Collaborative Digital History

Working with Librarians and Computer Scientists

American Historical Association

Denver, Colorado
6 January 2017

Ian Milligan
Assistant Professor
@ianmilligan1

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Five Minutes o' Fun

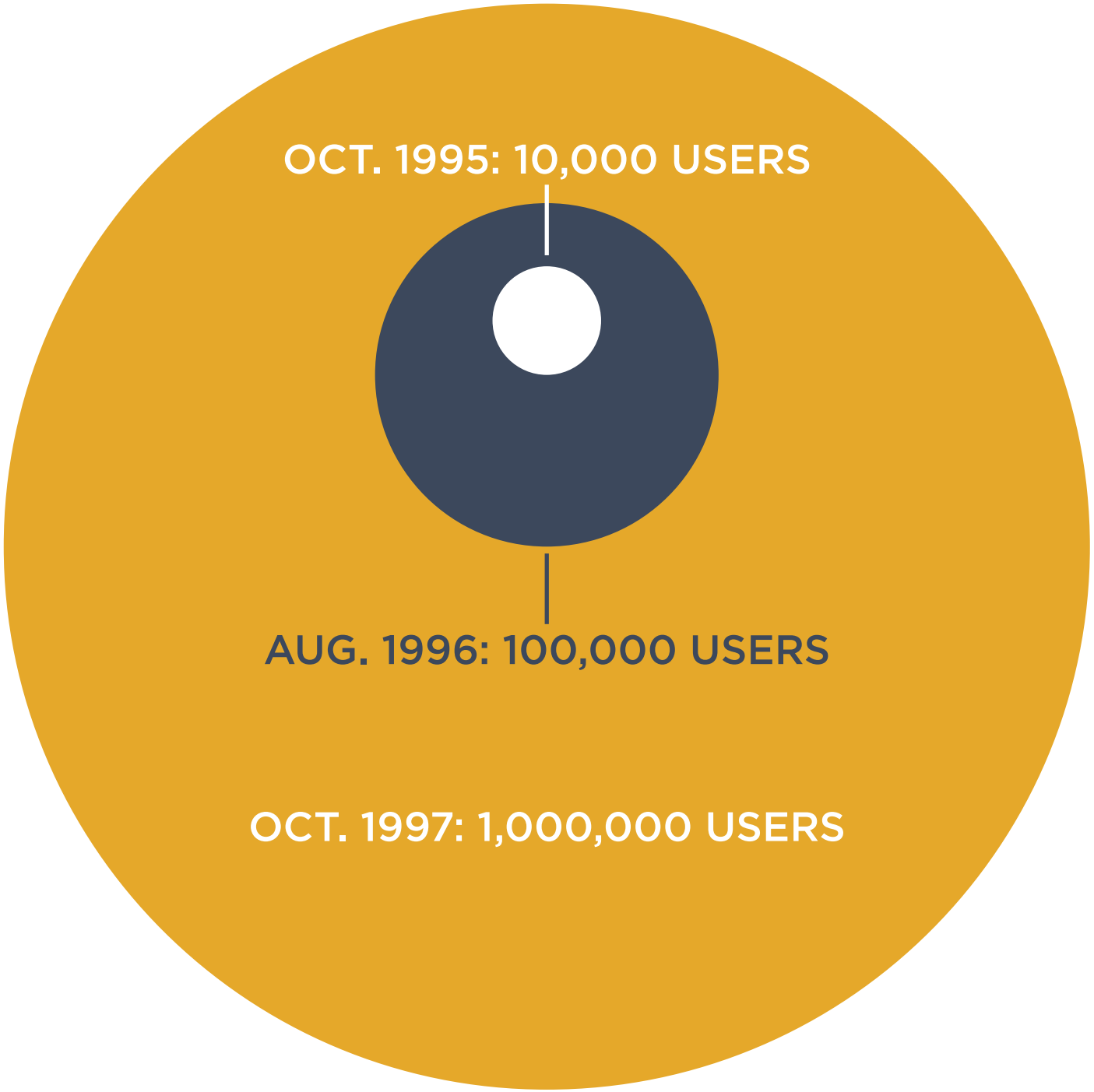
- The Problem
- Interdisciplinary Engagement
  - Computer Science
  - Library/Archives
- How to make a team work?

