

Collaborative Digital History

Working with Librarians and Computer Scientists

**American Historical Association
Denver, Colorado
6 January 2017**

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



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Five Minutes o' Fun

- The Problem
- Interdisciplinary Engagement
 - Computer Science
 - Library/Archives
- How to make a team work?

**We have a problem
facing our collective
cultural heritage**

Welcome to GeoCities Home

INTERNET ARCHIVE Wayback Machine 1,662 captures 22 Oct 96 - 6 Sep 15

TV TIME TUNNEL FOR A BLAST Click here for a Blast from TV's past!

GEOCITIES YOUR HOME ON THE WEB

REFRESH PARIS HERITAGE ATHENS

ENTER HERE INFORMATION NEIGHBORHOODS WHAT'S NEW WHAT'S COOL WHAT IS GEOCITIES?

* Free Home Pages & Free Member Email Advertiser Information

GeoCities Daily Audio Update -- Sponsored by IBM VoiceType Simply Speaking

Today's Cool Homestead Yosemite4273 Sunsets, coastal seagulls and flowers are part of the photographic fare at inedt's homepage.

GeoCities News of the Day - 12/20/96

GEOCITIES LIVE CHRISTMAS TREE! ON CAMERA!

Building a home page for the holidays? Submit your letters to Santa, favorite holiday recipes and other holiday cheer to our special NorthPole neighborhood. And share your holiday spirit with GeoCitizens around the world by helping us trim our virtual holiday tree!

Live at the GeoCities Mainstage

Check out our [schedule of events](#) and be a part of the next show...

Cigardude's Smoking Room - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Stop

Bookmarks Location: <http://www.geocities.com/NapaValley/1070/>

The Smoking Room

Welcome to the Cigar Dude's Smoking Room

Your Choices

- [Cigars](#)
- [Wine](#)
- [Beer](#)
- [Links](#)
- [Home](#)



