

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
SERIES I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

370

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL
SERIES I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

371

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
SERIES I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

372

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
SERIES I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

373

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series II
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

374

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series II
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Subseries F Democratic Party
BOX 30

JOHN J. BURNS LIBRARY
BOSTON COLLEGE

375

Scarcity

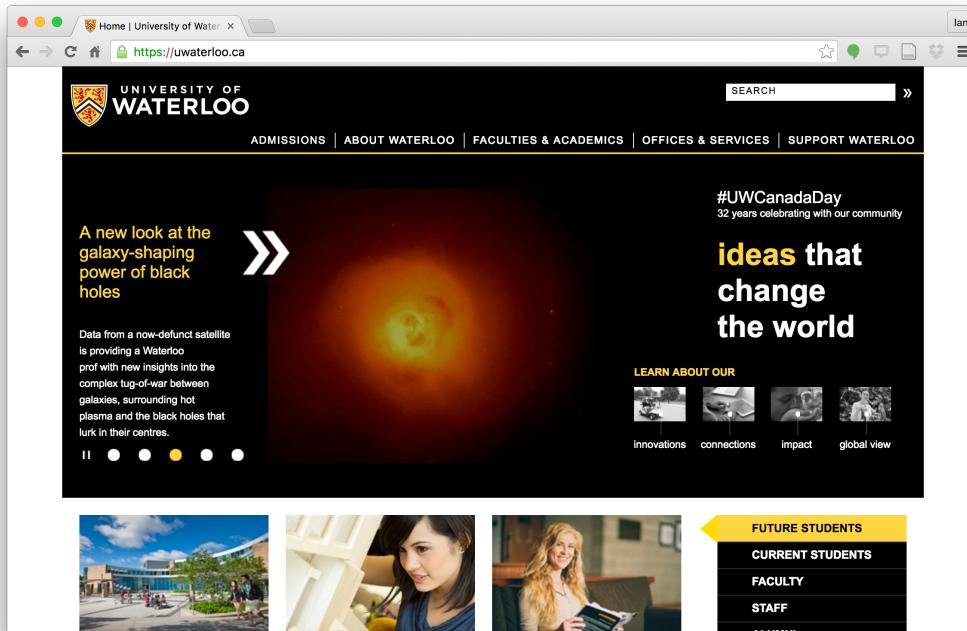




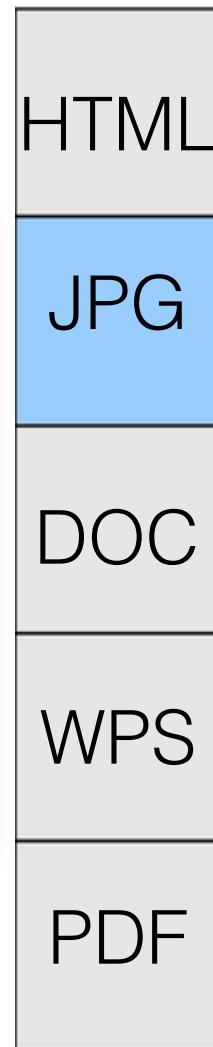
WebARChive (WARC) File
ISO 28500:2009

????

- Imagine your university's website:



WARC file



WARC record

Text header

Content block

[image/jpeg binary data]

WARC/1.0

WARC-Type: resource

WARC-Target-URI: file:/var/www/htdocs/images/logo.jpg

WARC-Date: 2006-09-19T17:20:24Z

WARC-Record-ID: <urn:uuid:92283950-ef2f-4d72-b224-f54c6ec9

Content-Type: image/jpeg

WARC-Payload-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZE

WARC-Block-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENM

Content-Length: 1662

...etc.

6 captures
9 Nov 96 - 25 Jun 04

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

Go

OCT

NOV
9
1996APR
2002Close
Help ?

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

You are viewing an archived web page, collected at the request of [Internet Archive Global Events](#) using [Archive-It](#). This page was captured on 5:07:27 Dec 03, 2011, and is part of the [Occupy Movement 2011/2012](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page. [Videos Metadata](#)

Welcome [login](#) | [signup](#)
Language [en](#) [es](#) [fr](#)

OccupyWallStreet

The revolution continues [worldwide!](#)

[News](#) [LiveStream](#) [#HowToOccupy](#) [Forum](#) [Chat](#) [User Map](#) [NYCGA](#) [About](#) [Donate](#)

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 [occupy movement: the liberation of vacant bank owned homes for those in need](#). The banks are

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.

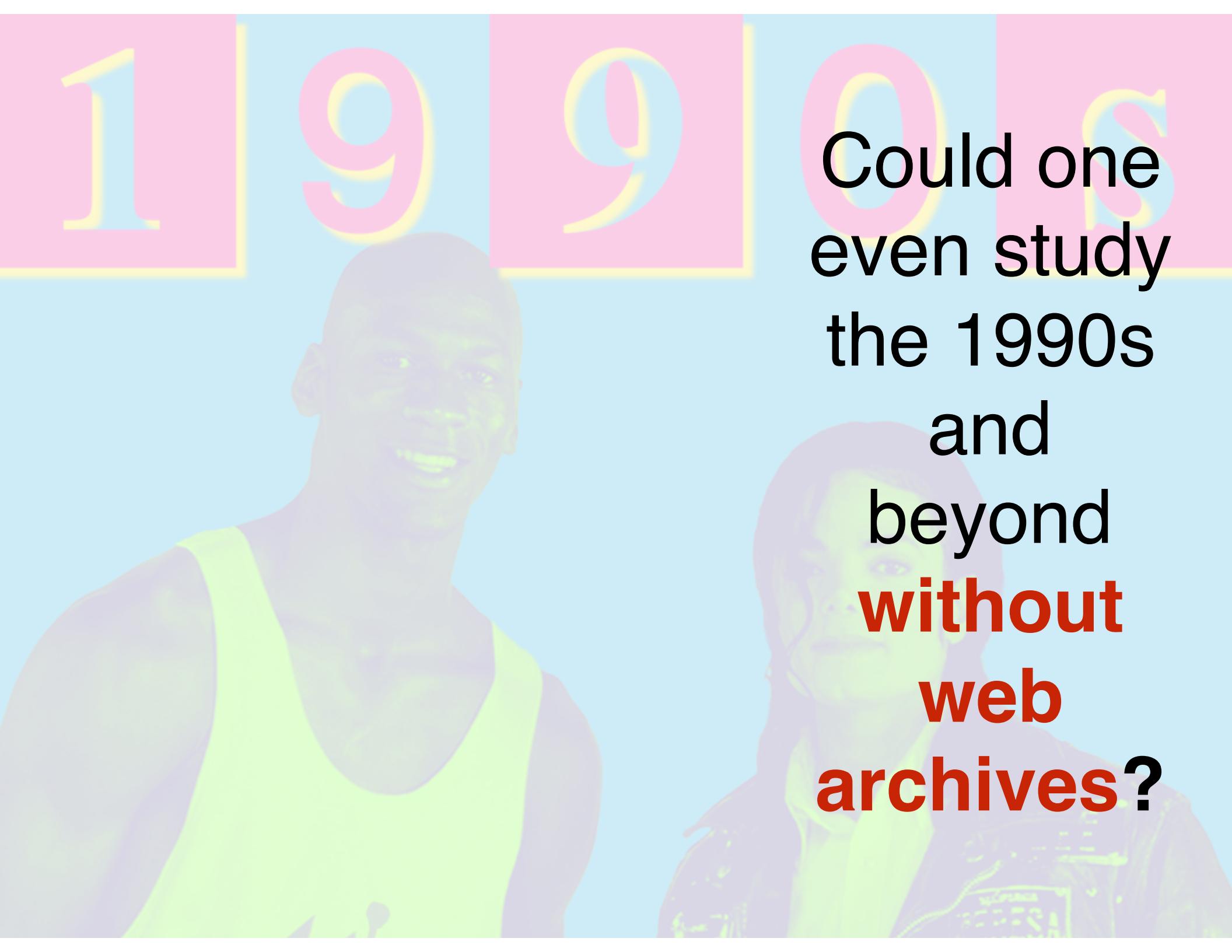
This #ows movement empowers real people to create real change from the bottom up. We want to see a [general assembly](#) in every backyard, on every street corner because we don't need Wall Street and we don't need politicians to build a better society.

the only solution is WorldRevolution

[Click here](#) for NYCGA committee meeting times.