We have a problem facing our collective cultural heritage


Welcome to GeoCities
web.archive.org/web/19961220165813/http://www.geocities.com/
http://www.geocities.com/
Go
OCT DEC FEB
20
1995 1996 1998
1,662 captures
22 Oct 96 – 6 Sep 15
Close
Help
TV TIME TUNNEL
CLICK HERE
FOR A BLAST FROM TV's PAST
Click here for a Blast from TV's past!
GEOCITIES
YOUR HOME ON THE WEB
ENTER HERE
INFORMATION
NEIGHBORHOODS
WHAT'S NEW
WHAT'S COOL
WHAT IS GEOCITIES?
PARIS
HEARTLAND
ATHENS
TIMES SQUARE
Our communities are home to the most popular collection of FREE HOME PAGES & E-MAIL on the web. Please join or visit one of our 29 neighborhoods today.
* Free Home Pages & Free Member Email
Advertiser Information
GeoCities Daily Audio Update -- Sponsored by IBM VoiceType Simply Speaking
HOT GEOSTORE
Geotoon of the week
Today's Cool Homestead
Yosemite4273
Sunsets, coastal seagulls and flowers are part of the photographic fare at inetd'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
Your Choices
Cigars
Wine
Beer
Links
Home
Welcome to the Cigar Dude's Smoking Room
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)


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

Scarcity

~~Scarcity~~

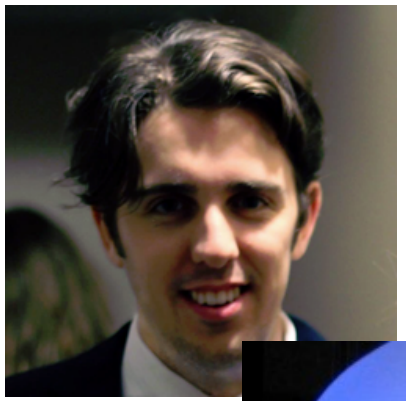
Abundance

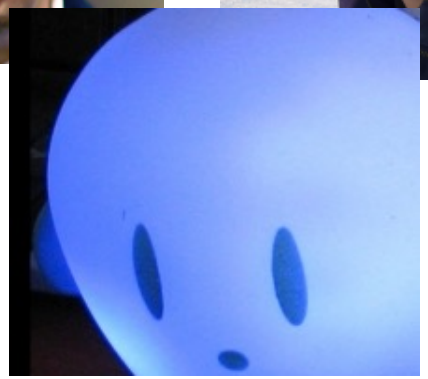
I can’t do it alone

Web Archives for Historical Research

Historians

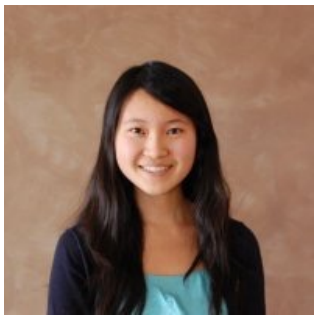






Computer Scientists





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

wahr
ianmilligan1

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

DIRECT MESSAGES (9)
alicez
jeremyw
ktmac
ktmack
nruest
poleary
ryandeschamps

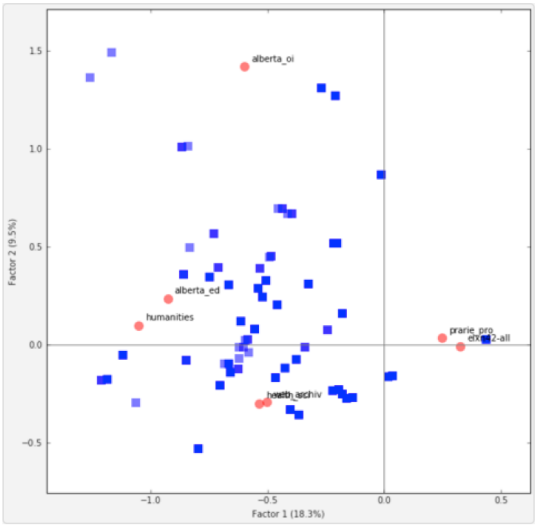
+ Invite People

#walk
3 members | Add a topic

Search

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.

