY OF WATERLOO

Historians are largely unprepared to engage with the quantity of digital sources that will fundamentally transform their trade.


CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Subseries F Democratic Party
BOX 28
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
373
CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series V
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Subseries F Democratic Party
BOX 29
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
374
CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series V
PARTY LEADERSHIP / ADMINISTRATIVE FILE
Subseries F Democratic Party
BOX 30
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
375


CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
374
CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series V
PARTY LEADERSHIP / ADMINISTRATIVE FILE
Democratic Party
BOX 30
JOHN J. BURNS LIBRARY
BOSTON COLLEGE

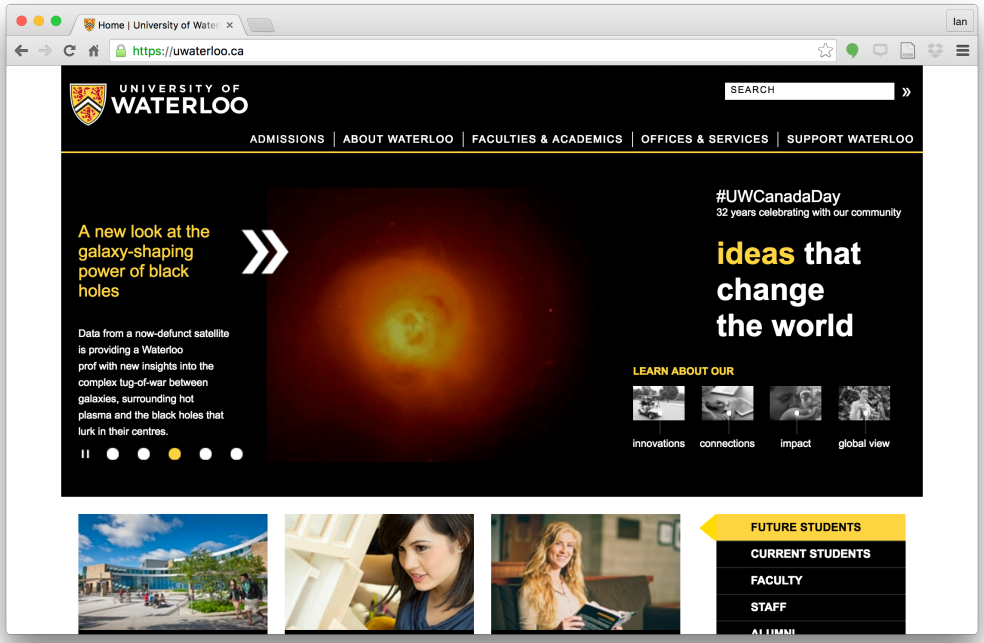
Scarcity

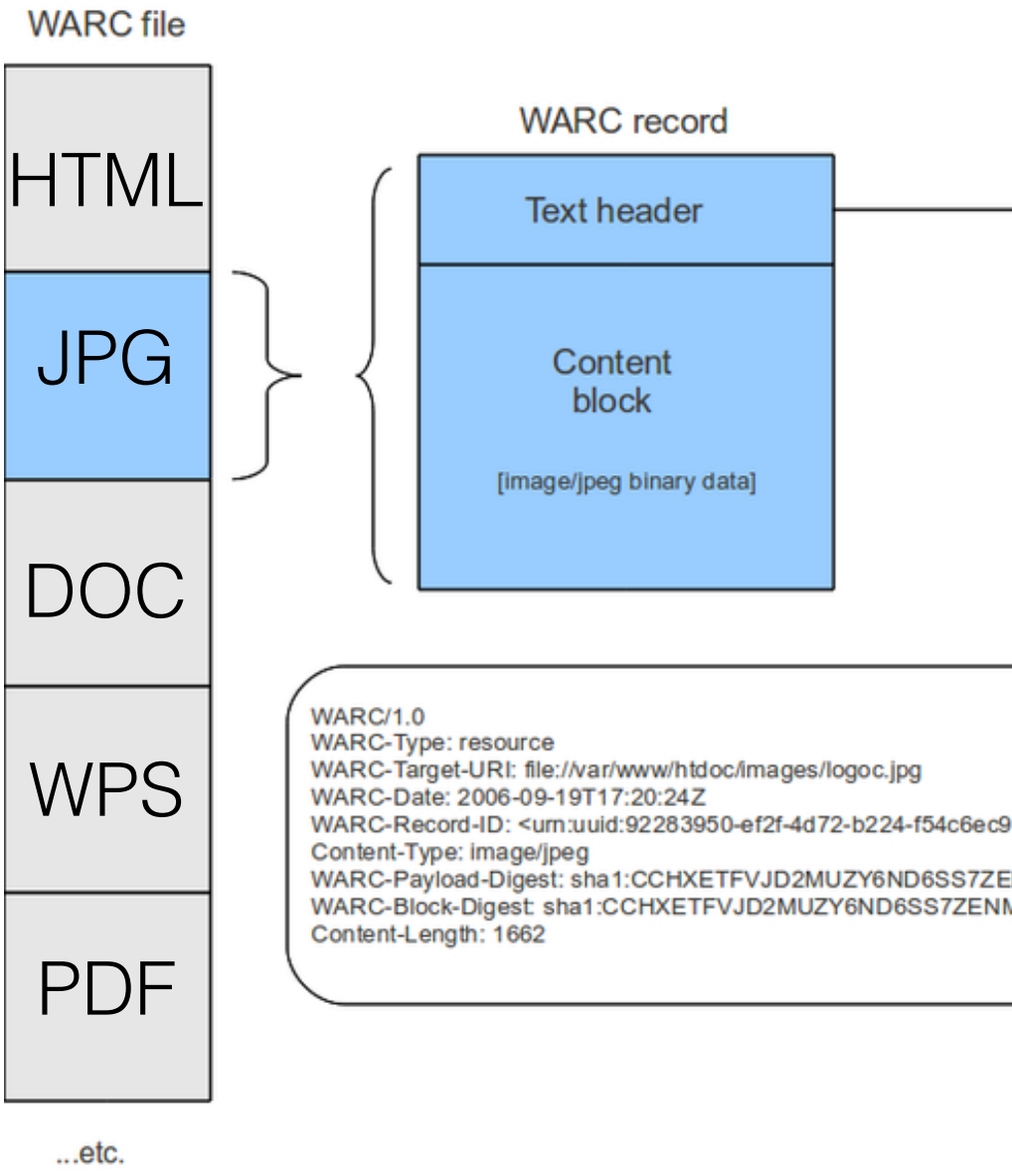
WebARChive (WARC) File

ISO 28500:2009

????

- Imagine your university's website:





Liberal Party of Canada - Hom

web.archive.org/web/19961109132852/http://www.liberal.ca/english/index.html

INTERNET ARCHIVE Wayback Machine

http://www.liberal.ca/english/menu.html Go

6 captures
9 Nov 96 - 25 Jun 04

OCT NOV APR
1995 9 1996 2002

Close Help

Liberal

October 1996

Welcome Cybernauts!

You have arrived at the home page of the Liberal Party of Canada. Perhaps you know who we are?

In fact, you may have read about us in your local newspaper, heard about us on your local radio or seen stories about us on your TV (especially if you're from Canada). We do get lots of exposure in the traditional media...unfortunately, it's all one-way communication. We Liberals are excited about the potential of the World Wide Web...the potential for interactive communication with you!

We are hoping that you will find our site interesting...we are certainly trying to make it so. And we are using your suggestions to improve as we grow. So please keep those e-mails coming!

On behalf of our leader Prime Minister Jean Chrétien, our President Senator Daniel Hays and our National Executive, welcome...enjoy your visit...and come back again!

Yours in surfing,

The WebMaster at LPC

This site is authorized by the Federal Liberal Agency of Canada

OccupyWallStreet

The revolution continues worldwide!

Farmers Join Occupy Wall Street, Calling for Food Justice

Posted 5 hours ago on Dec. 2, 2011, 6:21 p.m. EST by OccupyWallSt

As Wall Street's corrupt influence on the economy has grown, the corporate ownership of our food system has hurt the health and livelihood's of some of our most vulnerable communities. This *Sunday, December 4th* food justice activists and occupiers will be traveling from as far as Colorado, Iowa, Maine and Upstate New York to join together for the Occupy Wall Street FARMERS' MARCH.Through a day of dialogue, musical performances, and a march, farmers and their urban allies working for food justice in their communities will form alliances to fight and expose corporate control of the food supply.

Events throughout the day will call and inspire participants to fight against the corporate manipulation of the agriculture system. An industry that is responsible for using chemical toxins tied to soaring obesity rates, heart disease and diabetes and limiting access to affordable, wholesome food to the country's poorest citizens.

Read More...

30 Comments

Occupy Wall Street Goes Home

Posted 1 day ago on Dec. 1, 2011, 3:04 p.m. EST by OccupyWallSt

On December 6th Occupy Wall Street will join in solidarity with a Brooklyn community to re-occupy a foreclosed home. The day of action marks a national kick-off for a new frontier for the

General Inquiries:
general@occupywallst.org
Press Inquiries:
press@occupywallst.org
Press Phone: +1 (347) 292-1444
Help & Directions: +1 (516) 708-4777
Watch: The world we're building
Read: This call to action
Liberty Square Eviction Defense: Text "@occupyalert" to 23559 to receive alerts in the event of imminent emergency.