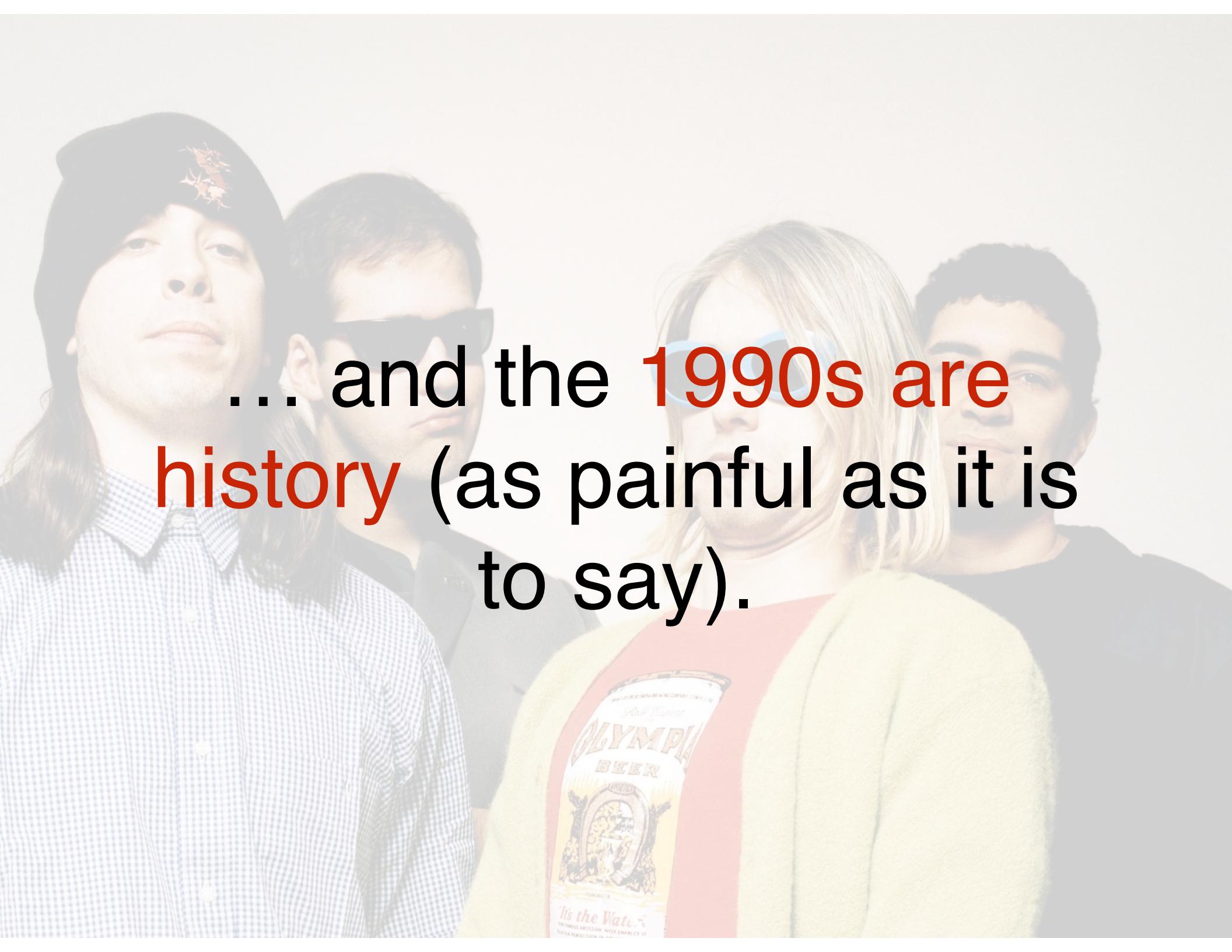
Scarcity Abundance





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

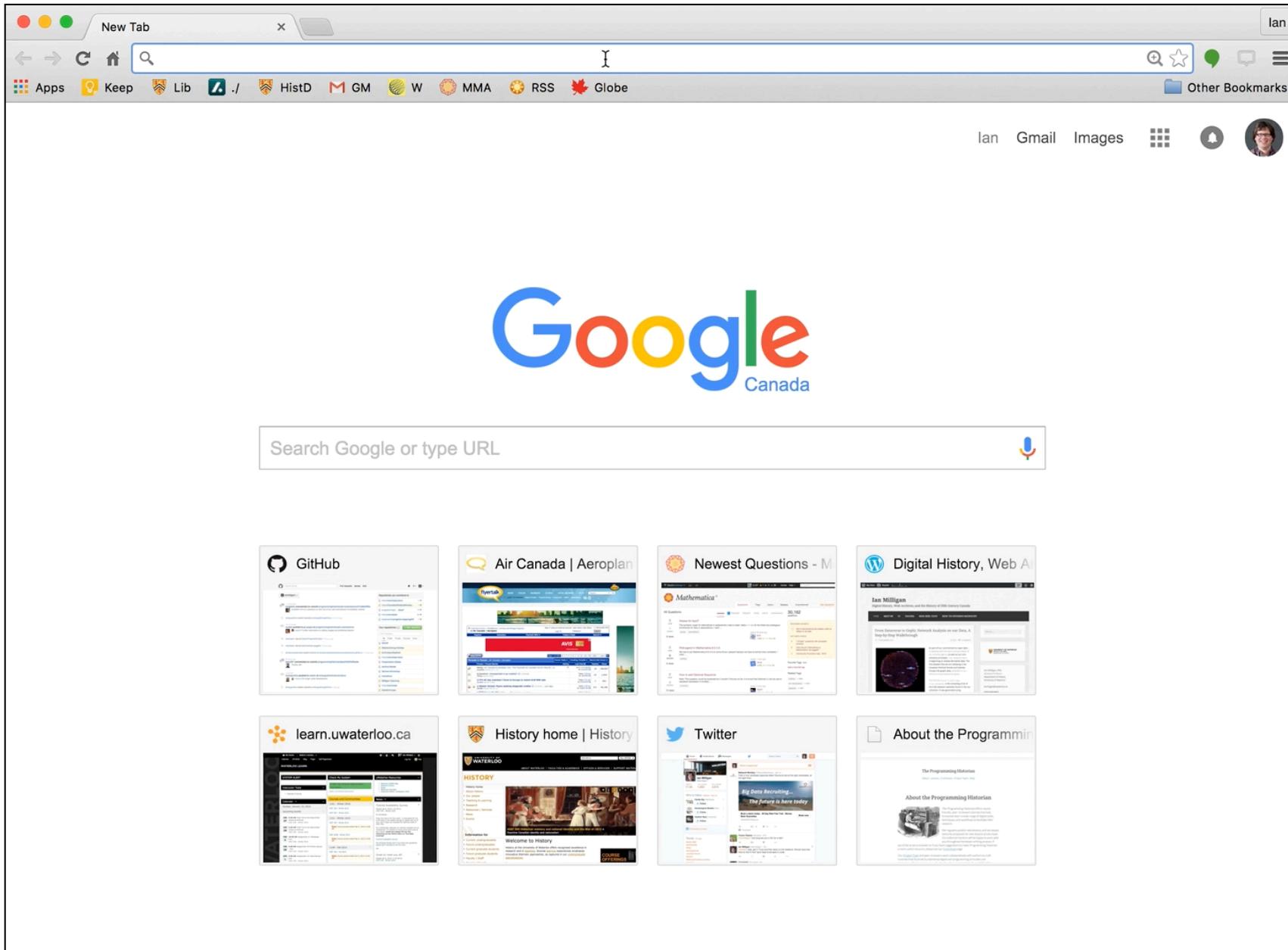
1990s



... 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!**

The screenshot shows a web browser displaying the Archive-It collection page for "Canadian Political Parties and Political Interest Groups". The URL in the address bar is <https://archive-it.org/collections/227?q=Stephen+Harper&page=1&show=Sites>. The page features the Archive-It logo and navigation links for HOME, EXPLORE, LEARN MORE, and CONTACT US. A banner at the top right states: "The leading web archiving service for collecting and accessing cultural heritage on the web. Built at the Internet Archive". Below the banner, the collection title "Canadian Political Parties and Political Interest Groups" is displayed, along with the collector information "Collected by: University of Toronto", the archival date "Archived since: Oct, 2005", and a brief description: "Canadian Political Parties and Political Interest Groups will archive the websites of all the national Canadian political parties, and a number of special interest groups across the political spectrum". A search bar at the bottom left allows users to enter a search term, and a search button is visible. The main content area shows a list of results for the query "Stephen Harper", with a total of 60,657 results. The results are sorted by "Best Match". Each result entry includes the title, URL, and a snippet of the archived page content.