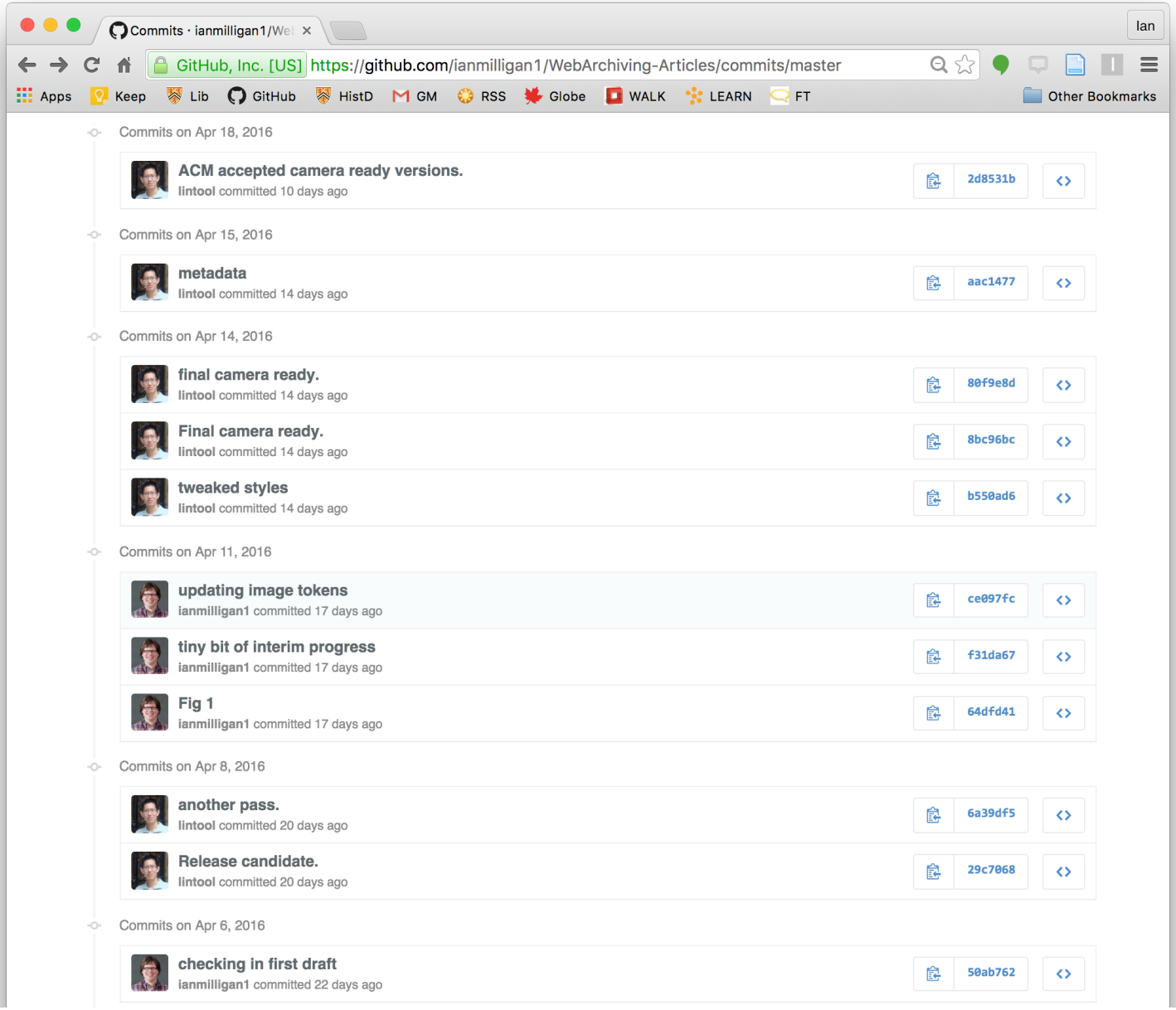
nruest 7:04 PM
@ianmilligan1: @ianmilligan1 `/data/cpp`
all copied over
1