You are visitor number 

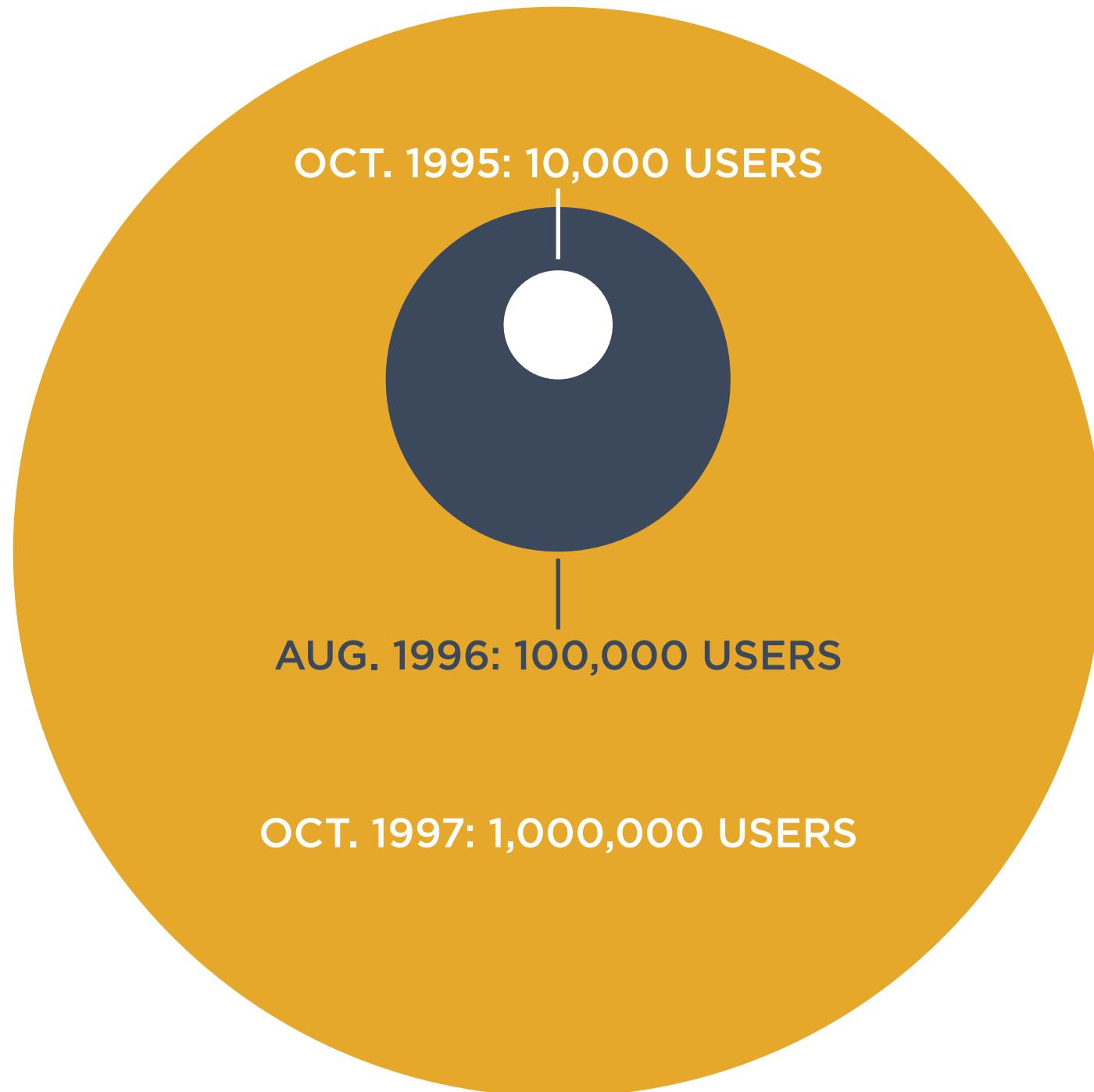
since June 5, 1996

The main purpose of this page is to give me a forum to voice my views and opinions on cigars, good beer and fine wine. It's also a pretty good way for me to learn HTML. This page was first created on May 8, 1996 and will take some time to evolve, so if you are into cigars you might want to check back every once in a while to see what's up. It is always nice to know what other people

Welcome to my home page, devoted to some of the finer pleasures in life: good cigar

Start Cigardude's Smoking ... 00:36

GEOCITIES USERS:



186 million
documents =

**BIG
DATA**

(for me, anyways)



Scarcity



Scarcity
Abundance



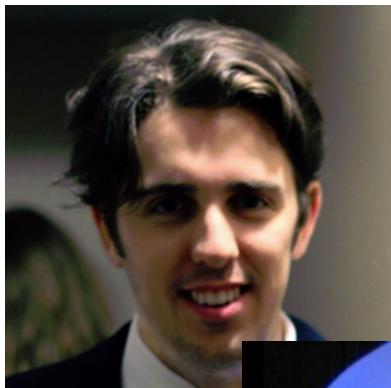
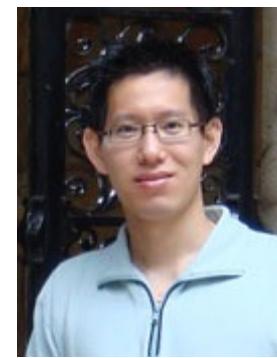
I can't do it alone

Web Archives for Historical Research

Historians



Computer Scientists



Librarians



Governance



Going it alone?

- Some values
 - Learning great skills!
 - Having the chops to talk to other people.
- But.. computer scientists and librarians are going to have the edge on sustainability and scalability
 - Shoddy code that's not sustainable;
 - Not optimized for large datasets (scalability);
 - Missing the diversity of disciplinary perspectives

You need a group!

Making Teams Work

- **Everybody needs to be happy**
 - Students need to be paid and represented on publications (and trained - we host Software Carpentry 2x year);
 - Computer Scientists need to present at conferences;
 - Librarians need to present at their conferences and publish in their journals;
 - And historians need to have material for monographs, articles, etc.;
- **Compromise - recognize that we are all scholars**

Constant Communication

Slack

#walk

3 members | Add a topic

May 19th

ryandeschamps 4:53 PM uploaded and commented on an image: [Pasted image at 2016-05-19, 4:53 PM](#)

“ This one plots the websites (leaves the names out for visibility) and provides a percentage representing the influence of the factors on the result. (kind of like an r-squared).

nruest 7:04 PM
@ianmilligan1: @ianmilligan1 [/data/cpp](#)

all copied over

1

ryandeschamps 9:44 PM
Excellent! Thanks so much!

May 20th

ryandeschamps 12:44 PM
@ianmilligan1: I am going to try and run a job that will give me image urls with counts organized by dates. It's the main object of the group's analysis, and I'm pretty sure you don't have anything like that yet, so I'm going to give it a shot. Feel free to kill the process if it's causing problems elsewhere though

W #1

AU #2

DN #3

I #4

D #5

DH #6

wahr ▾
ianmilligan1

CHANNELS (11)
general
geocities
github
random
rho
shine
twitter
uwaterloo

walk

warbase

DIRECT MESSAGES (9)