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



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://acoracosmopolite.com/>

Warcbase

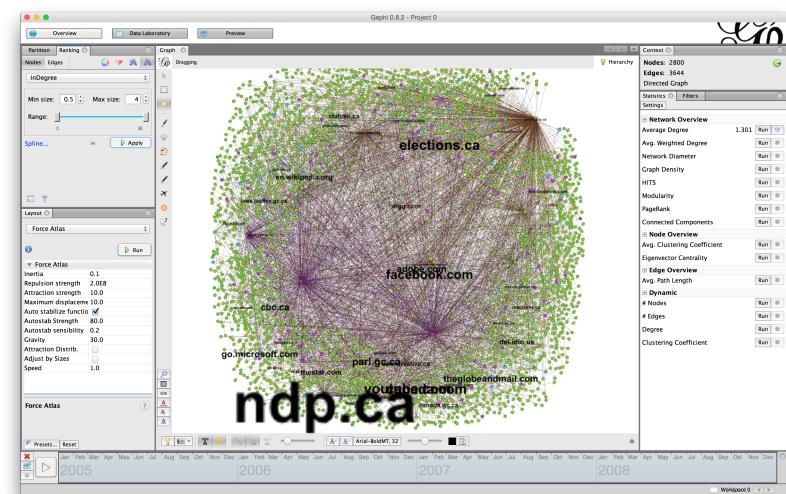
An open-source platform for managing web archives

<http://warcbase.org>

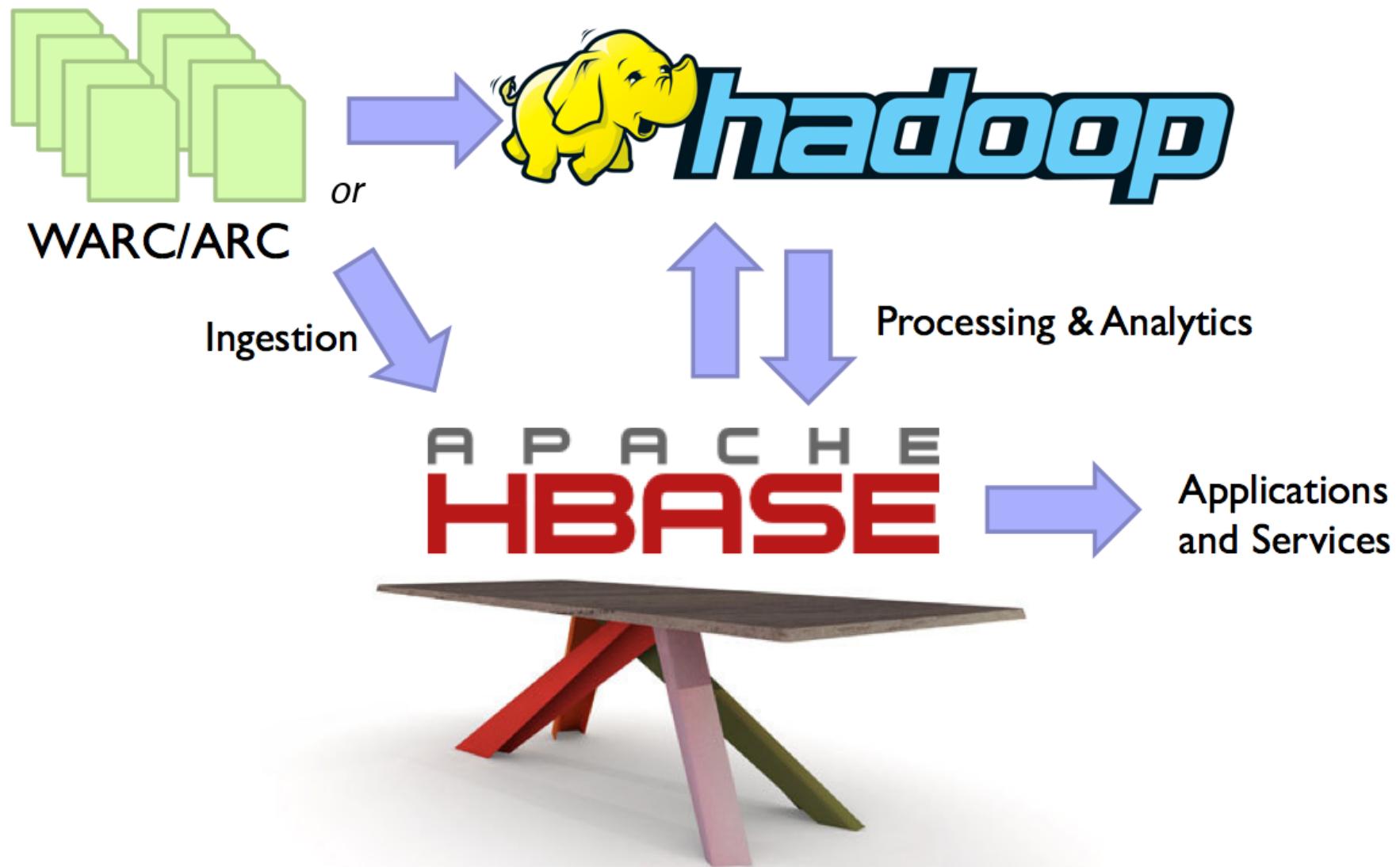
Two main facets

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

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



Warcbase



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 deprecated (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

The screenshot shows a web browser window with the title bar "Extracting Domain Level Plain Text". The address bar contains the URL "lintool.github.io/warcbase-docs/Spark-Extracting-Domain-Level-Plain-Text/". The page header includes links for "Warcbase", "Home", "Setup", "Web Archives", "Tweets", "Temporal Browsing", "Search", "Previous", "Next", and "GitHub". A sidebar on the left lists several options under the heading "Extracting Domain Level Plain Text": "All plain text", "Plain text by domain", "Plain text by URL pattern", "Plain text minus boilerplate", "Plain text filtered by date", "Plain text filtered by language", and "Plain text filtered by keyword". The main content area features a large heading "Extracting Domain Level Plain Text" and a sub-section "All plain text". It describes a script that extracts crawl date, domain, URL, and plain text from HTML files in sample ARC data. Below this is a code block:

```
import org.warcbase.spark.rdd.RecordRDD._  
import org.warcbase.spark.matchbox.{RemoveHTML, RecordLoader}  
  
RecordLoader.loadArchives("src/test/resources/arc/example.arc.gz", sc)  
  .keepValidPages()  
  .map(r => (r.getCrawlDate, r.getDomain, r.getUrl, RemoveHTML(r.getContent  
String)))  
  .saveAsTextFile("out/")
```

If you wanted to use it on your own collection, you would change "src/test/resources/arc/example.arc.gz" to the directory with your own ARC or WARC files, and change "out/" on the last line to where you want to save your output data.

Note that this will create a new directory to store the output, which cannot already exist.

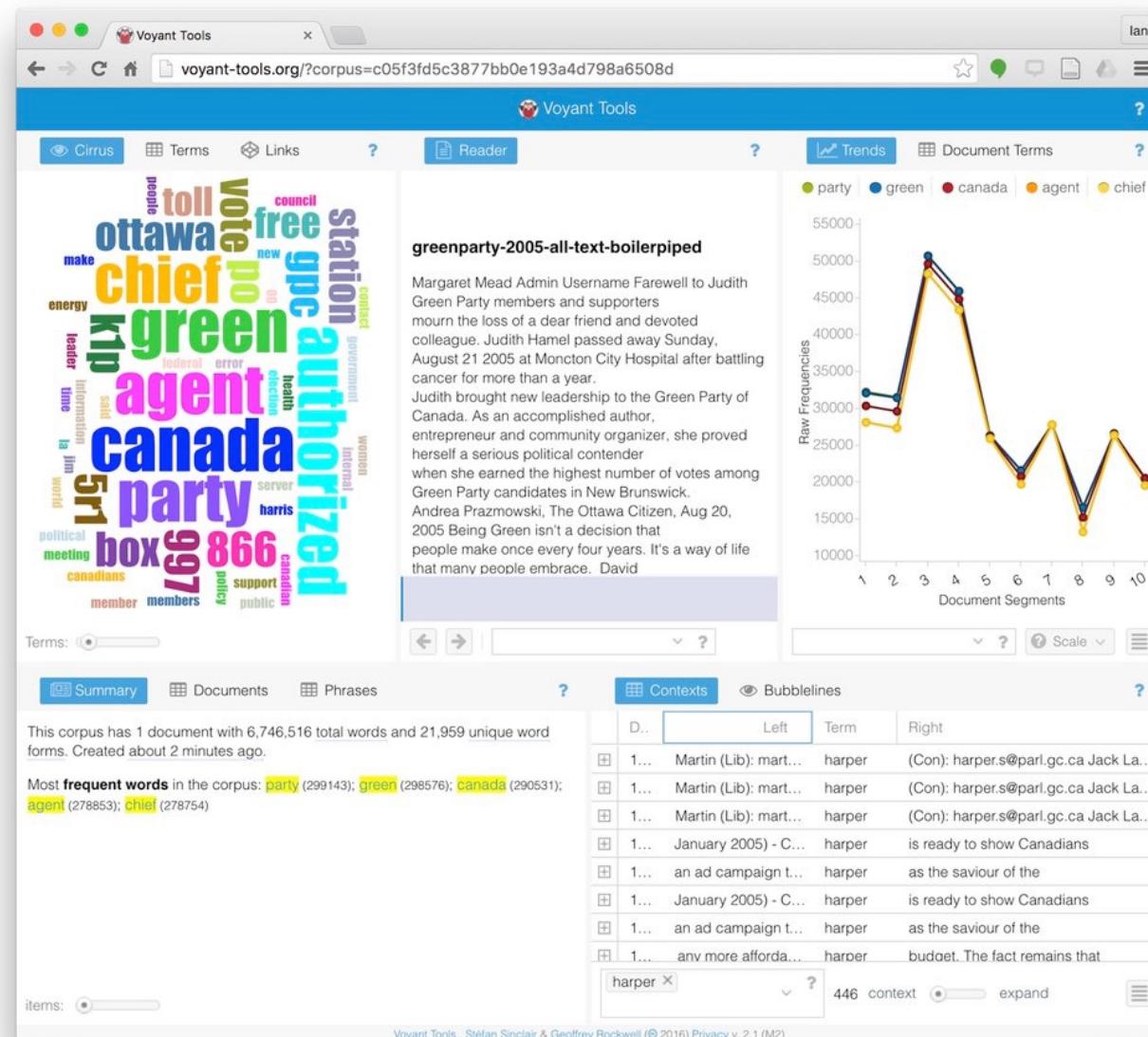
If you want to run it in your Spark Notebook, the following script will show in-notebook plain text:

```
val r = RecordLoader.loadArchives("/path/to/warcs", sc)  
.keepValidPages()  
.map(r => {
```