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

Constant Communication

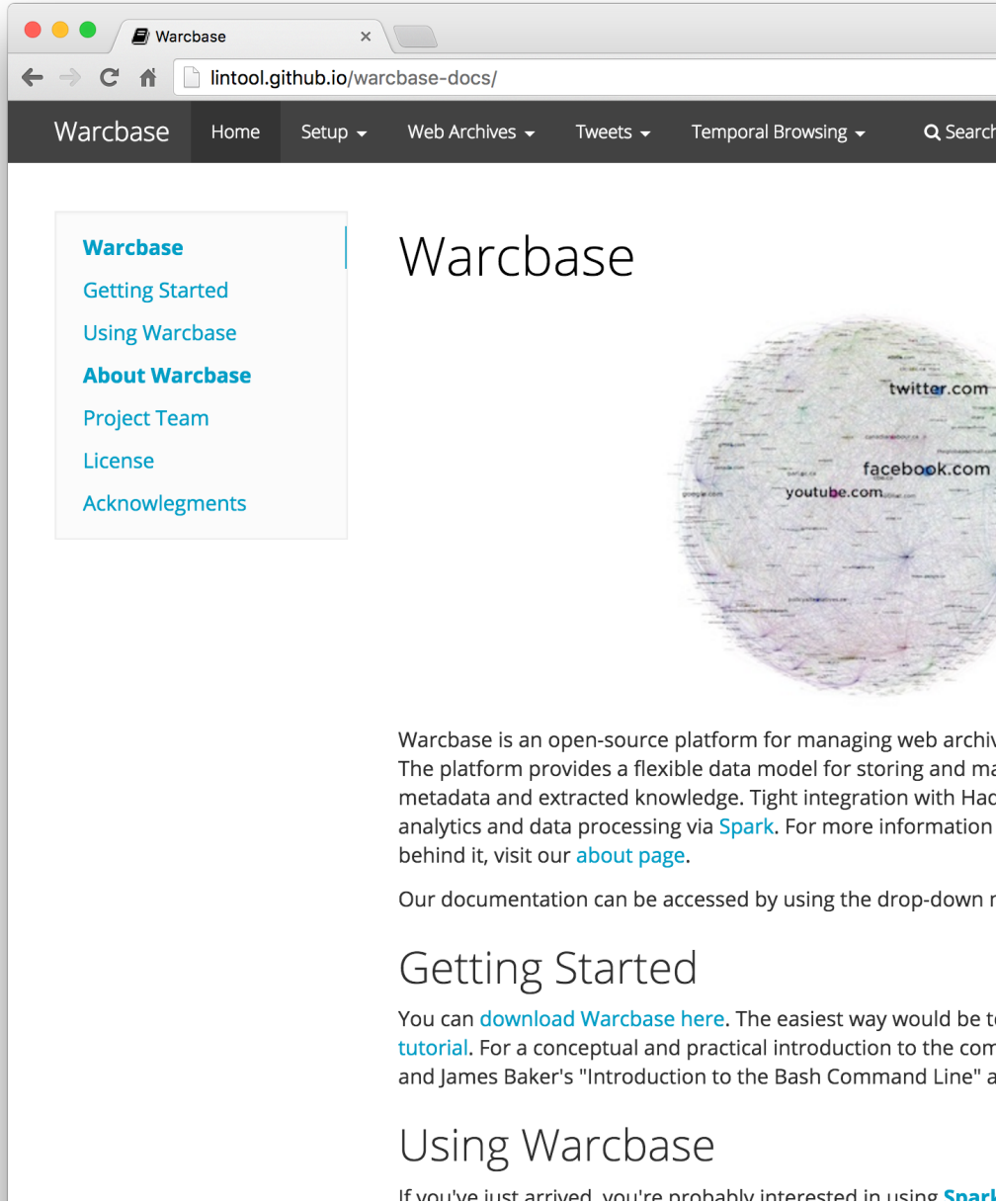


So what have we done?