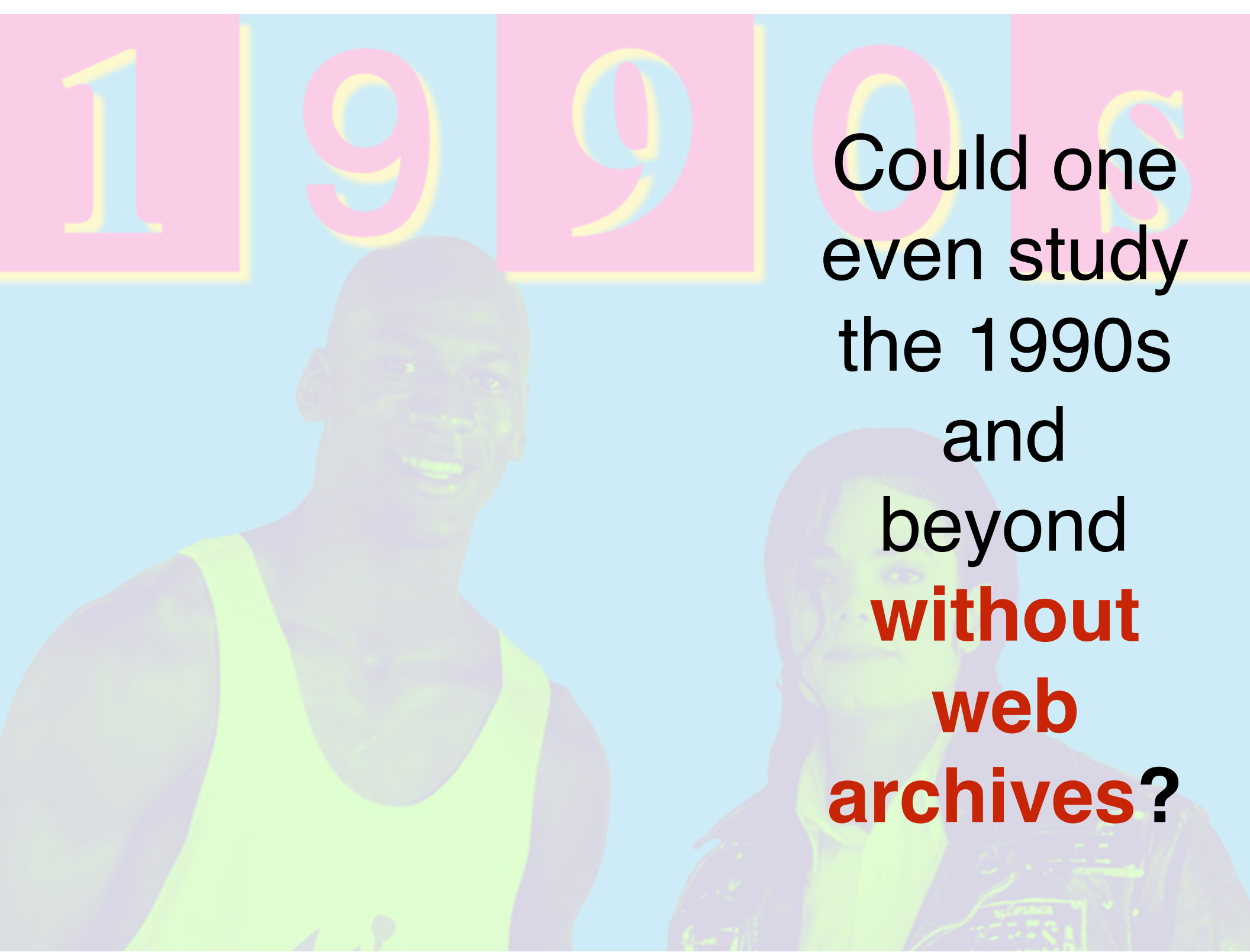
Occupy Wall Street is leaderless resistance movement with people of many colors, genders and political persuasions. The one thing we all have in common is that We Are The 99% that will no longer tolerate the greed and corruption of the 1%. We are using the revolutionary Arab Spring tactic to achieve our ends and encourage the use of nonviolence to maximize the safety of all participants.

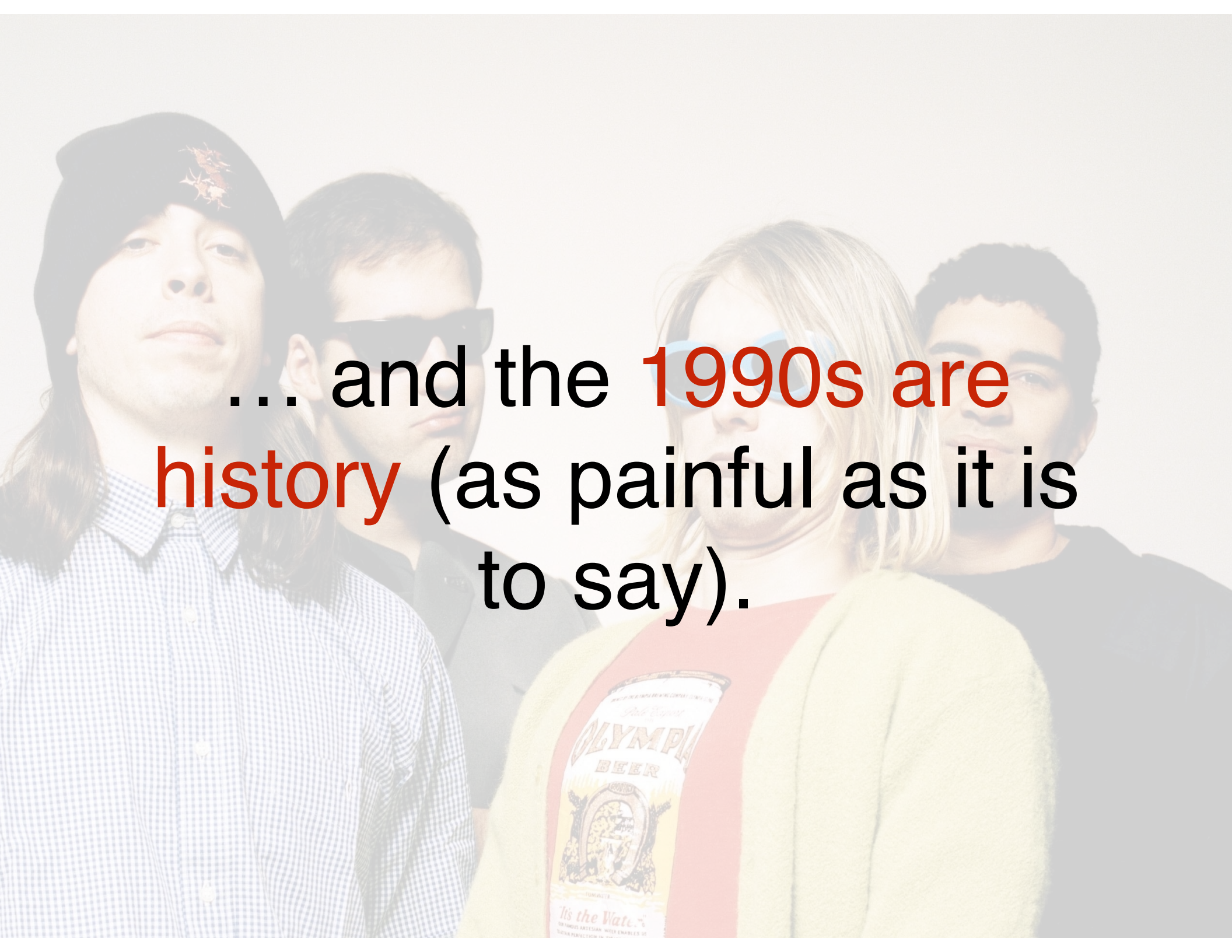
This #ows movement empowers real people to create real change from the bottom up. We want to see a general assembly in every backyard, on every street corner because we don't need Wall Street and we don't need politicians to build a better society.

the only solution is WorldRevolution

Click here for NYCGA committee meeting times.


~~Scarcity~~
Abundance

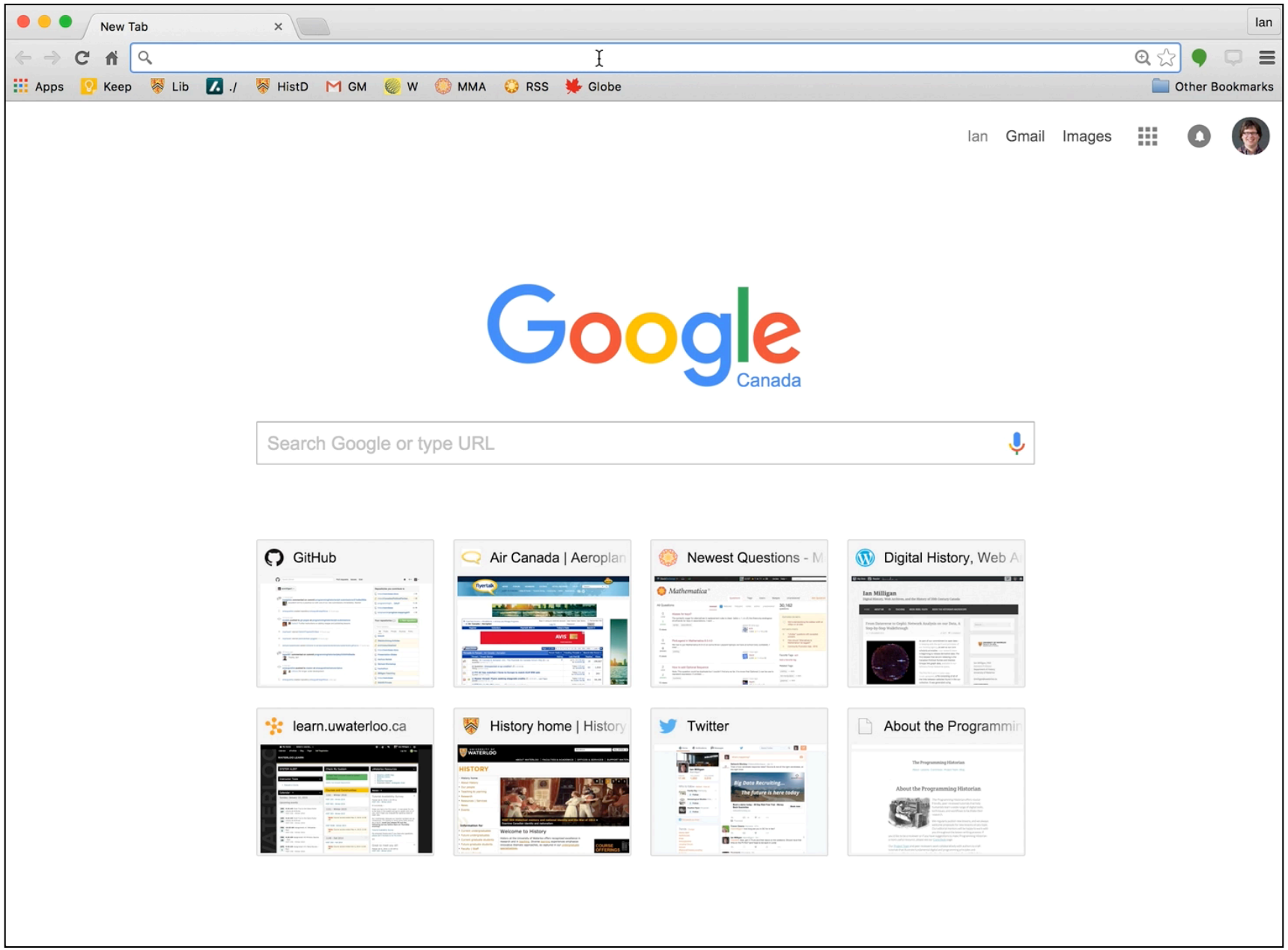

1990s
Could one even study the 1990s and beyond without web archives?



… and the 1990s are history (as painful as it is to say).

The decisions we make today will lay the foundations for how we work with born-digital cultural heritage.

So how do we currently use these resources?



This won't be enough!

… but how we will find the information we need?

Need to bring web archives into conversation with the digital humanist community (using standards-based tools).