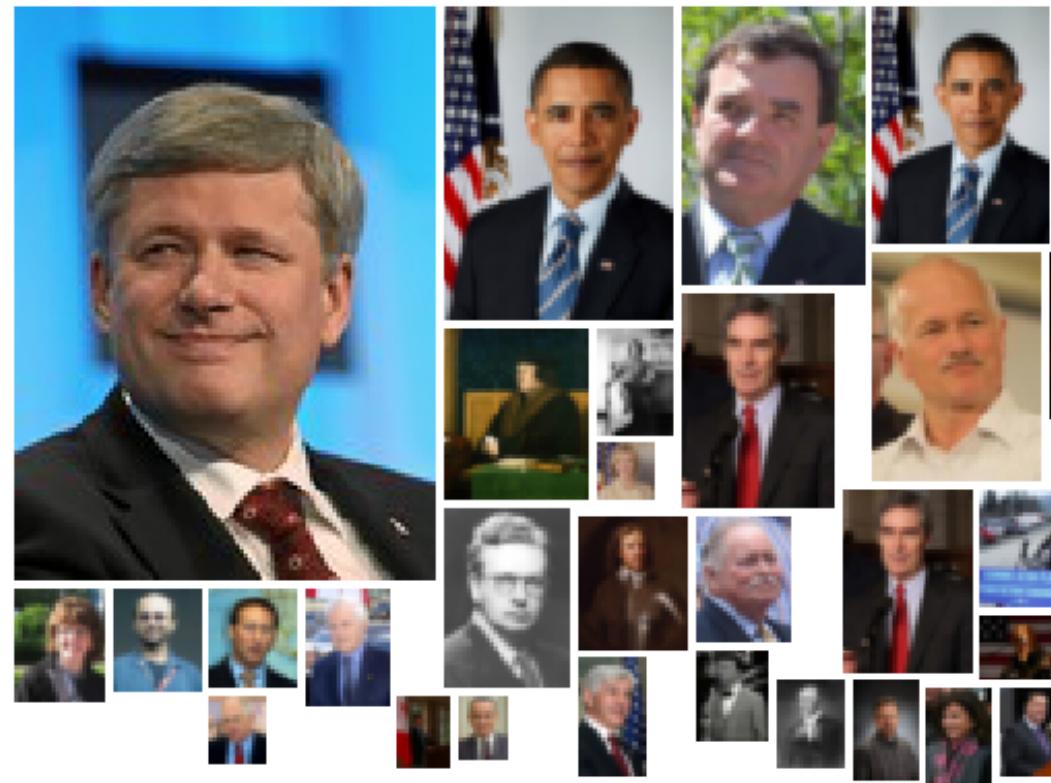
Extract all Text

```
1. i2millig@rho: ~/derivatives/cpp.all.plaintext (ssh)
Python      bash      bash      bash      i2millig@rho:... i2millig@rho:... bash
(20060222,liberal.ca,http://liberal.ca/bio_e.aspx?id=35049,Liberal Party of Canada HOME THE TEAM THE P
ARTY MEDIA CENTRE COMMISSIONS YOUR RIDING      Omar Alghabra www.omaralghabra.com Home > Mississauga--E
rindale Riding Map (PDF) Omar Alghabra came to Canada at a very young age, and immediately knew Canada
was his home. He was first elected 2006 as the Member of Parliament for Mississauga-Erindale. Mr. Al
ghabra is an experienced entrepreneur. For the past six years, he has worked for a large multinational
corporation, carrying out different responsibilities including quality assurance, project management, s
ales, contract management and management of a complete department handling a global mandate. Mr. Alghab
ra is an active member of his community. He is the former National President of the Canadian Arab Feder
ation (2004-2005) and a former member of the Community Editorial Board for the Toronto Star. (2003-2004
). Mr. Alghabra is currently a member of the Diversity Council for General Electric Canada and is activ
e in Junior Achievement for the Toronto Region. He was a member of the Multicultural Inter-Agency of Pe
ople from 2001 to 2002. Mr. Alghabra has a degree in Mechanical Engineering from Ryerson University and a
Masters in Business Administration (MBA) from York University.    Omar Alghabra 790 Burnamthorpe West,
Unit 10 905-276-2806   info@omaralghabra.ca Riding President Elias Hazineh Send an email
Home | News | Your Riding | Contact Us | français This website is the property of the
Liberal Party of Canada and may not be reproduced in whole or in part without express written permission.
© Liberal Party of Canada 2006. All rights reserved. Authorized by the registered agent for the Libe
ral Party of Canada. Privacy Policy)
(20060222,liberal.ca,https://liberal.ca/news_e.aspx?id=11470,Liberal.ca HOME THE TEAM THE PARTY MEDIA C
ENTRE COMMISSIONS YOUR RIDING  Celebrating our National Flag  February 15, 2006 February 15 is Nationa
l Flag Day in Canada, which marks the 41st anniversary of the first raising of the maple leaf flag on P
arliament Hill. Today is a celebration of our shared values, common citizenship and sense of pride in t
his great country we call home. The Canadian flag is one of the most recognizable symbols in the world
and flies proudly to remind us all of who we are and where we come from. The maple leaf's symbolic orig
ins date back to the beginning of our nation's history, while the red and white bars on the flag repres
ent strength and unity. Canada's flag was adopted in 1964 under the courageous leadership of Liberal Pr
ime Minister Lester B. Pearson. The idea of changing the Red Ensign which featured the Britain's Union
Jack, was very controversial at the time, with the Conservative Party strongly opposed to changing the
status quo. Facing strong Conservative resistance in the House of Commons, Pearson's minority governme
nt fought hard in the name of national unity and Canada's multicultural future to make the new flag a r
eality. In an impassioned speech to the House of Commons, Pearson said: "Mr. Speaker, it is for this g
eneration, for this Parliament, to give them and to give us all a common flag; a Canadian flag which, w
hile bringing together but rising above the landmarks and milestones of the past, will say proudly to t
he world and to the future: I stand for Canada." Thanks to Pearson's courageous leadership, Canadians a
cross this great nation celebrate our flag and what it stands for – a country and a citizenship that ar
e the envy of the world.
Home | News | Your Riding | Contact Us | fran
cais This website is the property of the Liberal Party of Canada and may not be reproduced in whole or
```

Extract all Text (voyant connector)