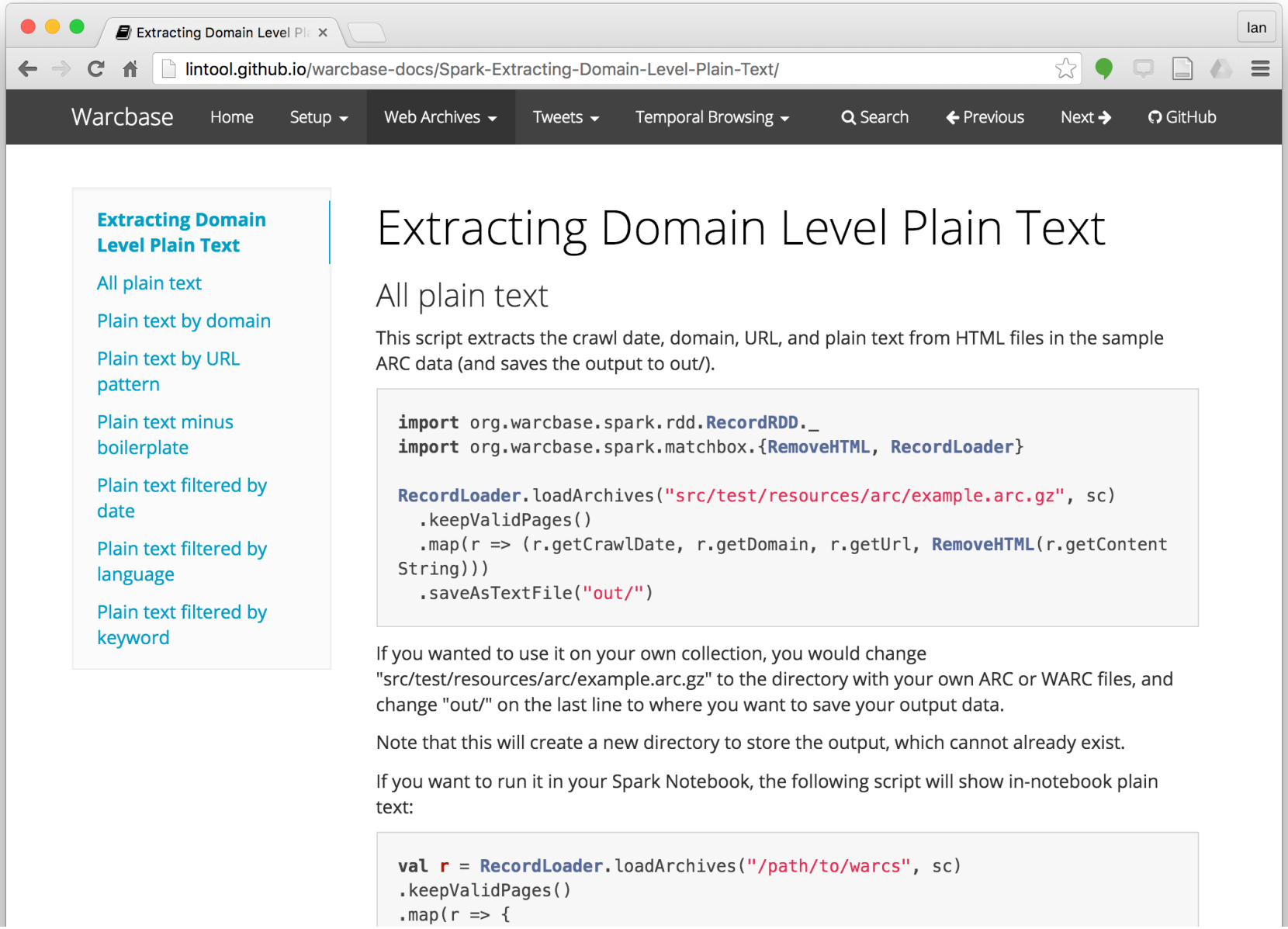
Case One: Warcbase

- **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)