Case study:
an Archive-It collection
(akin to special collections)

Pivotal Changes in Canadian Politics, 2005-2016

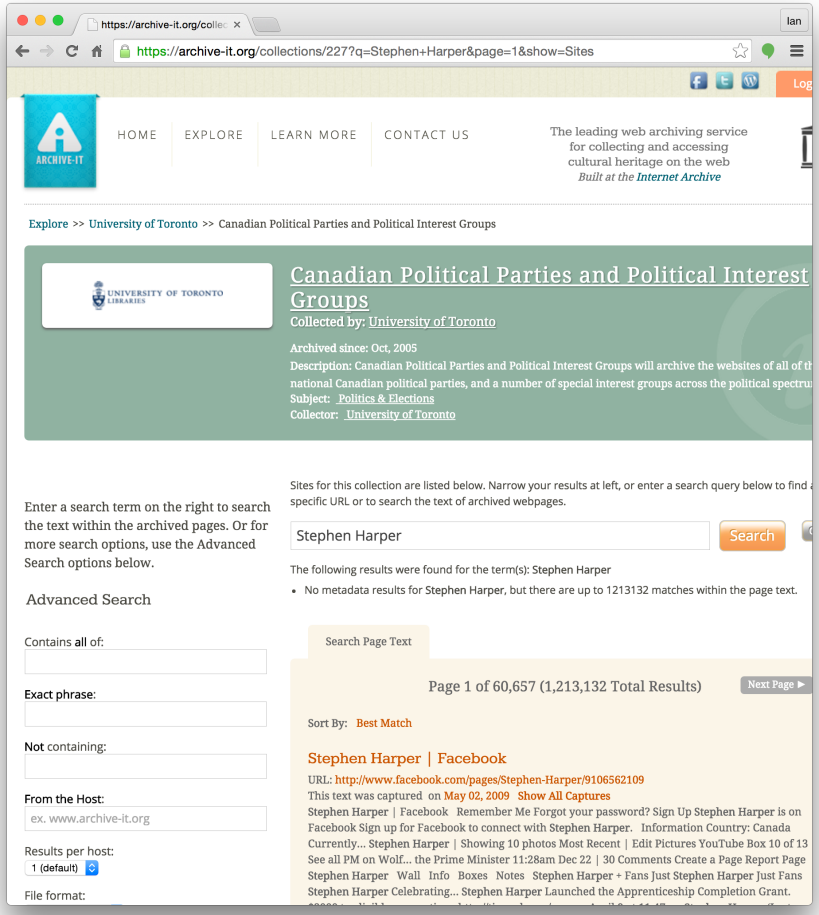
- Militarization of Canadian society?
- Change from ‘natural governing party’ of Liberals to Conservatives and back again
- Major policy changes on foreign policy, environment, etc.
- Sweeping change to how we understand our history
- How to measure?

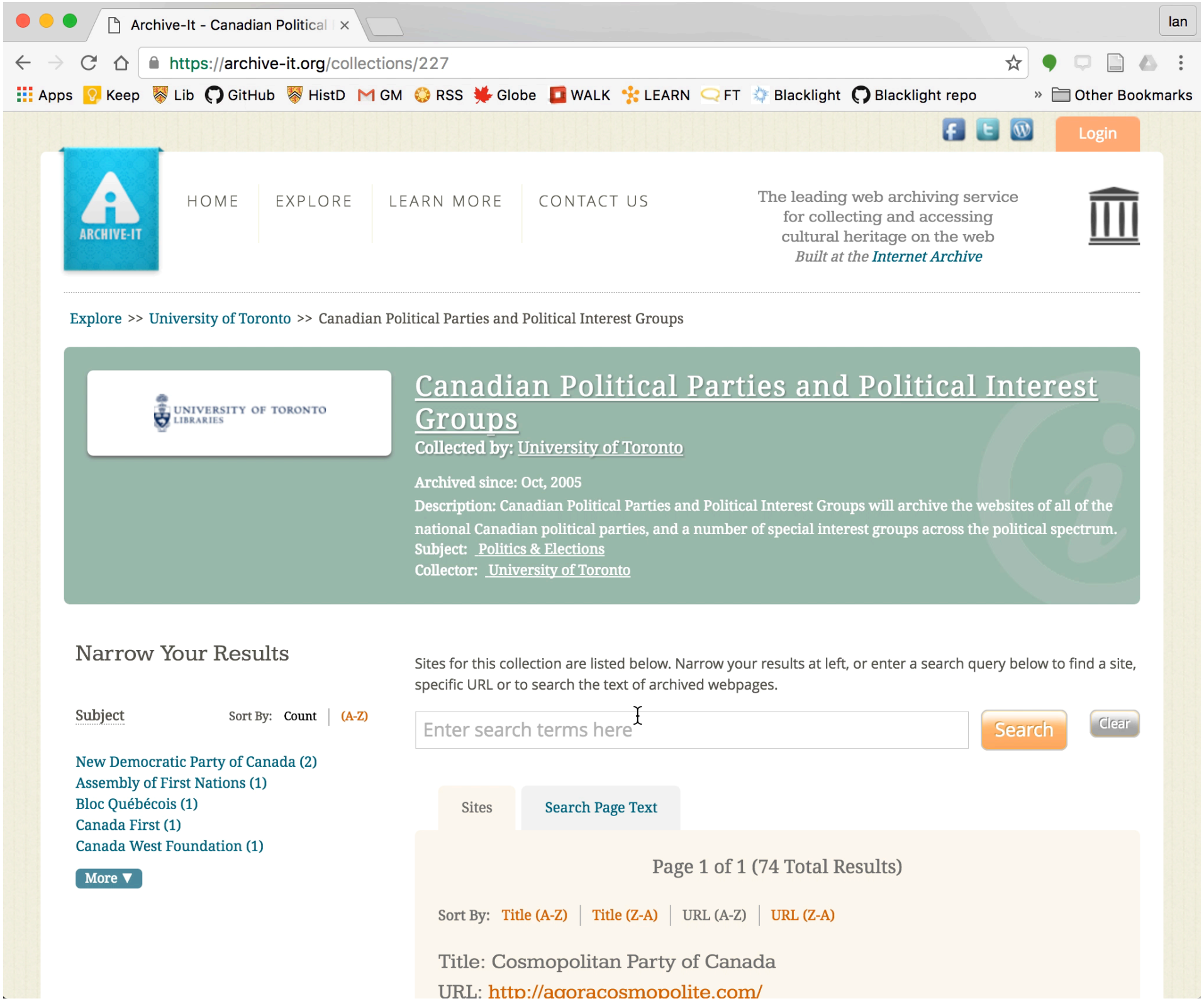
Canadian Political Parties & Political Interest Group Collection

- 50 Websites
  - All major political parties
  - Minor political parties
  - Political interest groups
- Collected quarterly between 2005 & present.

Current Interface

- Very limited - simple search engine, some advanced options; no facets
- Great collections.. but nobody uses them!

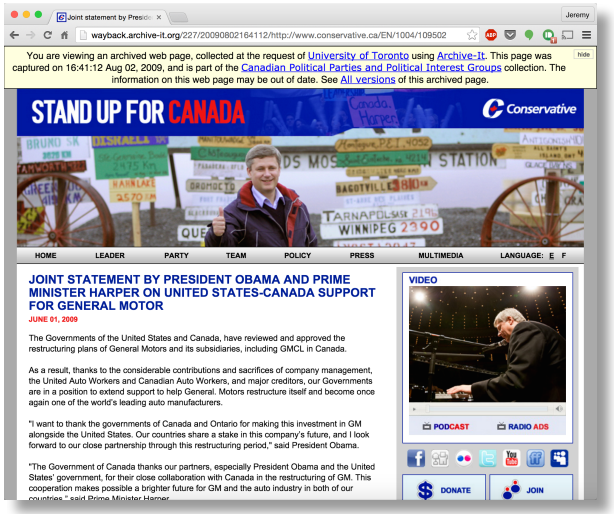



Archive-It - Canadian Political
Ian
https://archive-it.org/collections/227
Apps Keep Lib GitHub HistD GM RSS Globe WALK LEARN FT Blacklight Blacklight repo Other Bookmarks
Login
ARCHIVE-IT
HOME EXPLORE LEARN MORE CONTACT US
The leading web archiving service for collecting and accessing cultural heritage on the web
Built at the Internet Archive
Explore >> University of Toronto >> Canadian Political Parties and Political Interest Groups
UNIVERSITY OF TORONTO LIBRARIES
Canadian Political Parties and Political Interest Groups
Collected by: University of Toronto
Archived since: Oct, 2005
Description: Canadian Political Parties and Political Interest Groups will archive the websites of all of the national Canadian political parties, and a number of special interest groups across the political spectrum.
Subject: Politics & Elections
Collector: University of Toronto
Narrow Your Results
Subject 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 left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.
Enter search terms here
Search
Clear
Sites
Search Page Text
Page 1 of 1 (74 Total Results)
Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)
Title: Cosmopolitan Party of Canada
URL: http://agoracosmopolite.com/

Warcbase

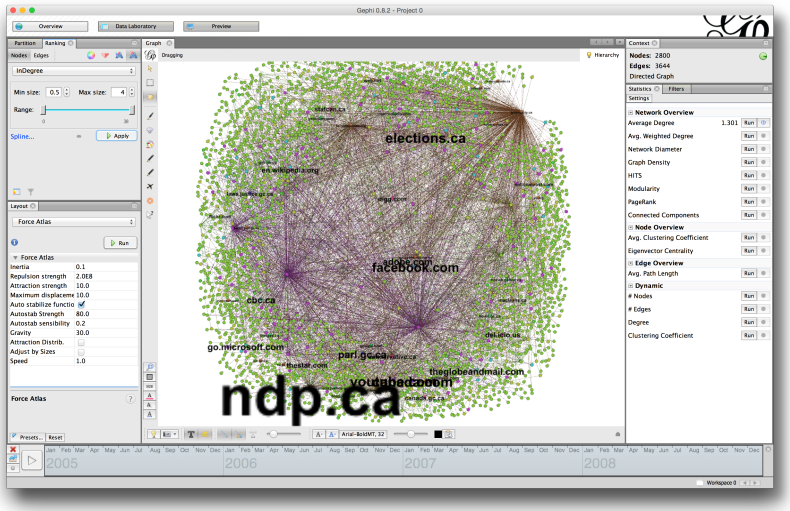
An open-source platform for managing web archives