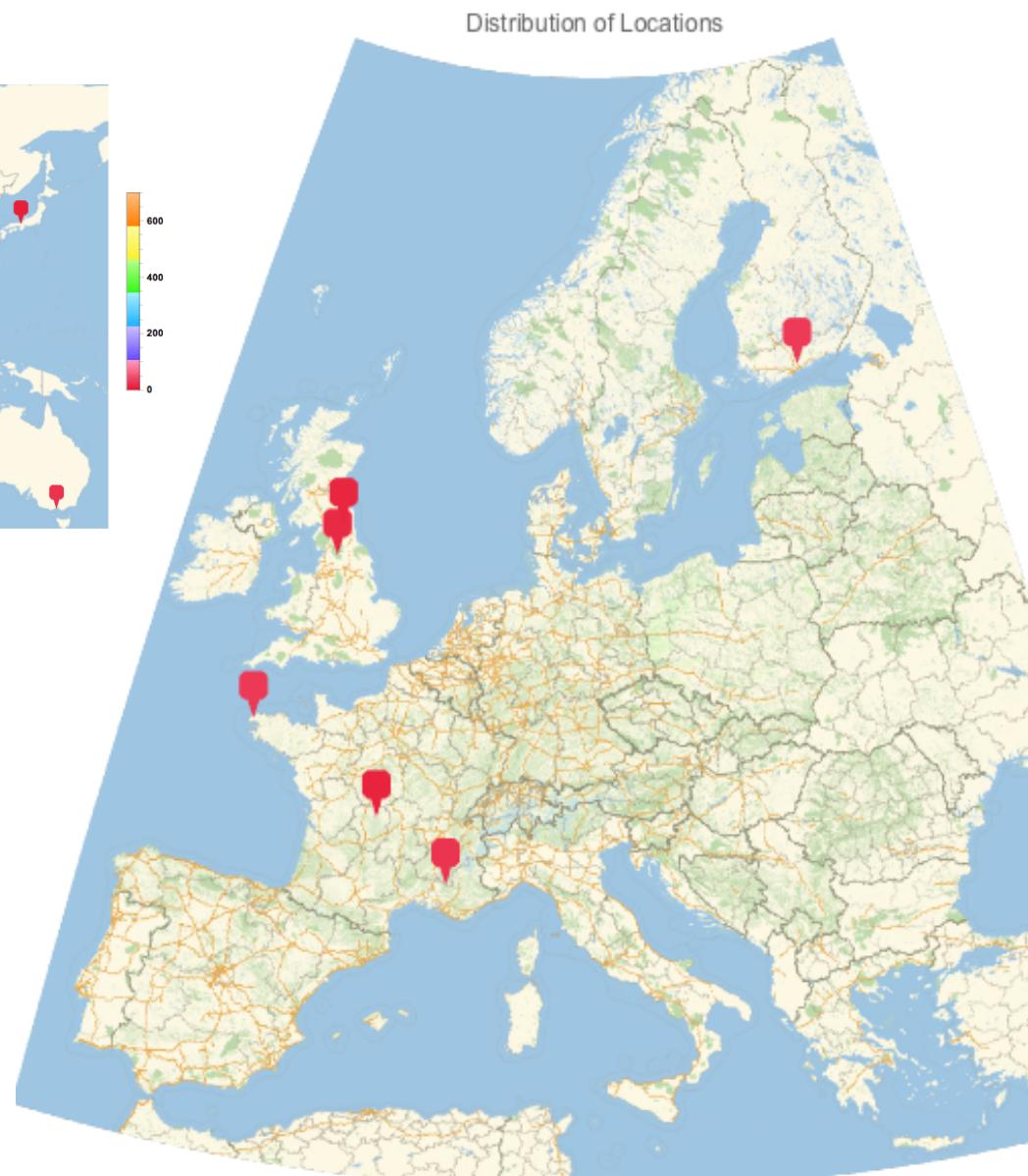
Extract Entities



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

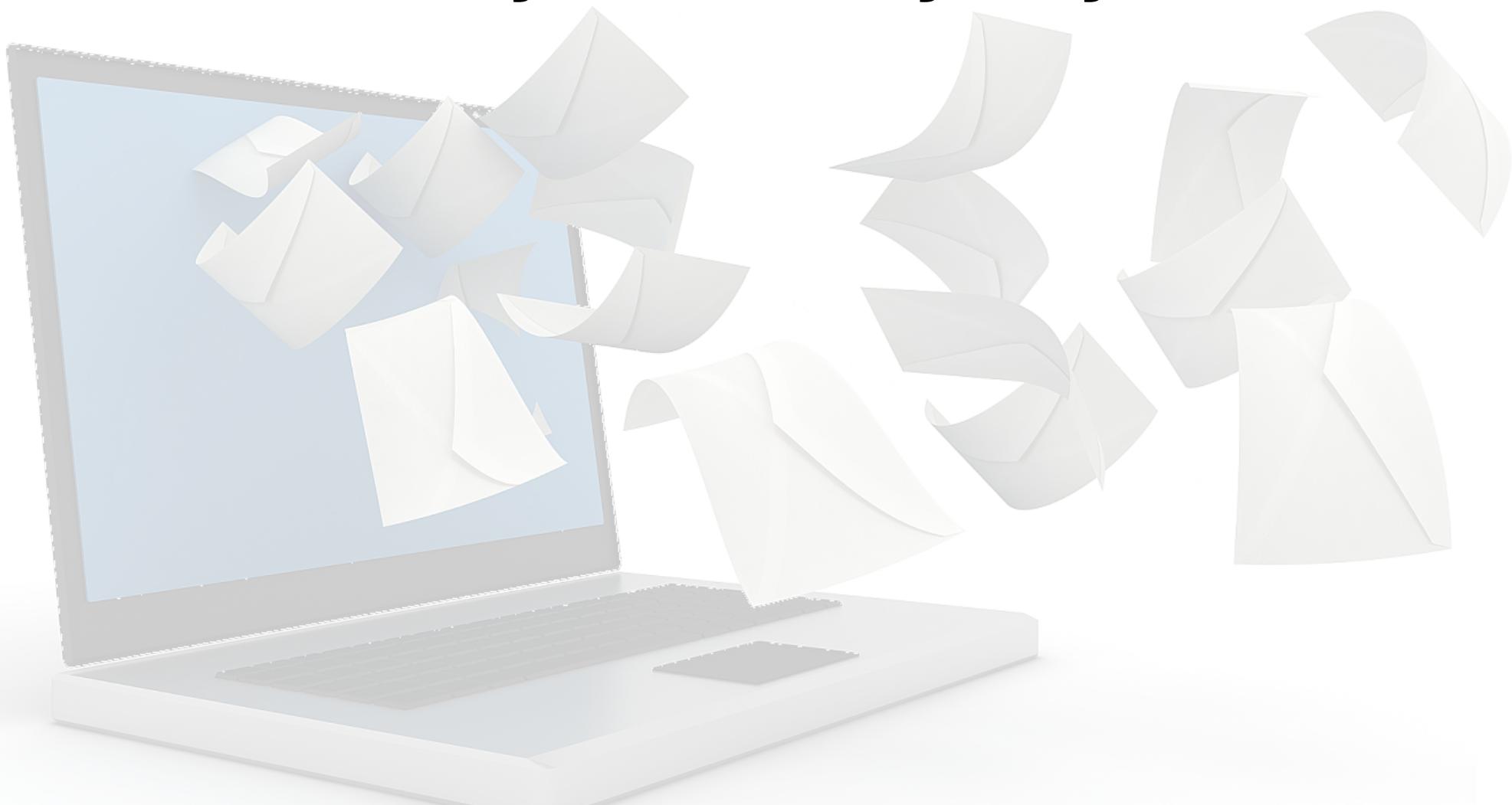
Hi Tony –

See you after class?

Ian

Tells you nothing!

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



[log out & save](#) • [log out & delete](#)

Lookup Contacts

Charge

Nodes [A] [S]Links [Q] [W]

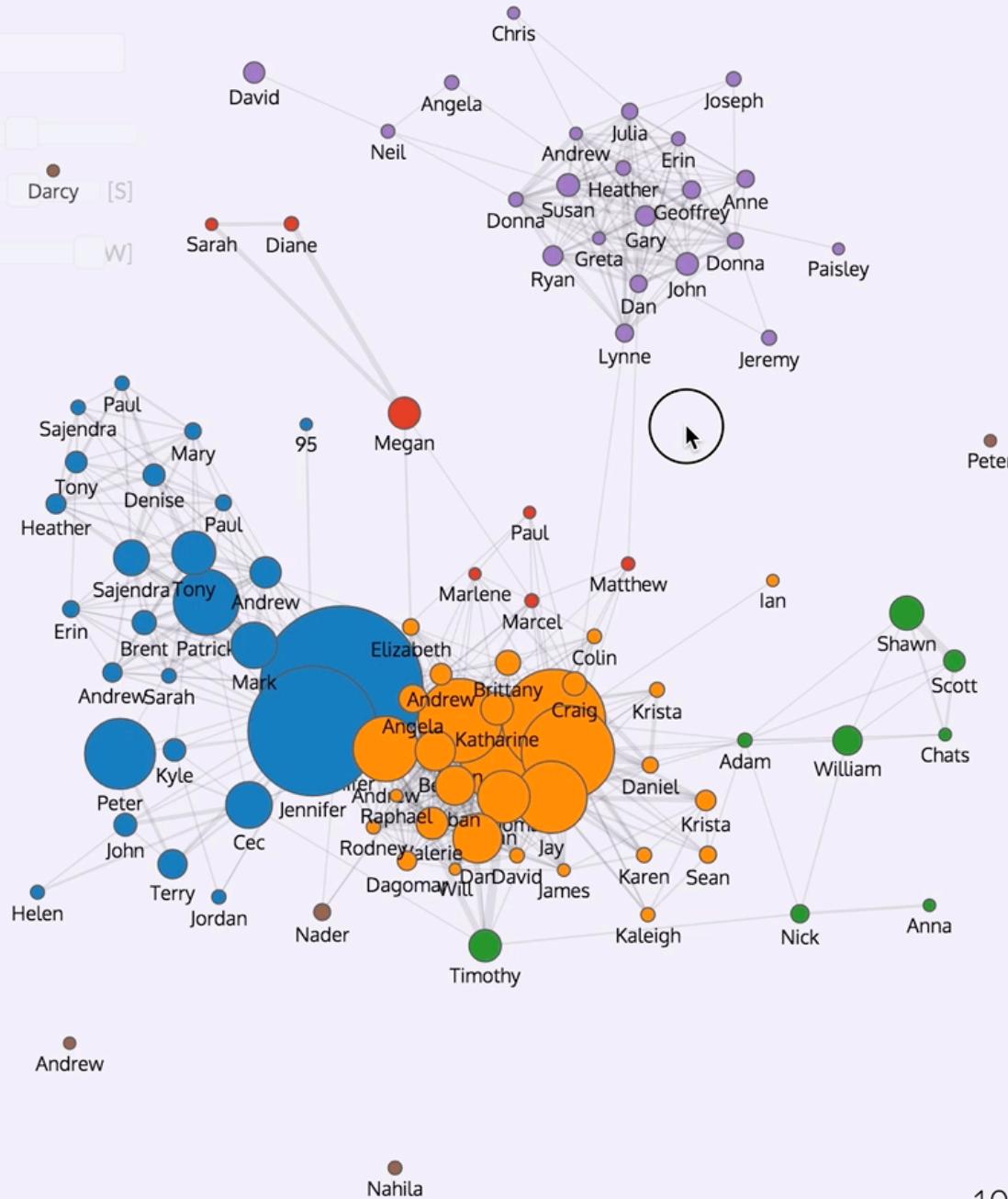
Take a snapshot

[-with labels](#)[-without labels](#)

Feedback?