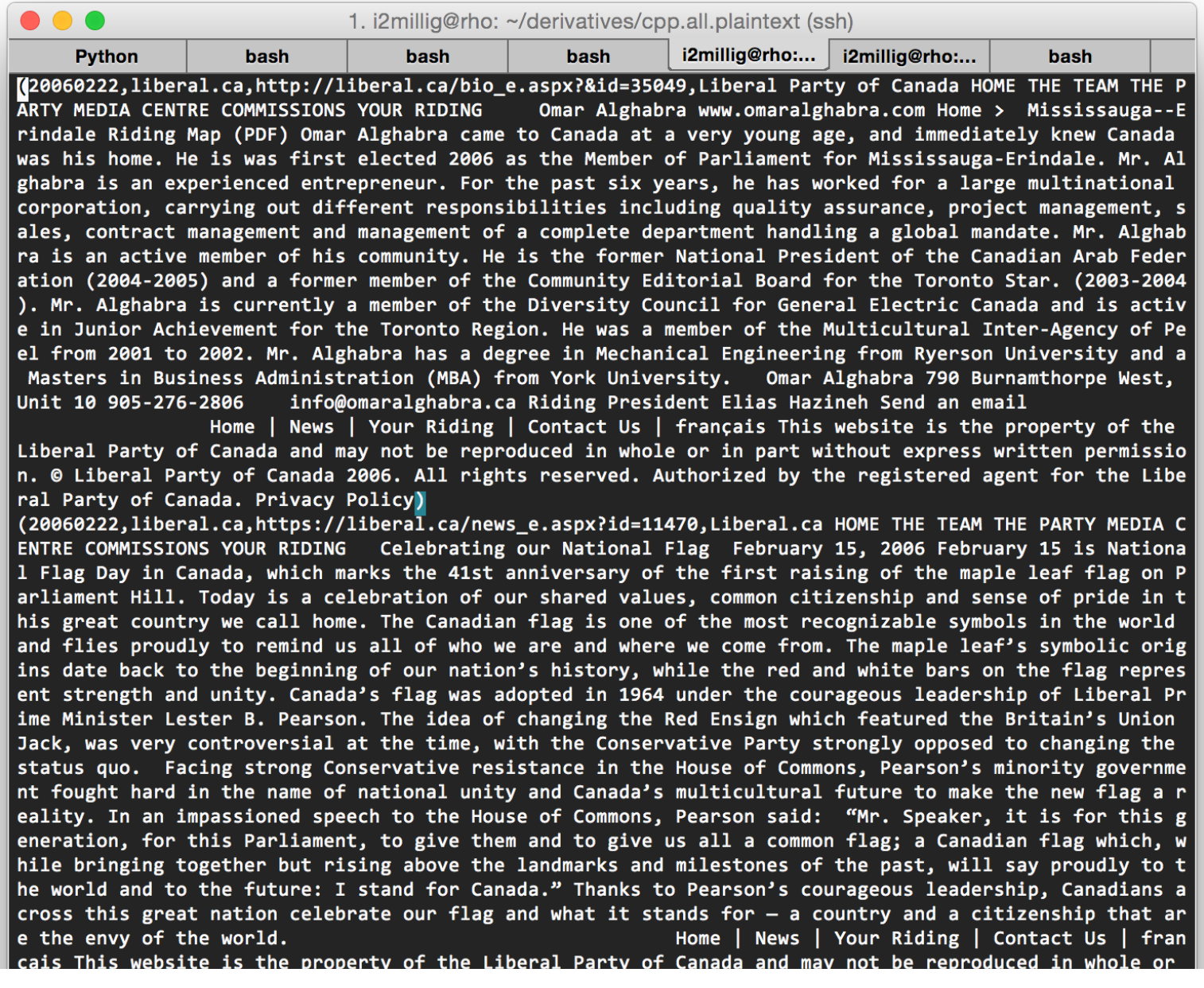
- Currently using it on the **GeoCities** and **Canadian Politics** web archives