alicez
jeremyw
ktmac
ktmack
nruest
poleary
ryandeschamps

+ Invite People

Constant Communication

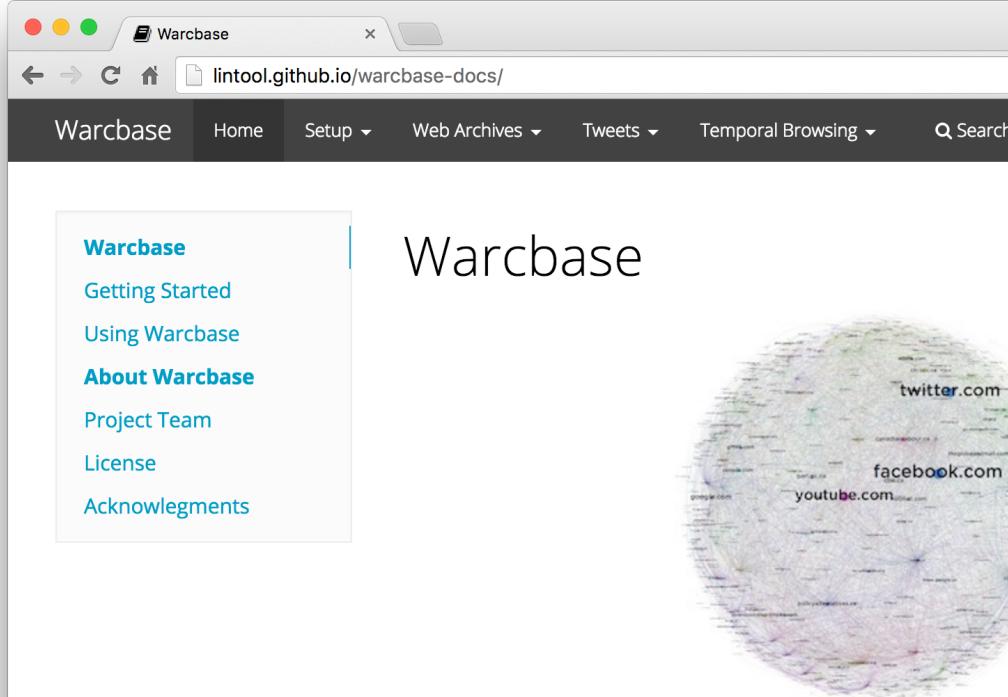
The screenshot shows a GitHub commit history for a repository named 'ianmilligan1/WebArchiving-Articles'. The commits are listed by date, from April 6 to April 18, 2016. Most commits were made by a user named 'lintool'.

- Commits on April 18, 2016:**
 - ACM accepted camera ready versions. (Commit 2d8531b)
- Commits on April 15, 2016:**
 - metadata (Commit aac1477)
- Commits on April 14, 2016:**
 - final camera ready. (Commit 80f9e8d)
 - Final camera ready. (Commit 8bc96bc)
 - tweaked styles (Commit b550ad6)
- Commits on April 11, 2016:**
 - updating image tokens (Commit ce097fc)
 - tiny bit of interim progress (Commit f31da67)
 - Fig 1 (Commit 64dfd41)
- Commits on April 8, 2016:**
 - another pass. (Commit 6a39df5)
 - Release candidate. (Commit 29c7068)
- Commits on April 6, 2016:**
 - checking in first draft (Commit 50ab762)

So what have we
done?

Case One: Warcbase

- **Jimmy Lin** (main developer, CS/lead), **Ian Milligan** (co-lead, history), **Jeremy Wiebe** (history/PhD), **Alice Zhou** (computer science, undergrad), **Youngbin Kim** (computer science, undergrad), **Nick Ruest** (librarian @ York)
- Currently using it on the **GeoCities** and **Canadian Politics** web archives



The screenshot shows a web browser window for the Warcbase documentation site at lintool.github.io/warcbase-docs/. The page title is "Warcbase". The navigation bar includes links for Home, Setup, Web Archives, Tweets, Temporal Browsing, and Search. A sidebar on the left contains links for Warcbase (Getting Started, Using Warcbase, About Warcbase), Project Team, License, and Acknowledgments. The main content area features a large circular network visualization with various nodes labeled with domain names like twitter.com, facebook.com, youtube.com, and google.com.

Warcbase

Warcbase is an open-source platform for managing web archives. The platform provides a flexible data model for storing and managing web pages, their metadata and extracted knowledge. Tight integration with Hadoop allows for distributed analytics and data processing via [Spark](#). For more information about the architecture behind it, visit our [about page](#).

Our documentation can be accessed by using the drop-down menu in the top right corner of the page.

Getting Started

You can [download Warcbase here](#). The easiest way would be to [clone the GitHub repository](#) and follow the [tutorial](#). For a conceptual and practical introduction to the command-line interface, see [the Bash Command Line tutorial](#) and James Baker's "Introduction to the Bash Command Line" article.

Using Warcbase

If you've just arrived, you're probably interested in using [Spark](#) or [Hadoop](#) to process your web archive. You can also use the [command-line interface](#) to interact with Warcbase directly.

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.