[Compose](#)

Jeannine

[All](#) • [Past Year](#) • [Past Month](#) • [Past Week](#)10.5 years
26 Sep 2004 - 12 Mar 2015

Ian Milligan

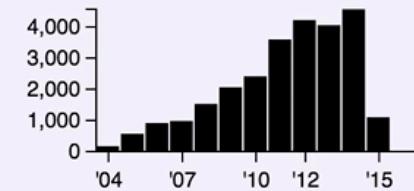


729 collaborators

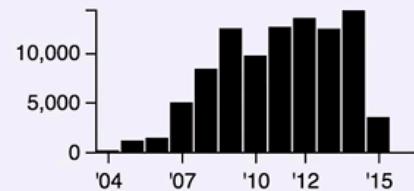
121,854 emails

My Stats[Top Collaborators](#)

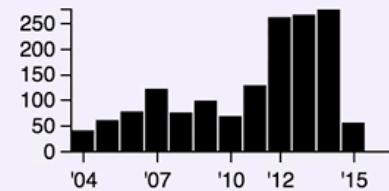
Emails Sent



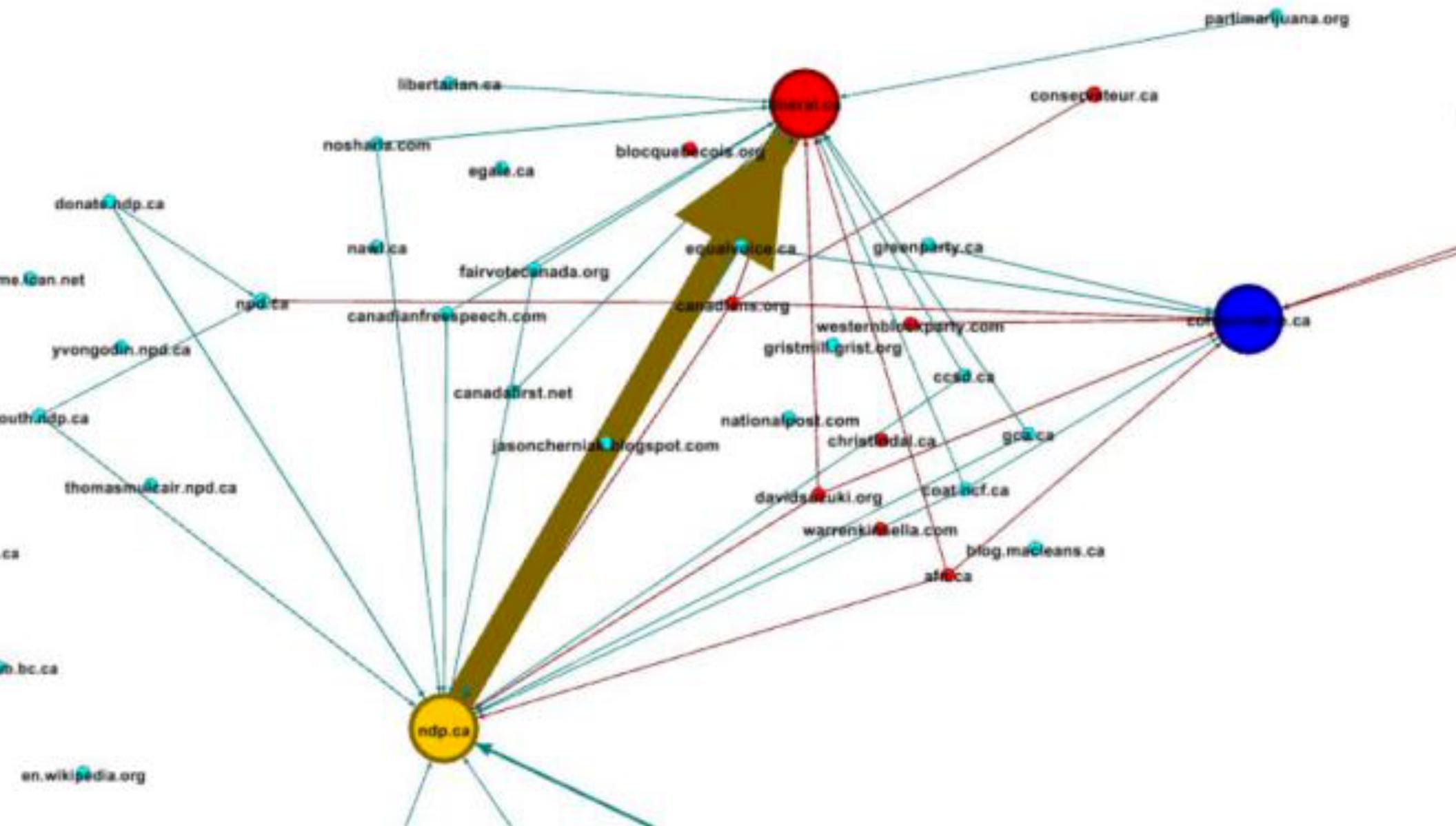
Emails Received



New Collaborators



2005 Canadian Federal Election



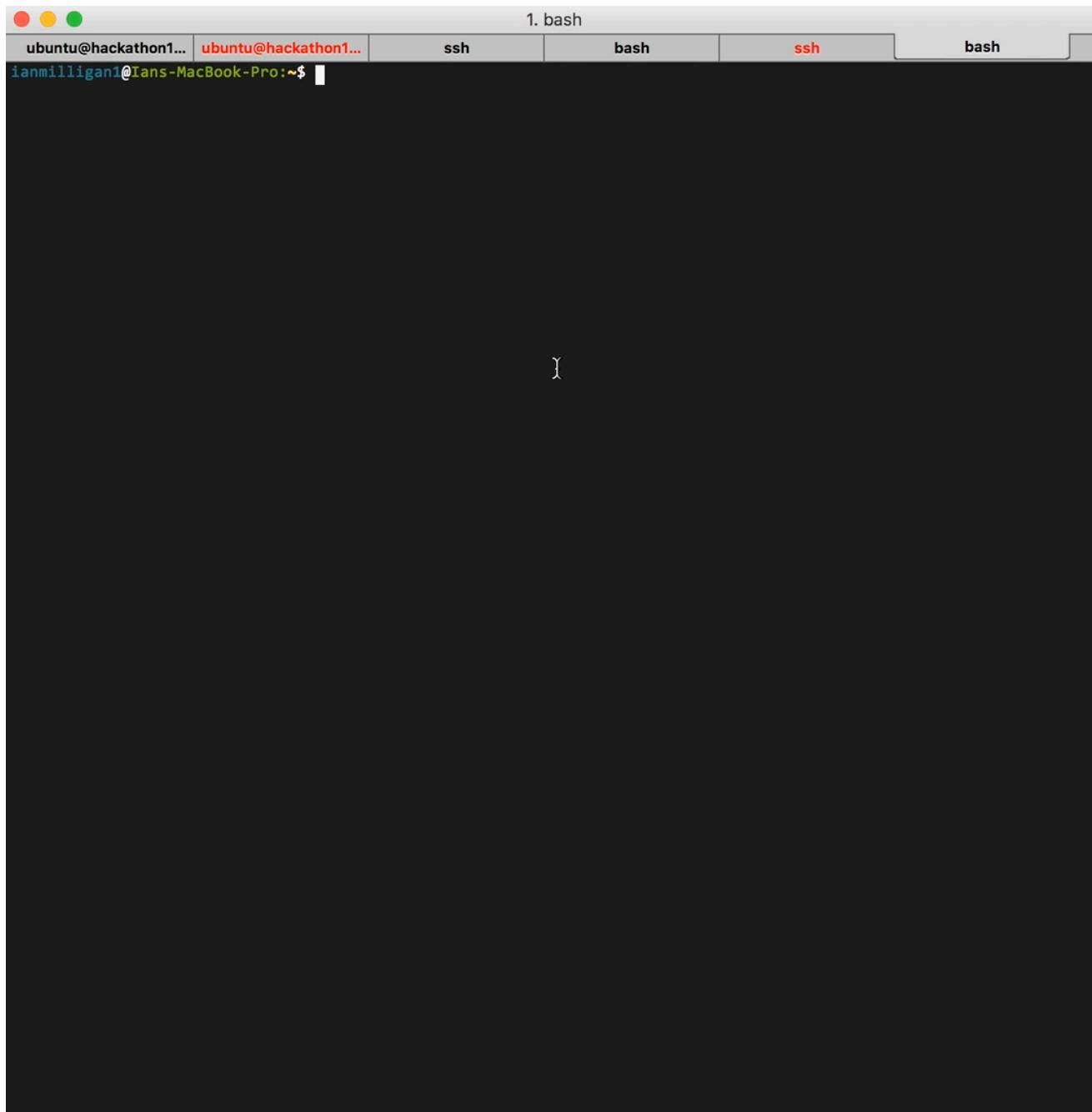
**So how do you use
warcbase in your own
work?**