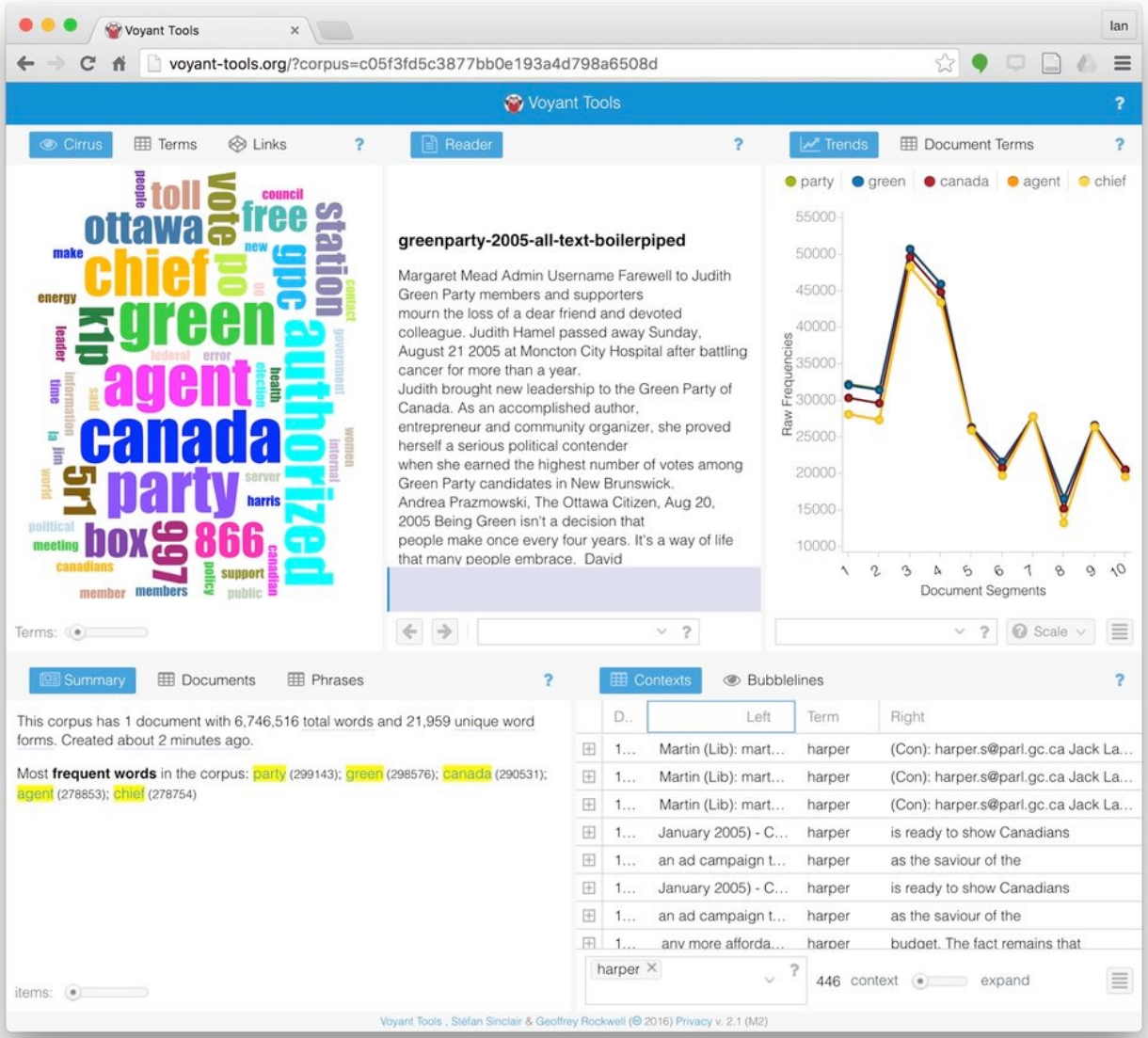
docs.warcbase.org