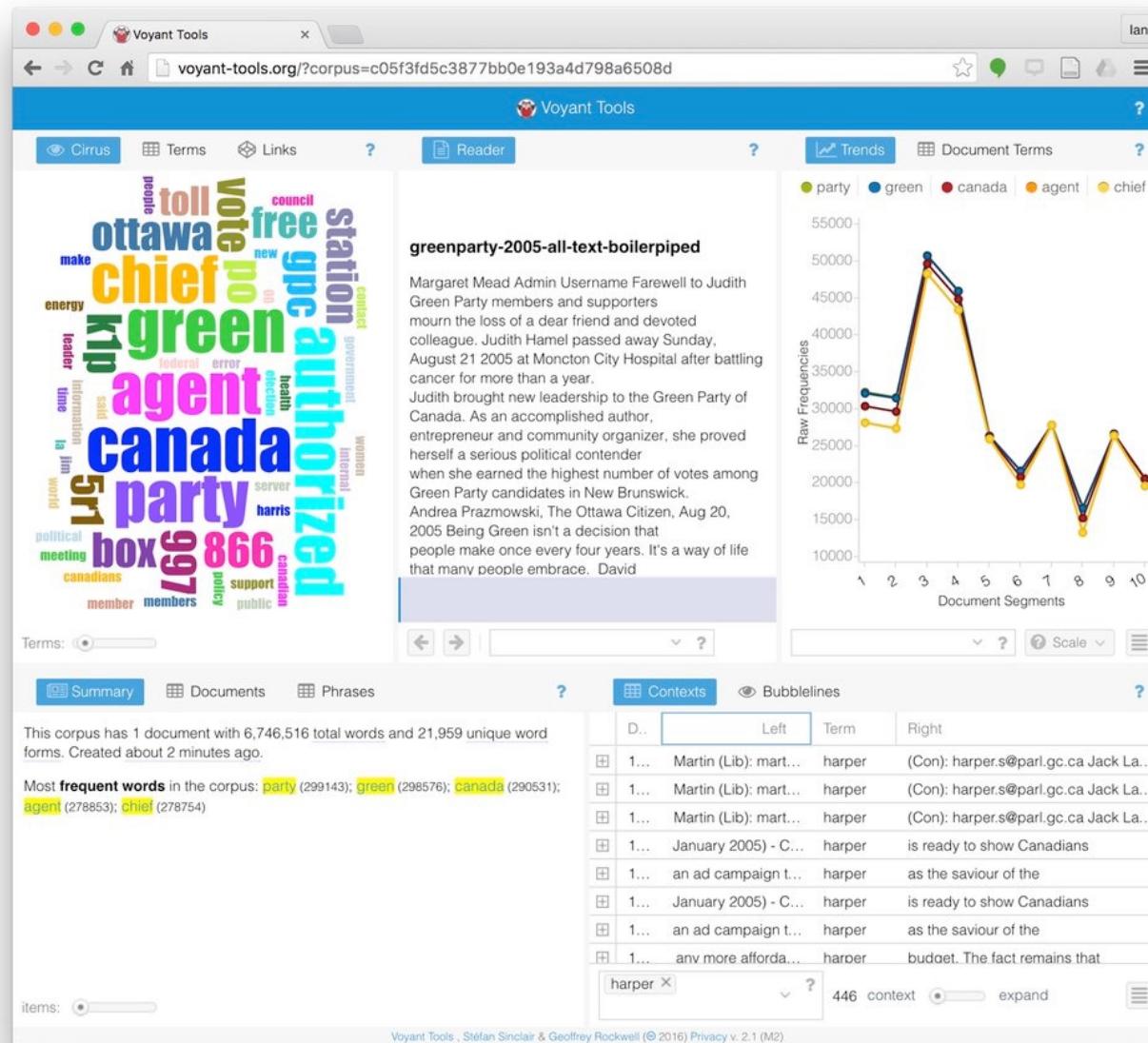
If you want to run it in your Spark Notebook, the following script will show in-notebook plain text:

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

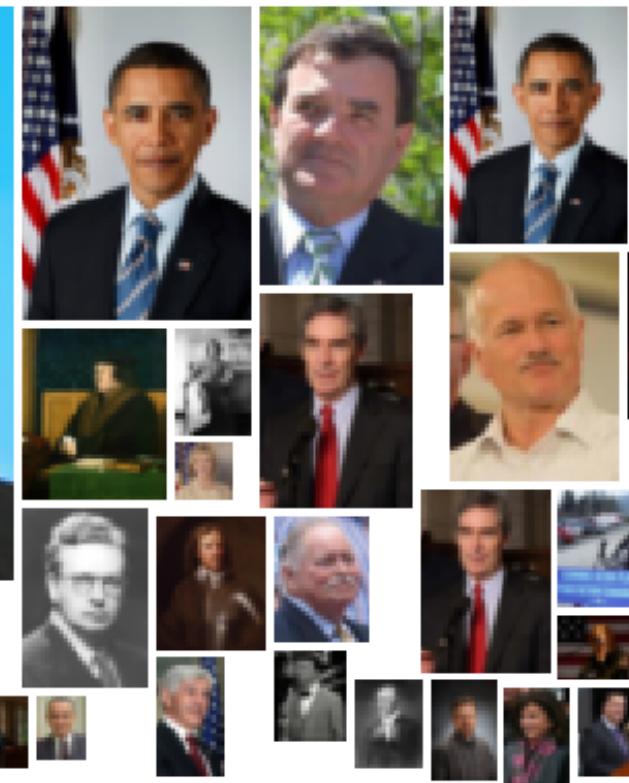
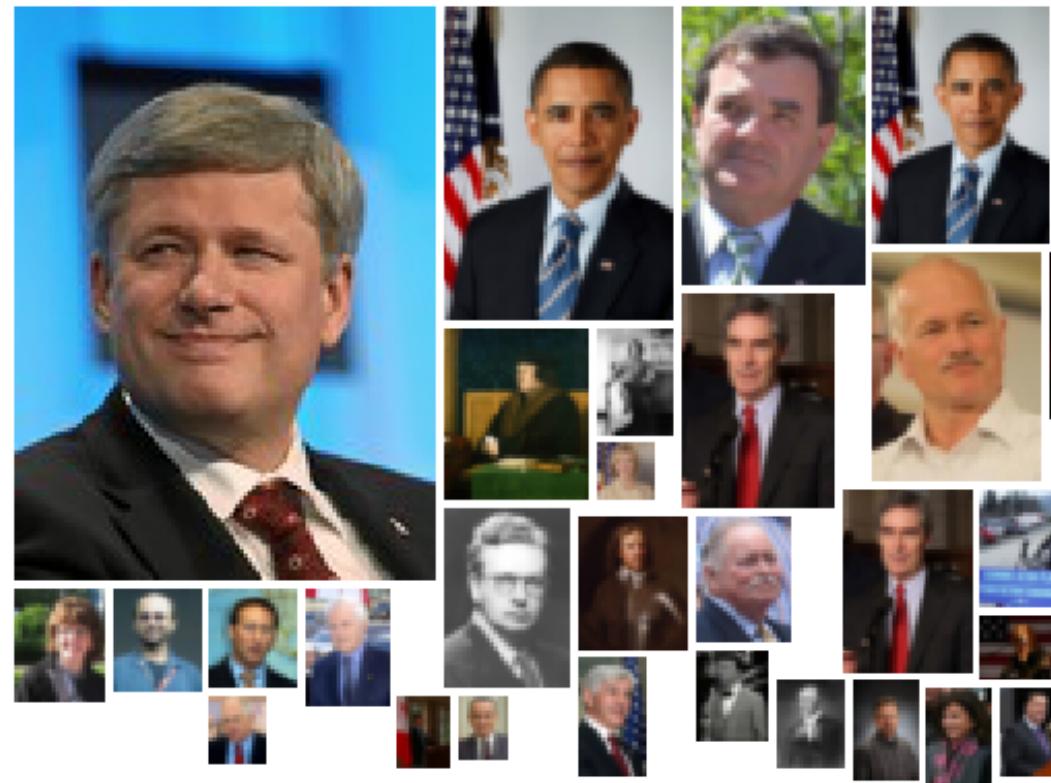
Extract all Text

```