Step One: Ingest data

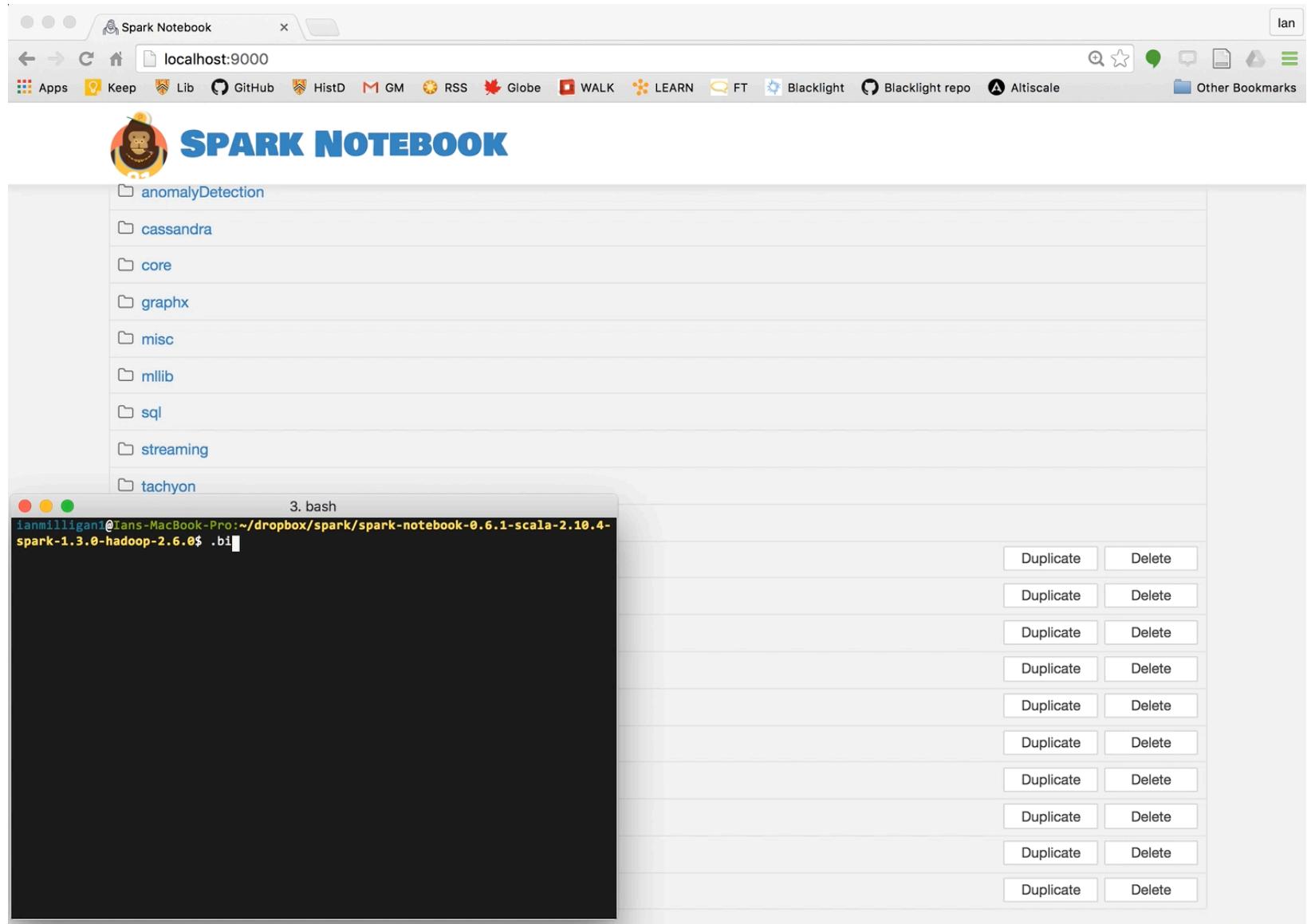


```
i2millig@rho: /mnt/vol1/data_sets/geocities/warc$ ssh
bash                                bash          i2millig@rho: /mnt/vol1/data...
GEOCITIES-20091029114235-00191-1a400110.us.archive.org.warc.gz
GEOCITIES-20091029115416-00171-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029123854-00172-crawling08.us.archive.org.warc.gz
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-20091029165037-00194-1a400110.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-20091029180018-00182-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029182725-00183-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029185058-00184-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-2009102922459-00199-1a400110.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/warc$ du -h
4.1T .
i2millig@rho:/mnt/vol1/data_sets/geocities/warc$
```

Step Two: Basic Shell Analysis



Step Two: Basic Analytics



Step Three: Filtering a Corpus

```
import org.warcbase.spark.matchbox.{ExtractDomain, ExtractLinks, RecordLoader}
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*www\\.", ""), ExtractDomain(f._2).replaceAll("^\\s*www\\.", ""))))
  .filter(r => r._2 != "" && r._3 != "")
  .countItems()
  .filter(r => r._2 > 5)
  .saveAsTextFile("cpp.sitelinks")
```