http://warcbase.org



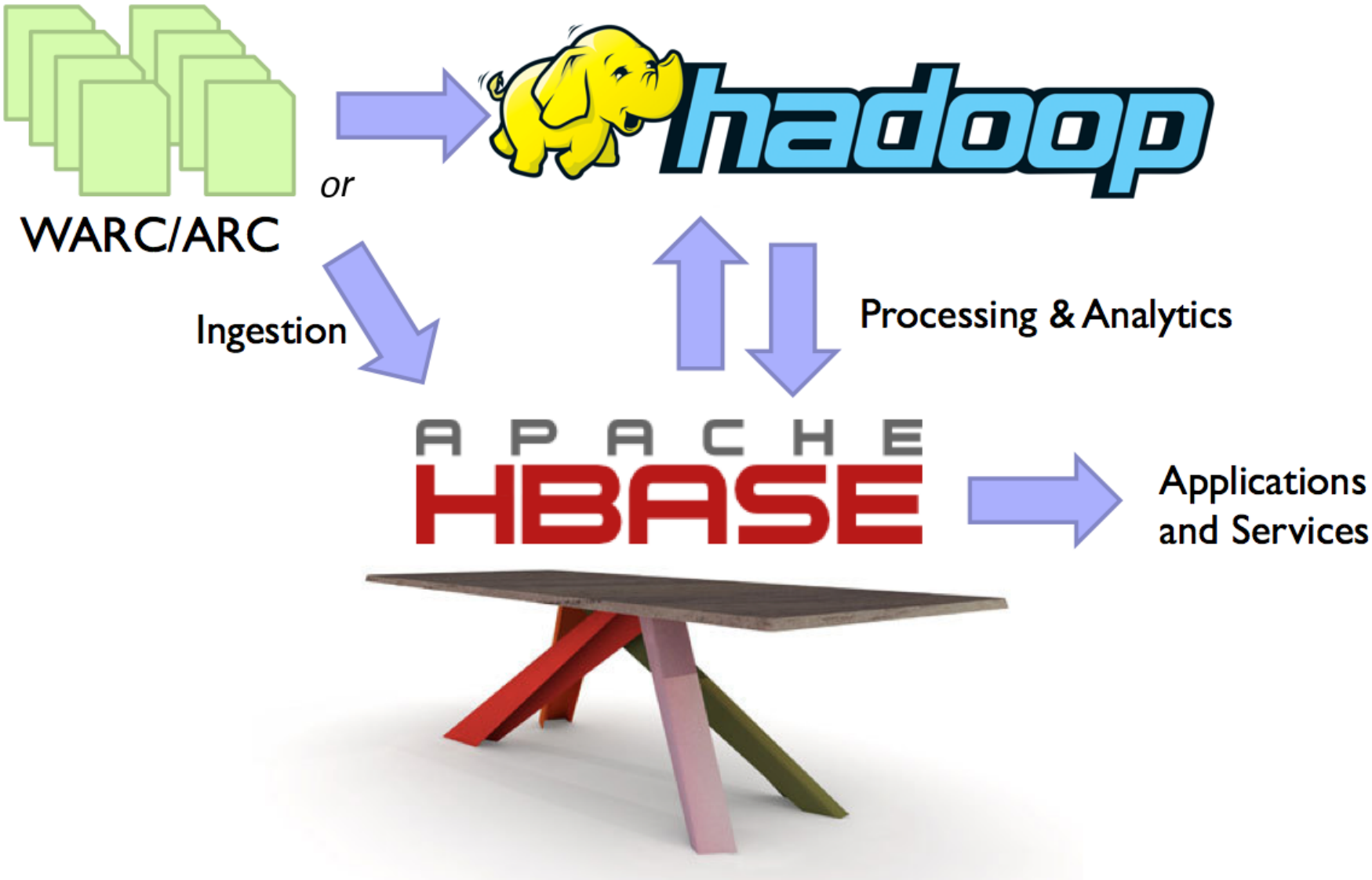
Two main facets

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



Funded by Mellon, SSHRC, NSERC, and Government of Ontario.

Warcbase


WARC/ARC
or
hadoop
Ingestion
Processing & Analytics
APACHE
HBASE
Applications
and Services

Warcbase

- **Scalable**
  - From Raspberry Pi, to laptop, to powerful desktop, to single-node beefy server, to cluster
- **Very powerful**
  - *Trantor cluster*: 1.2PB of disk, 25 compute nodes totalling 3.2TB memory and 300 current-generation Intel cores.

Analysis: Scripting

- Previously, Pig scripts ran Hadoop MapReduce jobs
  - Now depreciated (except for the DH 2016 abstract!)
- Transitioned to Spark, scripts written in Scala
  - Better performance
  - API and libraries for Python (*PySpark*) – hope to facilitate adoption

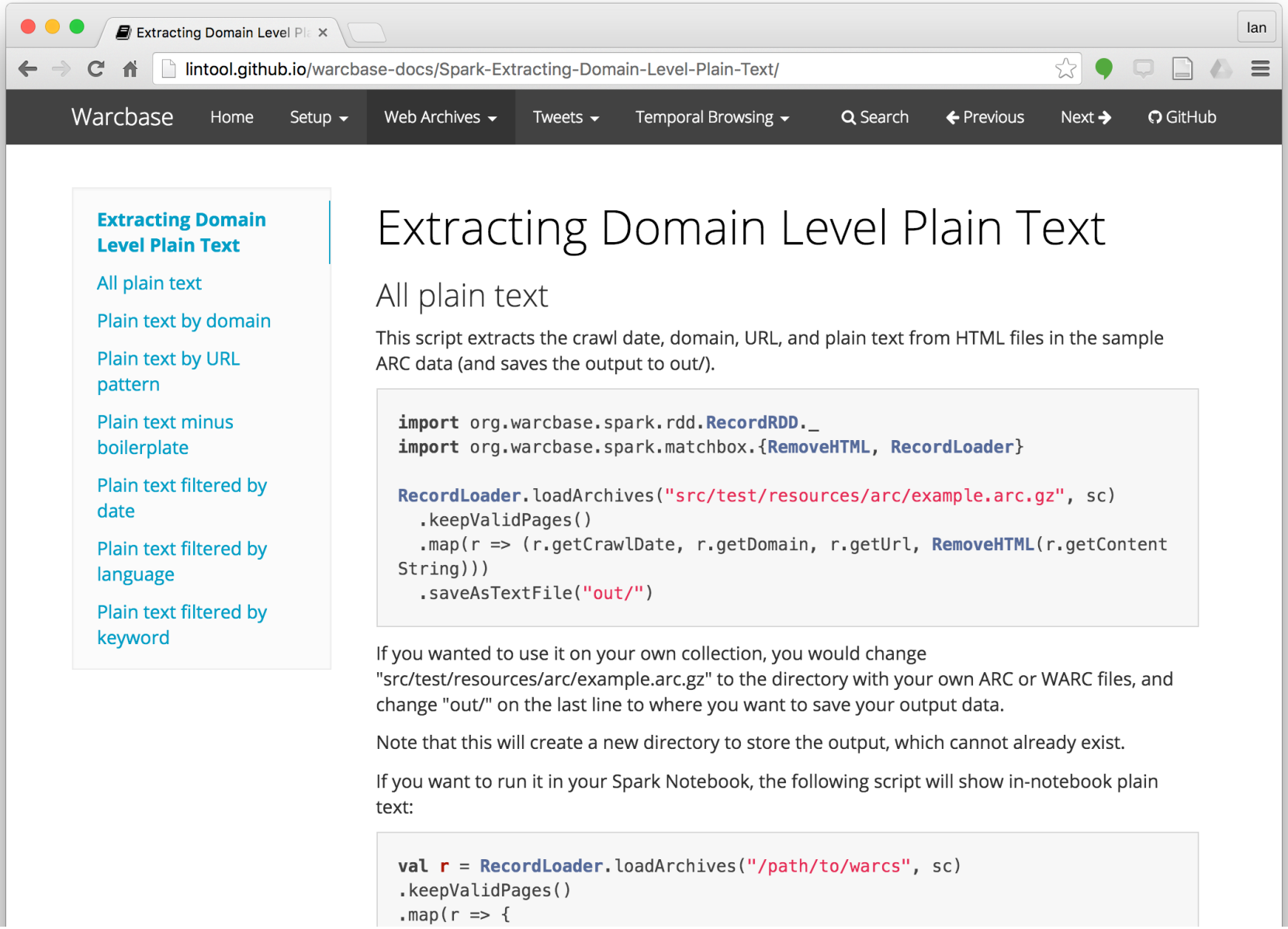
A Simple Script

Domain level plain text extraction (Spark)

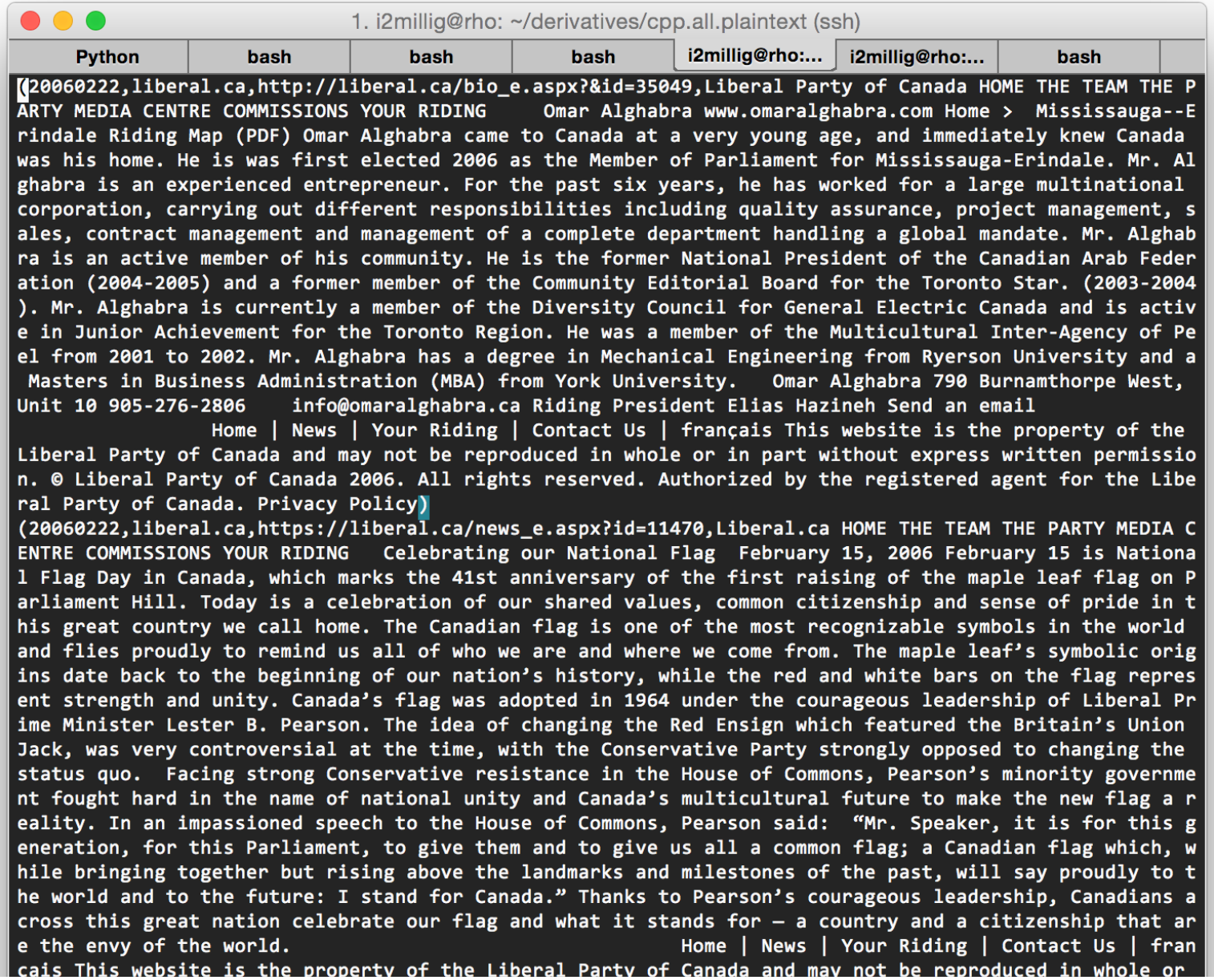
```
import org.warcbase.spark.matchbox.{RemoveHTML, RecordLoader}
import org.warcbase.spark.rdd.RecordRDD._

RecordLoader.loadArchives("src/test/resources/arc/example.arc.gz", sc)
  .keepValidPages()
  .keepDomains(Set("greenparty.ca"))
  .map(r => (r.getCrawlDate, r.getDomain, r.getUrl, RemoveHTML(r.getContent
String)))
  .saveAsTextFile("out/")
```