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



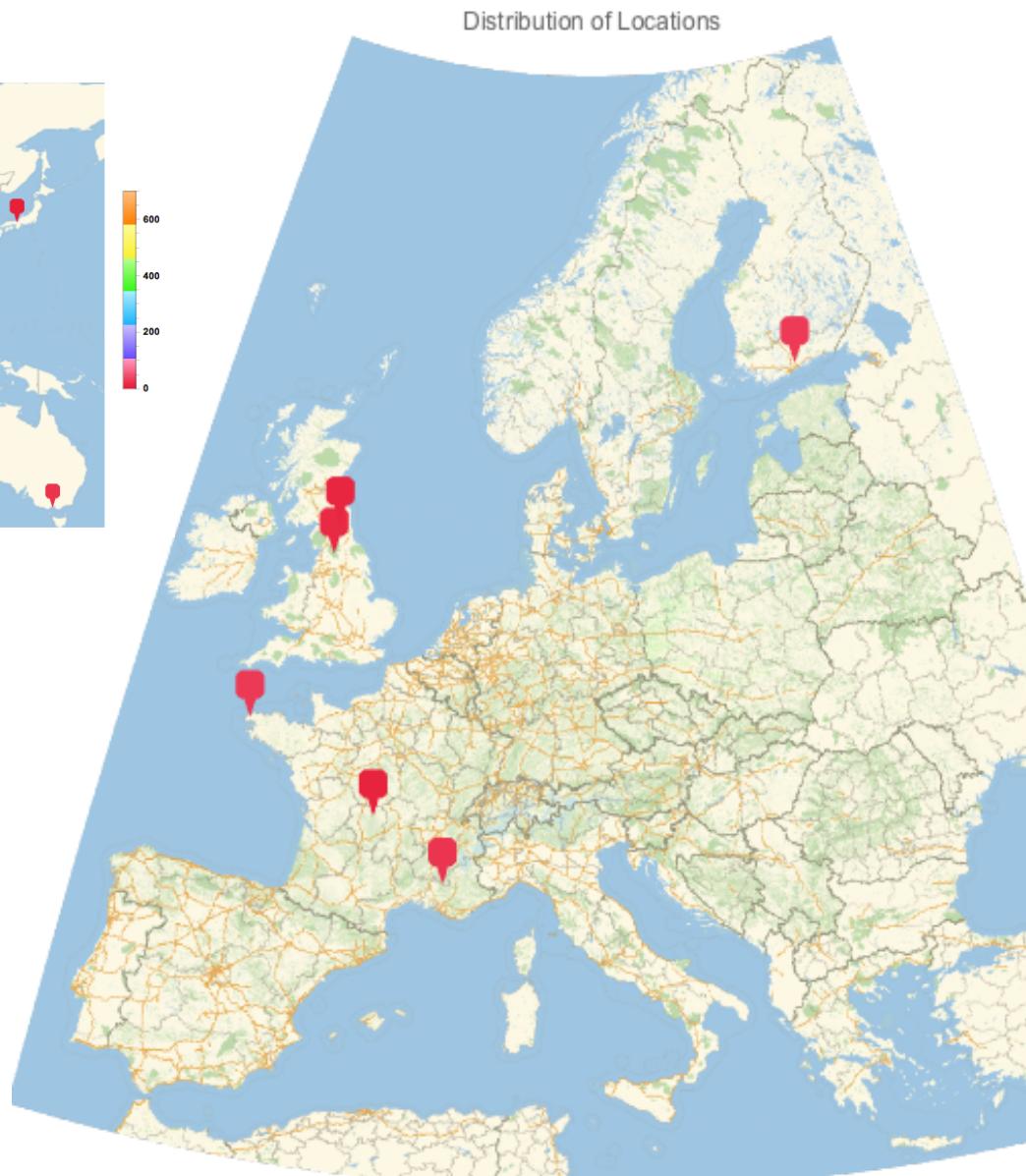
Extract Entities

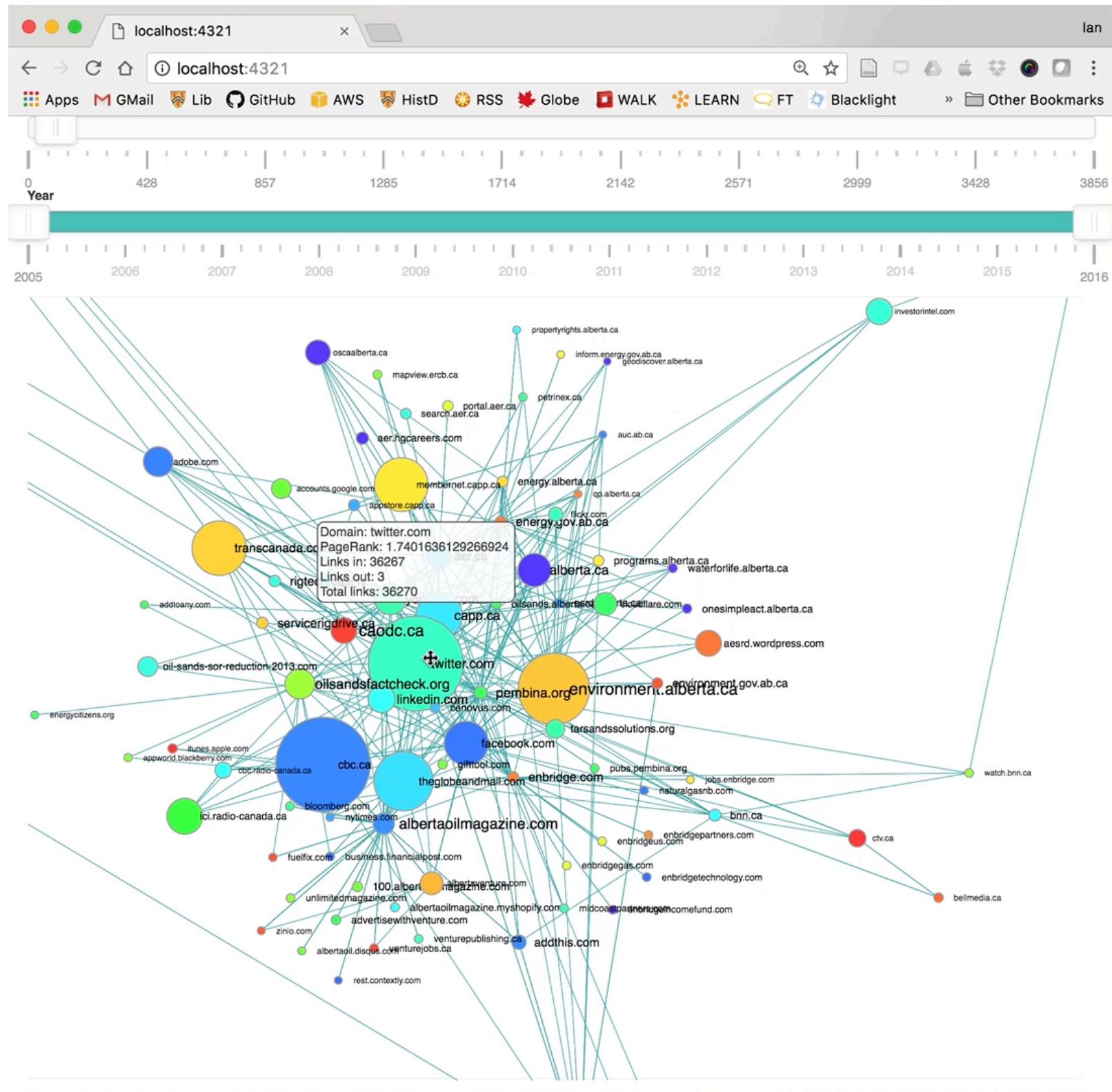


Extract Entities