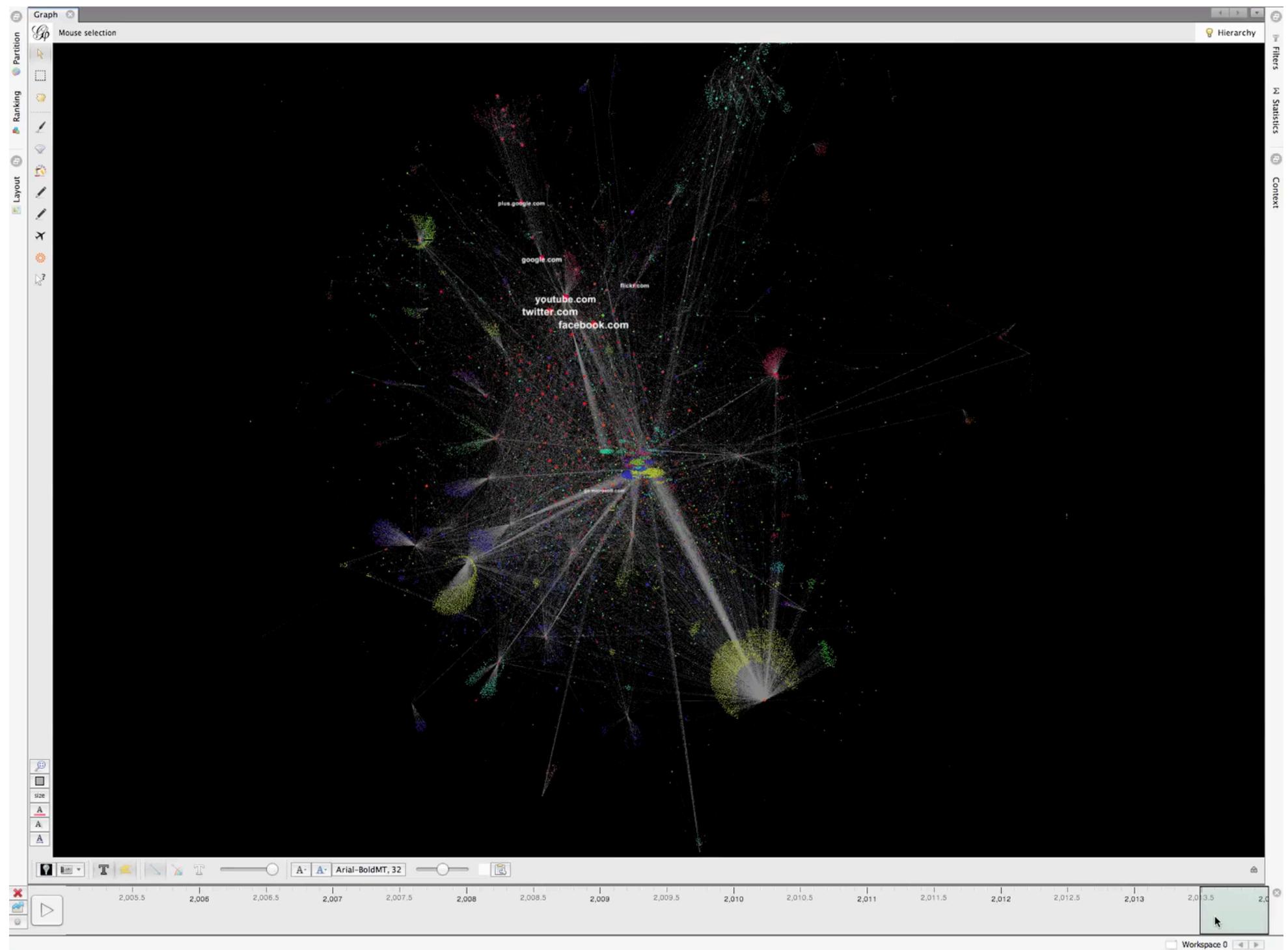
A Link Graph

Step Three: Filtering a Corpus

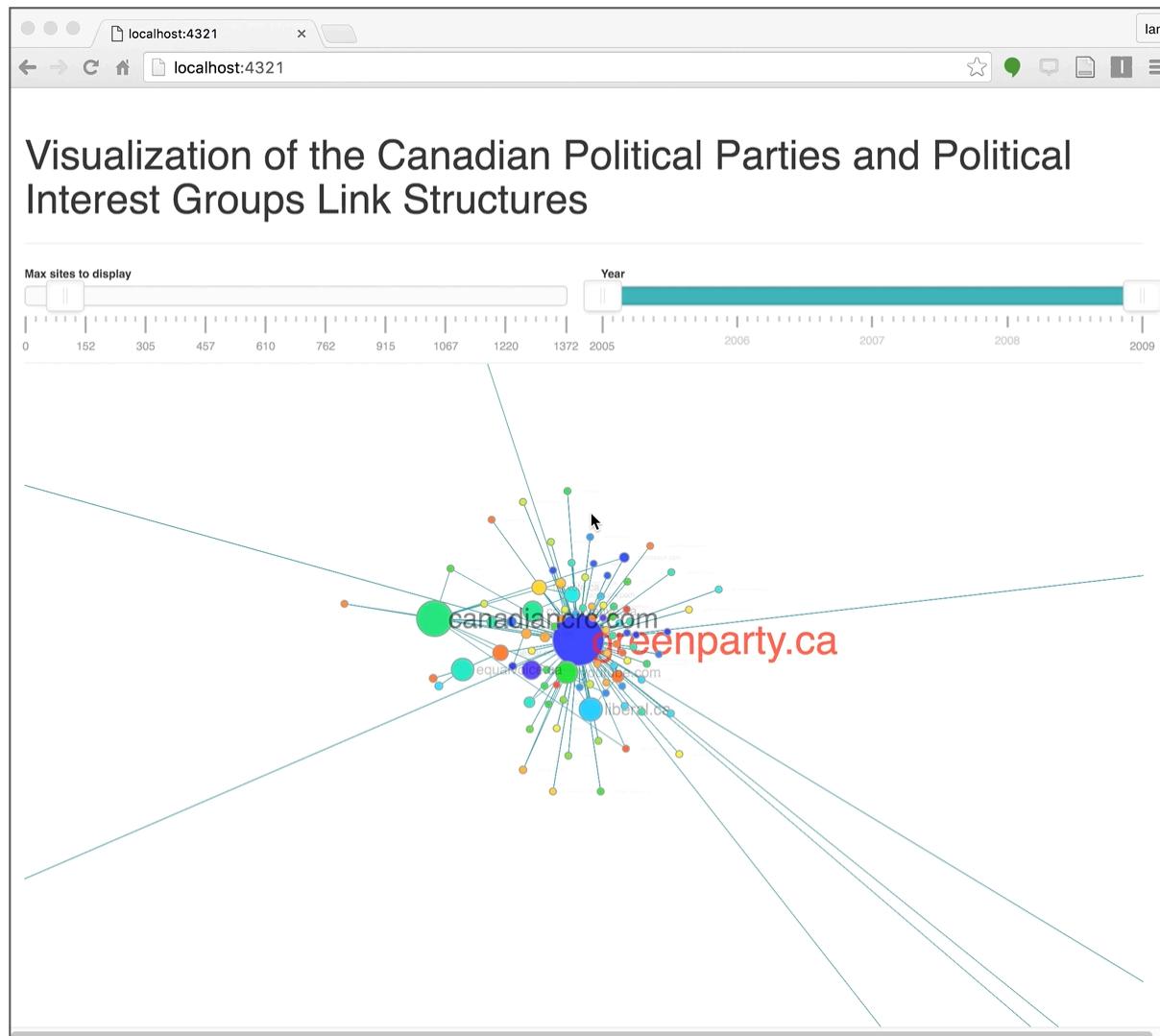
```
((20080612, liberal.ca, liberal.ca), 1832983)
((20060326, ndp.ca, ndp.ca), 1801775)
((20060426, ndp.ca, ndp.ca), 1771993)
((20060325, policyalternatives.ca, policyalternatives.ca), 1735154)
```

Results

Step Four: Visualize to select
sub-corpora



In-browser test visualizations



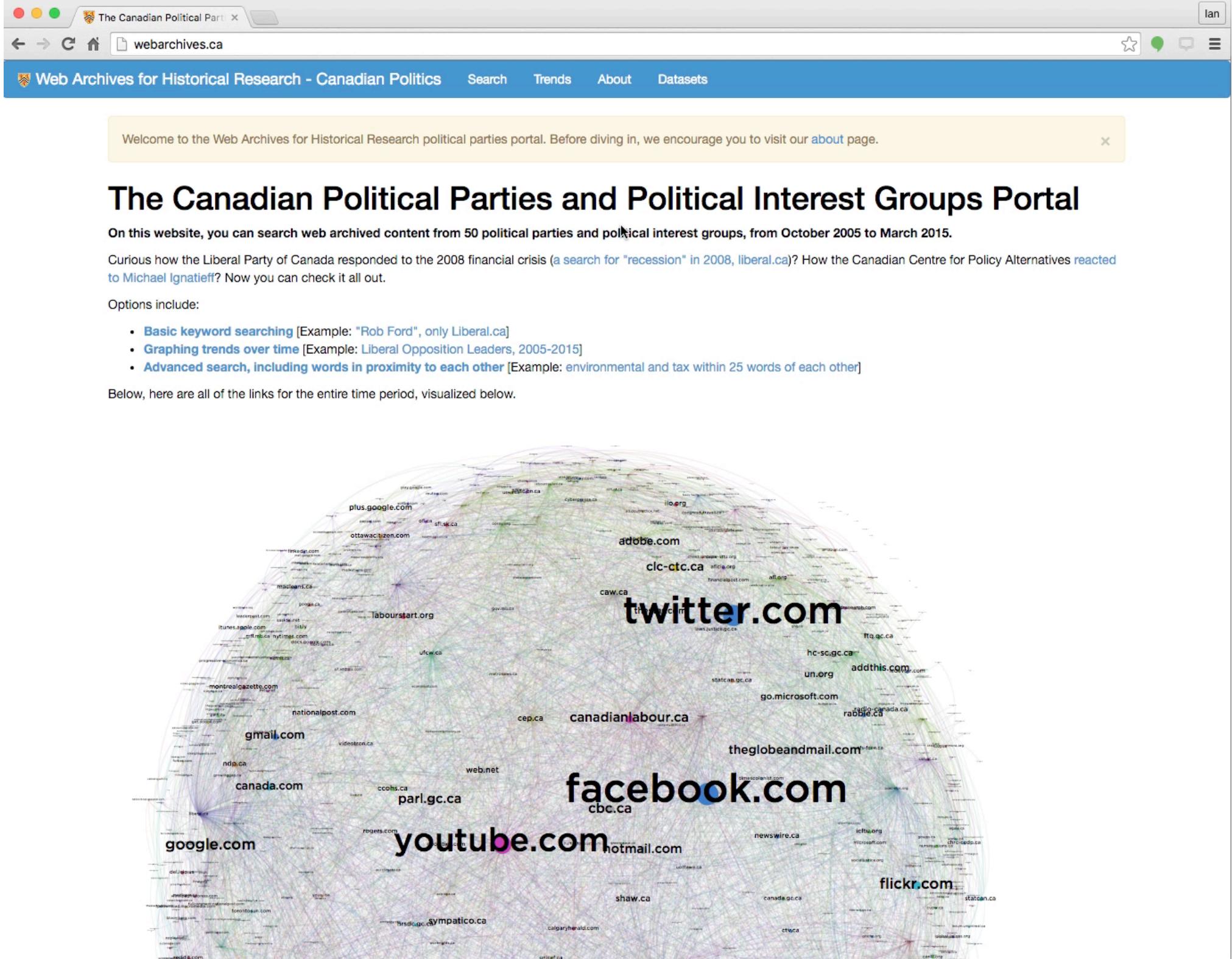
Step Five: Text Analysis

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 Warchbase too!



(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



compute | **calcul**
canada | canada



UNIVERSITY OF
WATERLOO

THE
ANDREW W.
MELLON
FOUNDATION



NSERC
CRSNG

Thanks/Questions!