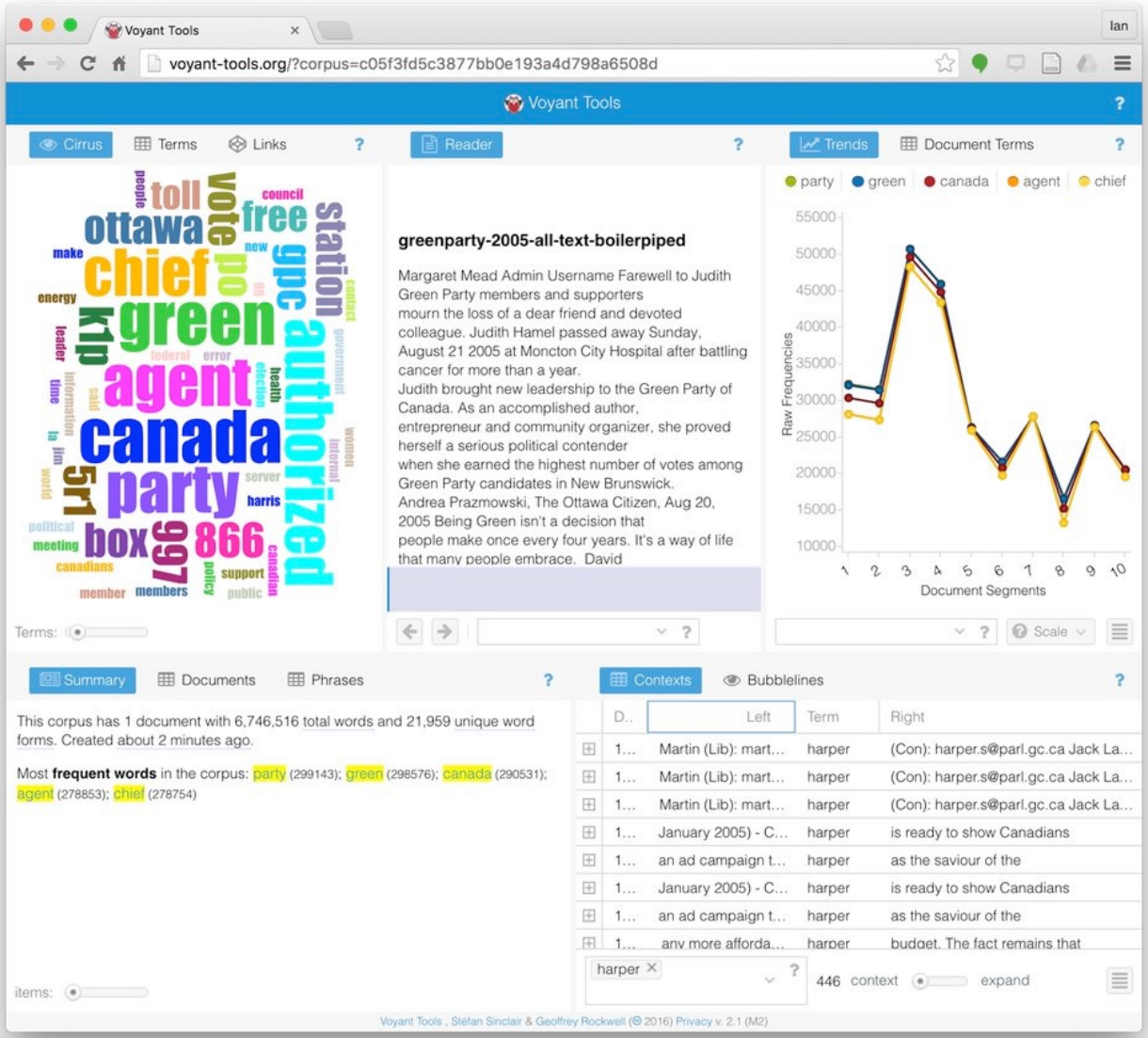
docs.warcbase.org



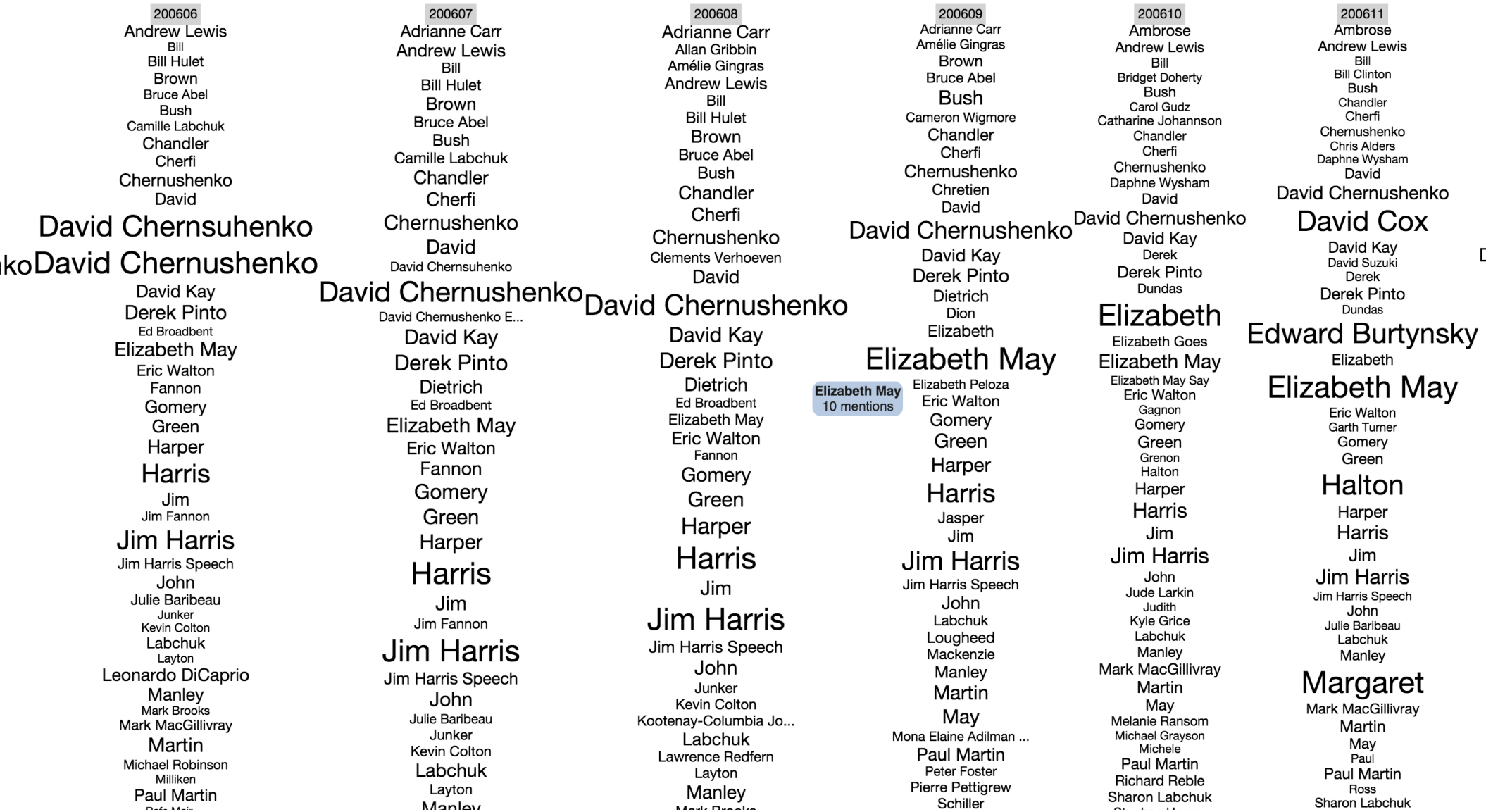
Extract all Text



Extract all Text (voyant connector)