```
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,
```





Outputs

- *ACM Journal of Computing and Cultural Heritage* (warcbase) - **computer science**
- *ACM/IEEE Joint Conference on Digital Libraries* (Solr indexer) - **library/computer science**
- Web Archives for Longitudinal Knowledge (WALK) Project (warcbase use case) - **digital humanities**
- *International Journal of Humanities and Arts Computing* (IJHAC), *Digital Studies* (humanities) - **digital humanities/history**
- *Code4Lib* conference and journal (libraries) - **library**

Example Two: Web Archives for Longitudinal Knowledge (WALK) Project

WALK

- **Web Archives for Longitudinal Knowledge (WALK)**
- **Ian Milligan** (Co-PI, UW) + **Nick Ruest** (Co-PI, York), w/ **Geoff Harder**, **Todd Suomela**, **Sonya Betz**, **Peter Binkley**, **Geoffrey Rockwell** (Alberta), **Jefferson Bailey** (Internet Archive), and **John Simpson** (Compute Canada).
- 20 TB of Web Archives/Six Institutions on our server
- Common portals



Welcome to the Web Archives for Longitudinal Knowledge (WALK) portal. Before diving in, we encourage you to visit our [about](#) page.

Web Archives for Longitudinal Knowledge (WALK) Portal

This website is home to the **Web Archives for Longitudinal Knowledge (WALK) Project**, an envisioned Canadian national Web Archiving portal. Spearheaded by the [University of Waterloo](#), [York University](#), and the [University of Alberta](#), we plan to bring together interested Canadian partners to provide access to their collections.

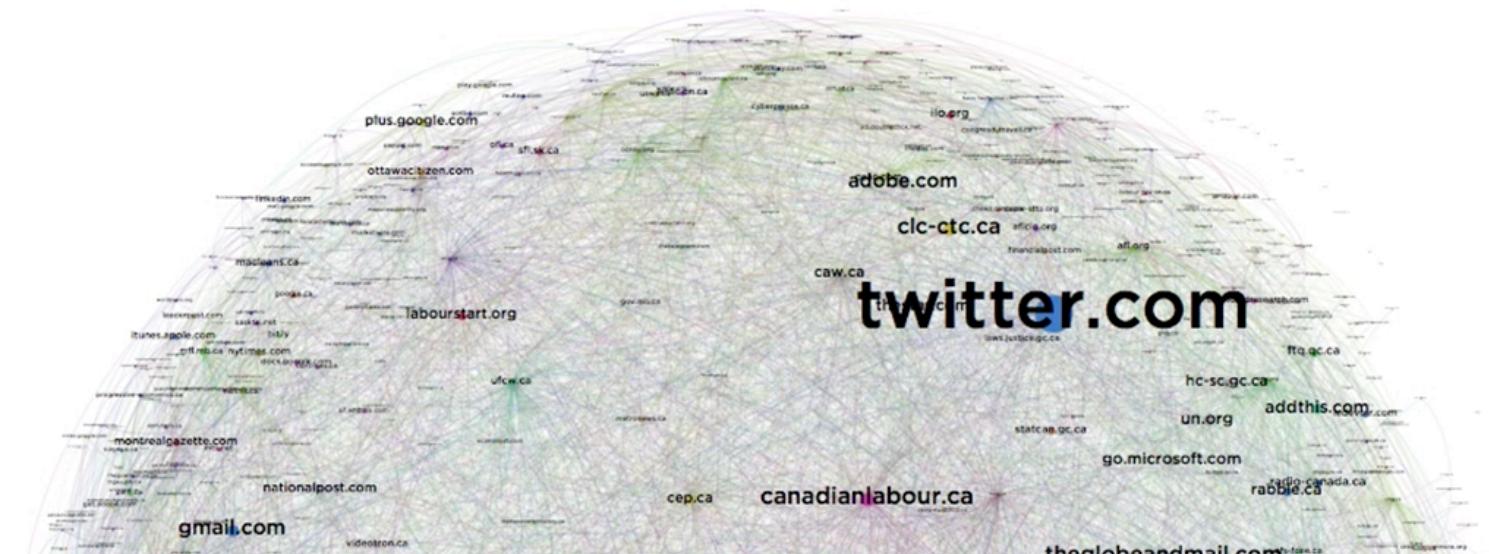
Currently, this is a prototype site providing access to one such archive, the University of Toronto's Canadian Political Parties and Political Interest Groups collection. This website allows you to search content from 50 political parties and political interest groups, from October 2005 to March 2015.

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]

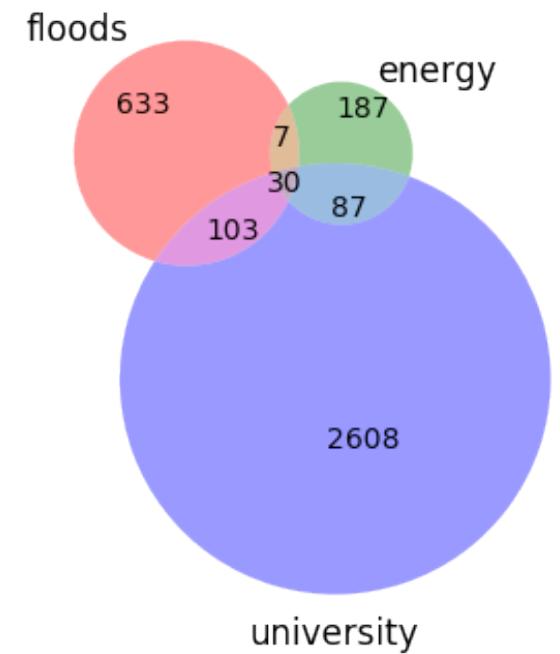
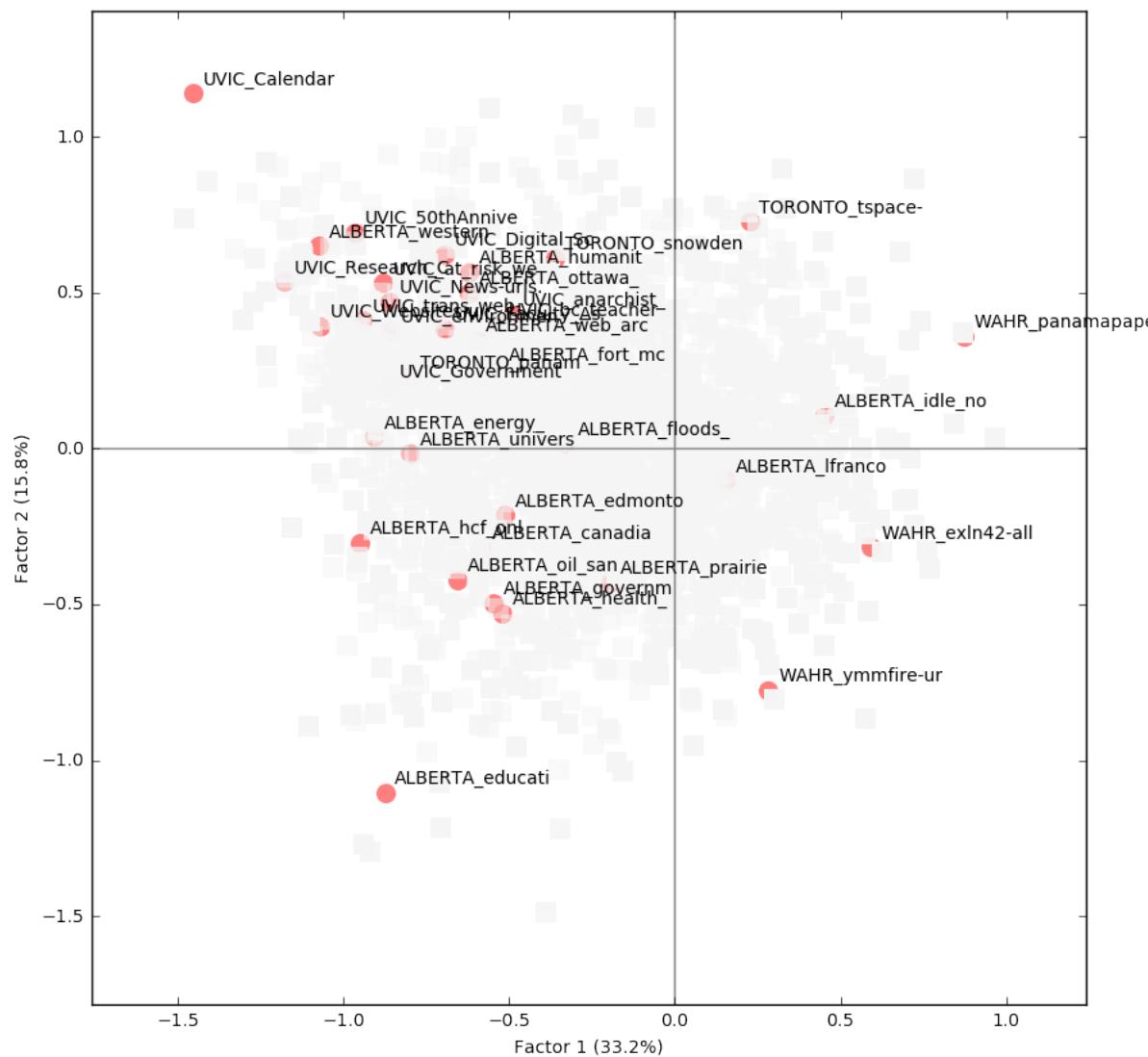
Below, here are all of the links for the entire time period, visualized below.





**Bringing together it all into an
interface like this, central hub
for Web Archives in Canada.**

Exploring collection coverage and curatorial models



**Only possible because CS,
historians, and librarians work
together, communicate together,
and publish together.**



Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada



compute
canada | **calcul**
canada



UNIVERSITY OF
WATERLOO

Thanks!

Ian Milligan
Assistant Professor
@ianmilligan1



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History