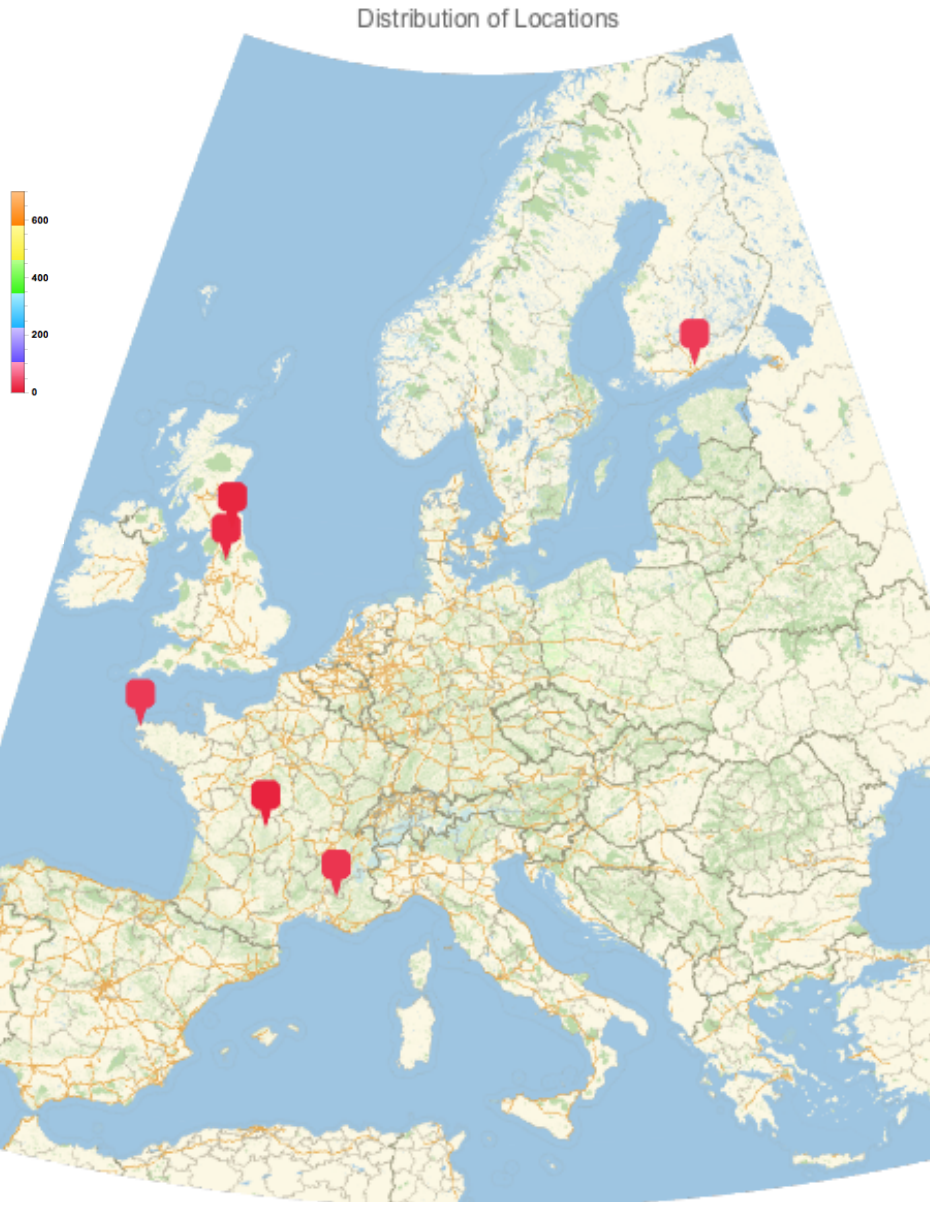


Extract Entities



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,
```

And a move away from content and towards structured metadata

An Example

Imagine one e-mail

Hi Tony –

See you after class?

Ian

Tells you nothing!

But what if I e-mailed him every Friday? Or every day?

immersion.media.mit.edu

log out *& save* • log out *& delete*

Lookup Contacts

Charge

Nodes [A] [S]

Links [Q] W]

Take a snapshot
- with labels
- without labels

Feedback?
Compose

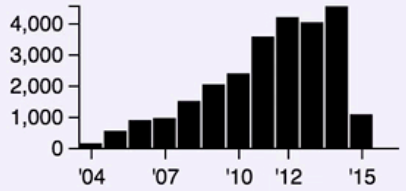
Ian Milligan

729 collaborators

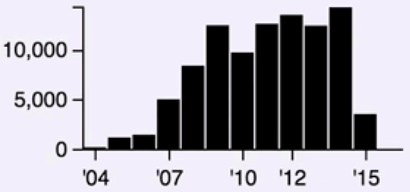
121,854 emails

My Stats | Top Collaborators

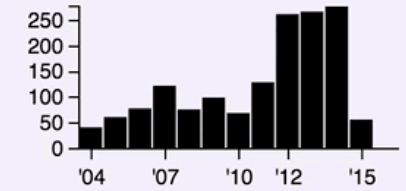
Emails Sent

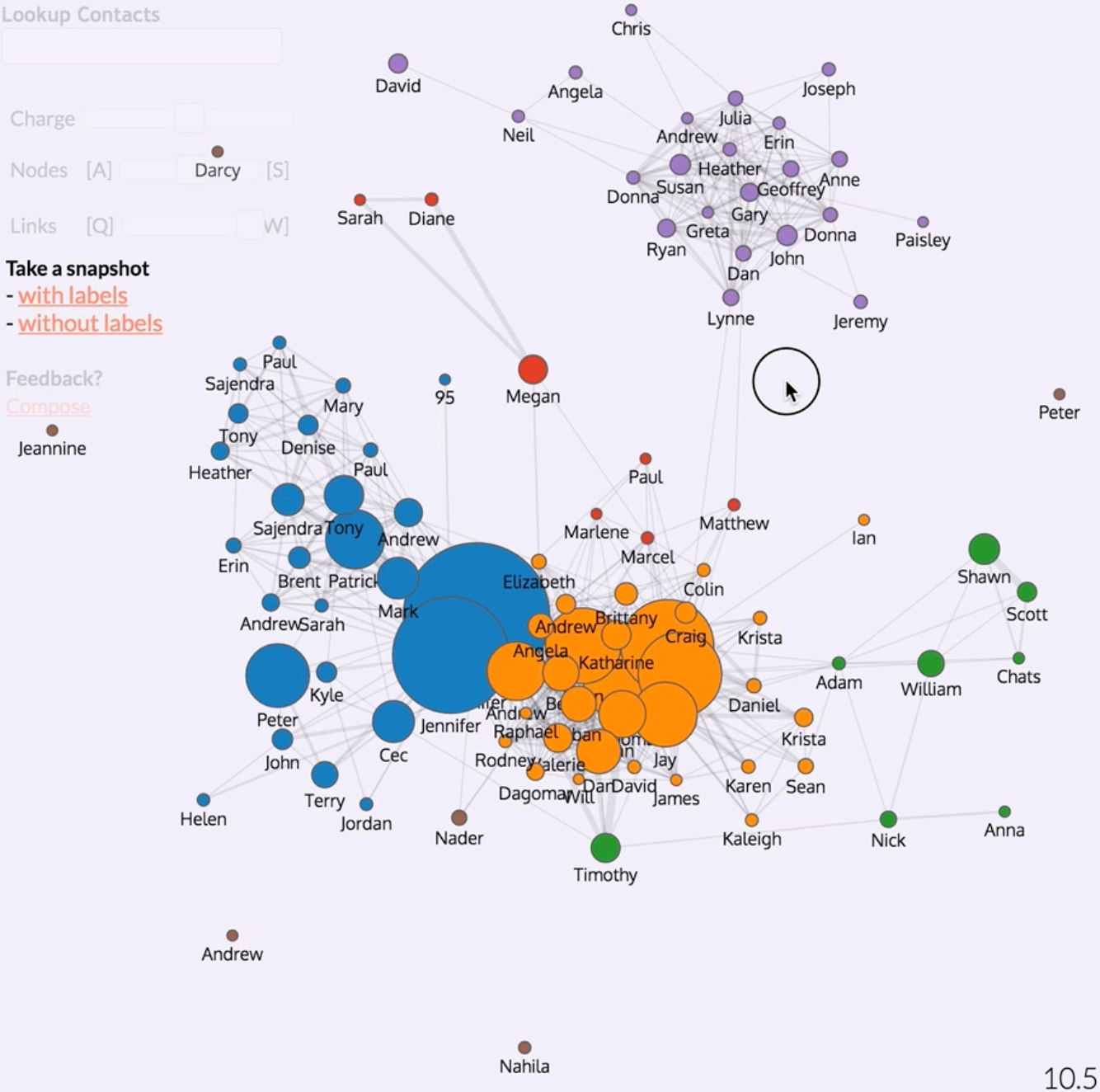


Emails Received



New Collaborators



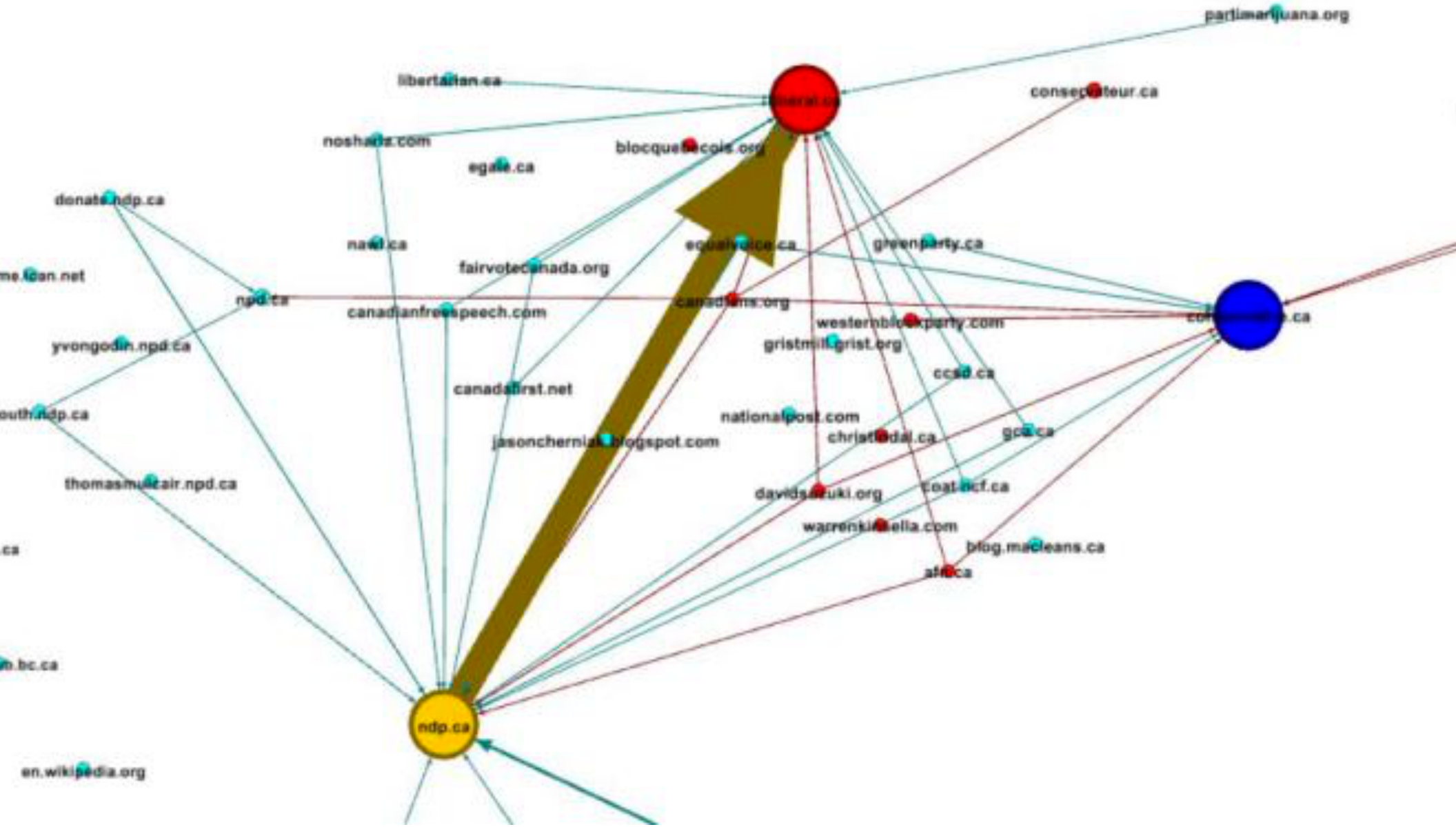


10.5 years

26 Sep 2004 - 12 Mar 2015

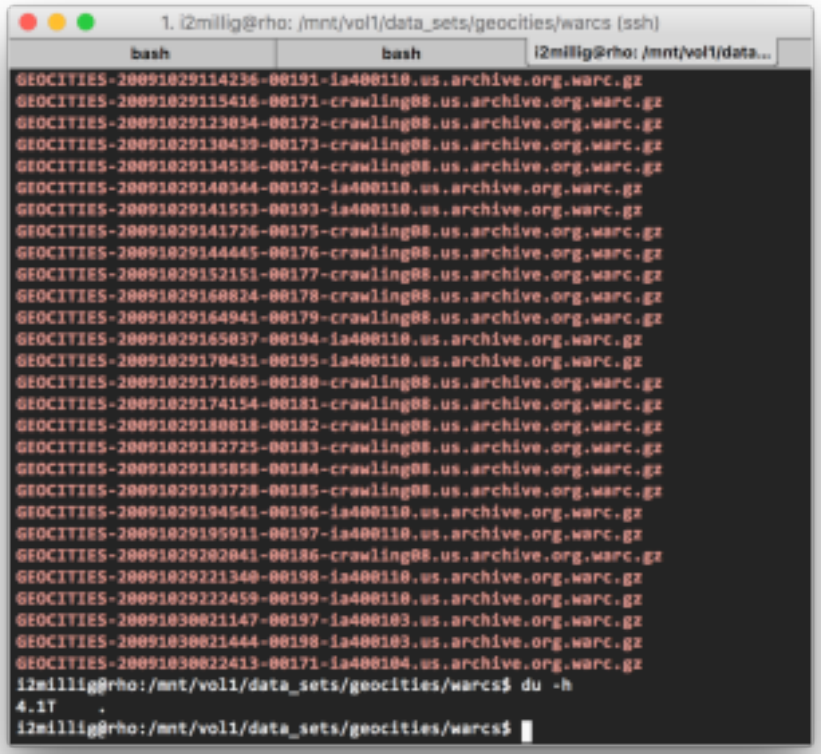
All • Past Year • Past Month • Past Week

2005 Canadian Federal Election

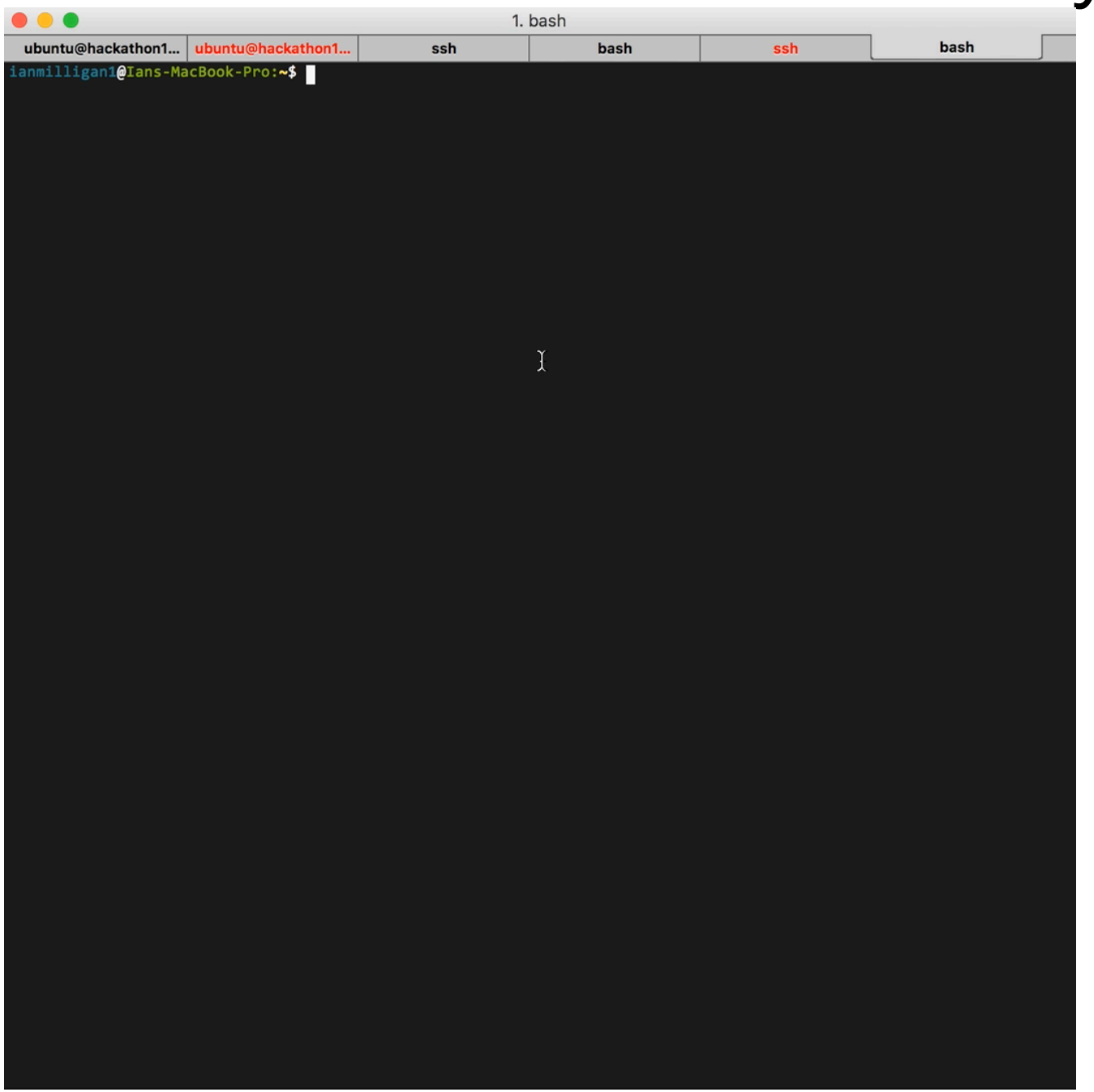


So how do you use warcbase in your own work?

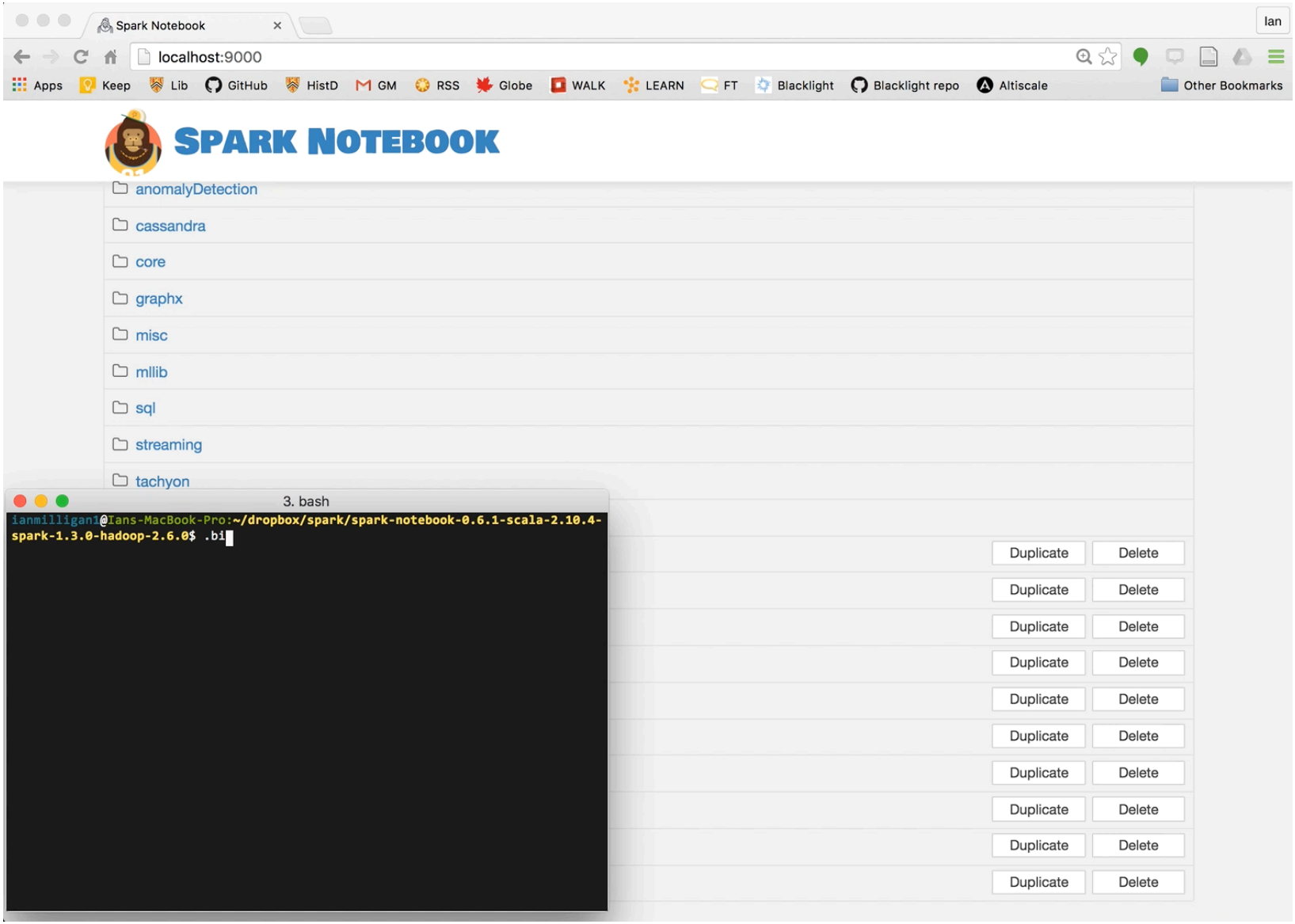
Step One: Ingest data



Step Two: Basic Shell Analysis



Step Two: Basic Analytics



Step Three: Filtering a Corpus

```
import org.warcbase.spark.matchbox.{ExtractDomain, ExtractLinks, RecordLoad
er}
import org.warcbase.spark.rdd.RecordRDD._

RecordLoader.loadArchives("/path/to/arc", sc)
  .keepValidPages()
  .map(r => (r.getCrawlDate, ExtractLinks(r.getUrl, r.getContentString)))
  .flatMap(r => r._2.map(f => (r._1, ExtractDomain(f._1).replaceAll("^\\s*w
ww\\.", ""), ExtractDomain(f._2).replaceAll("^\\s*www\\.", ""))))
  .filter(r => r._2 != "" && r._3 != "")
  .countItems()
  .filter(r => r._2 > 5)
  .saveAsTextFile("cpp.sitelinks")
```

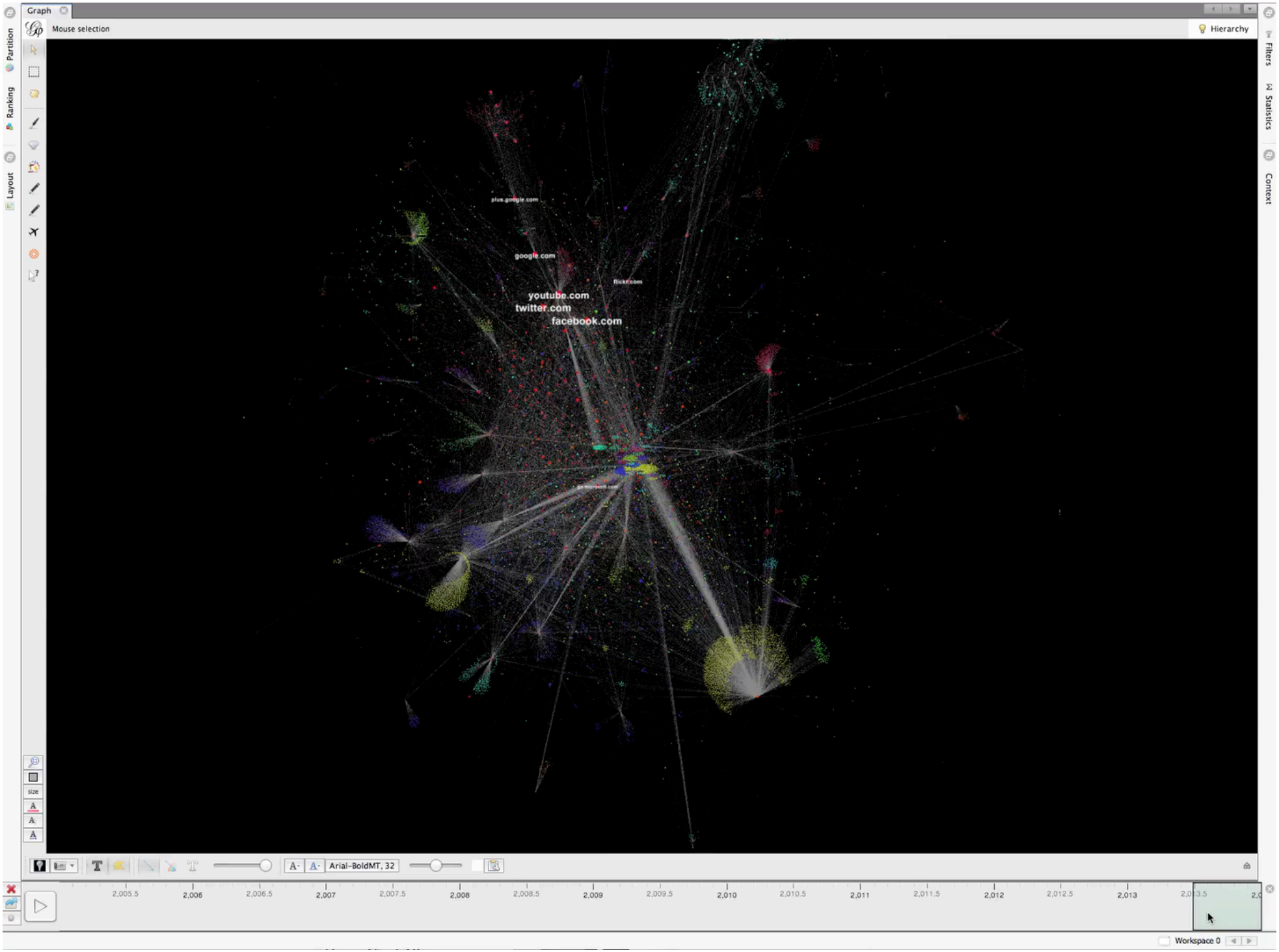
A Link Graph

Step Three: Filtering a Corpus

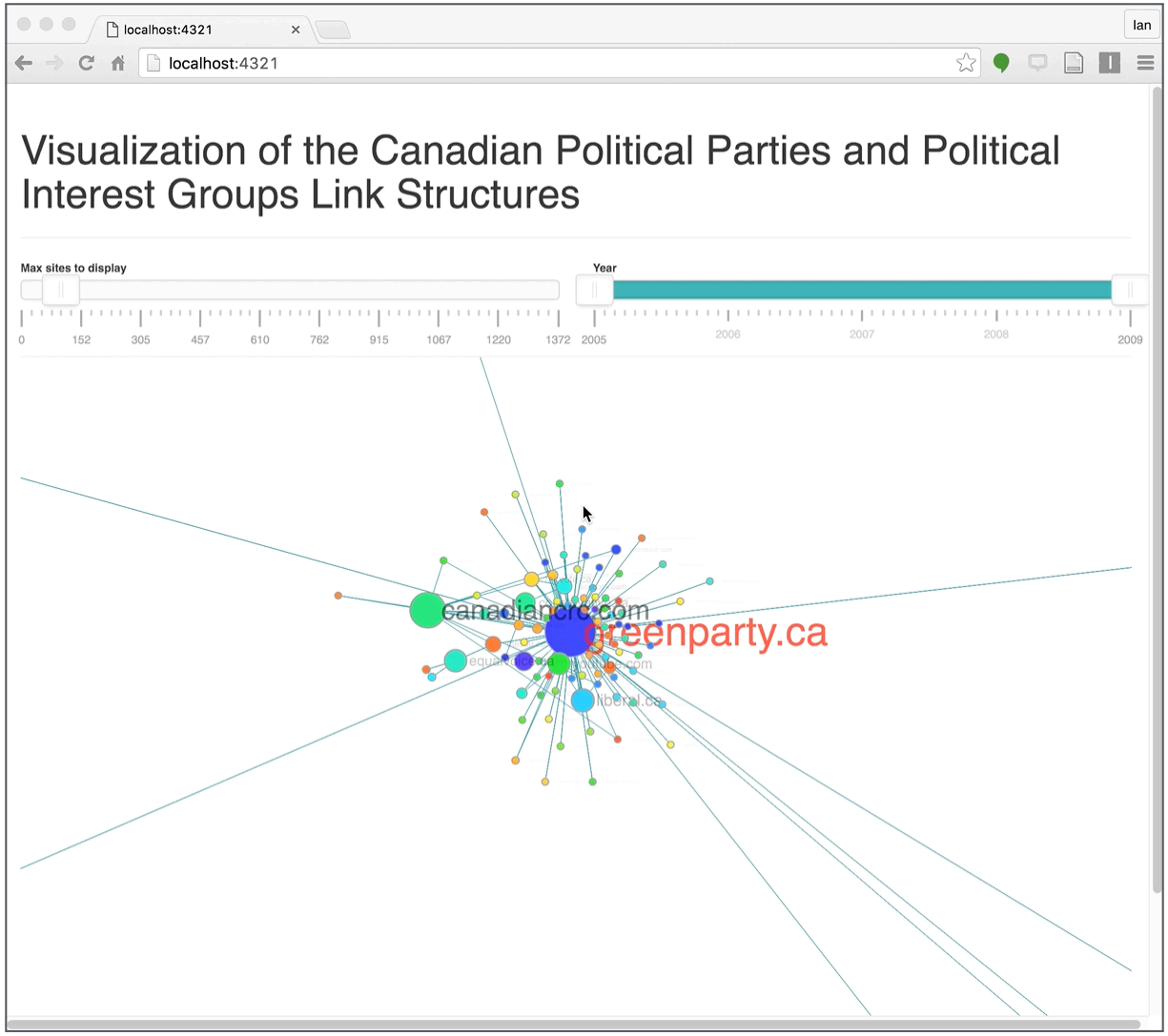
```
((20080612,liberal.ca,liberal.ca),1832983)
((20060326,ndp.ca,ndp.ca),1801775)
((20060426,ndp.ca,ndp.ca),1771993)
((20060325,policyalternatives.ca,policyalternatives.ca),1735154)
```

Results

Step Four: Visualize to select sub-corpora


Graph
Mouse selection
Hierarchy
Partition
Ranking
Layout
Filters
Statistics
Context
plus.google.com
google.com
flickr.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

In-browser test visualizations



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!

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

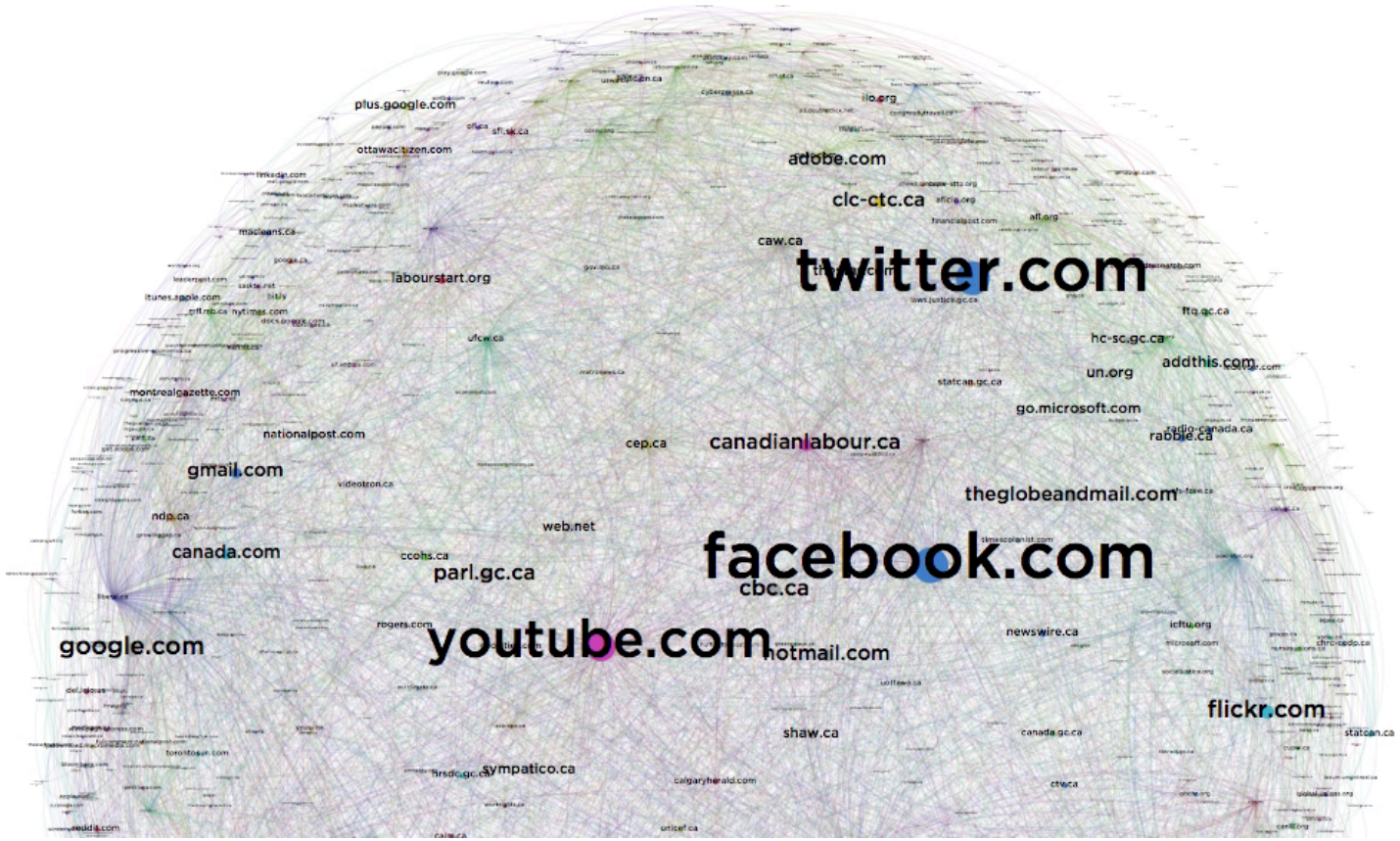
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

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.



(for more on Shine/ WebArchives.ca/etc., come to our panel in CC at 11:30)

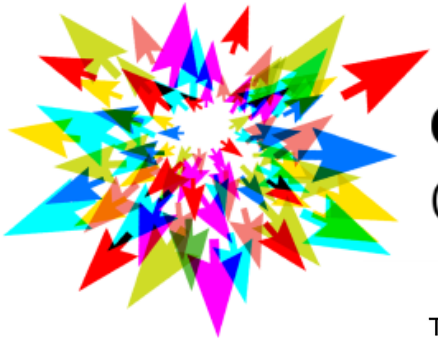
Warcbase = A general purpose platform for unlocking web archives.

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Ontario



compute canada | calcul canada

UNIVERSITY OF WATERLOO

THE ANDREW W. MELLON FOUNDATION

NSERC CRSNG

Thanks/Questions!