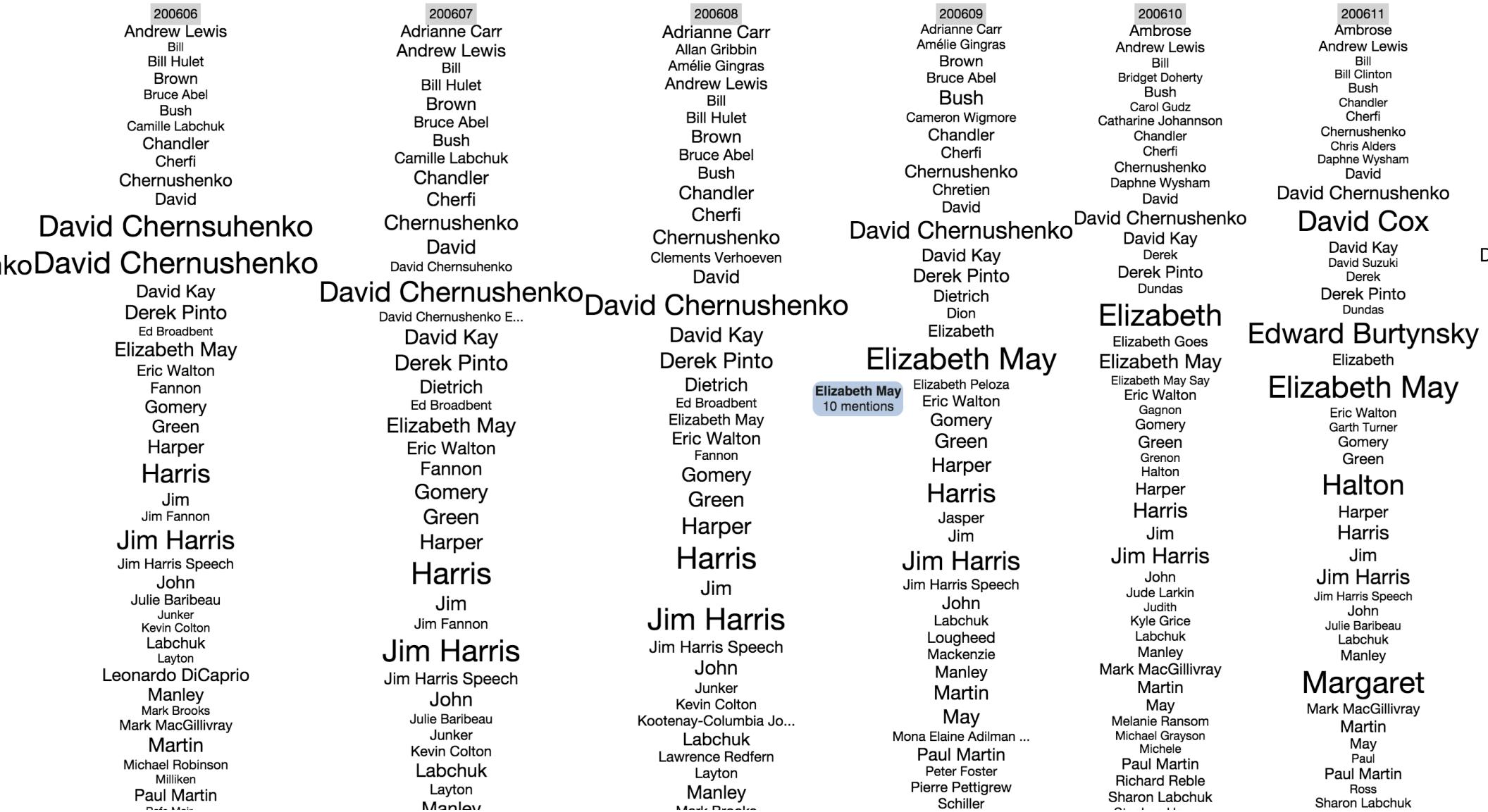
Extract all Text



Extract all Text

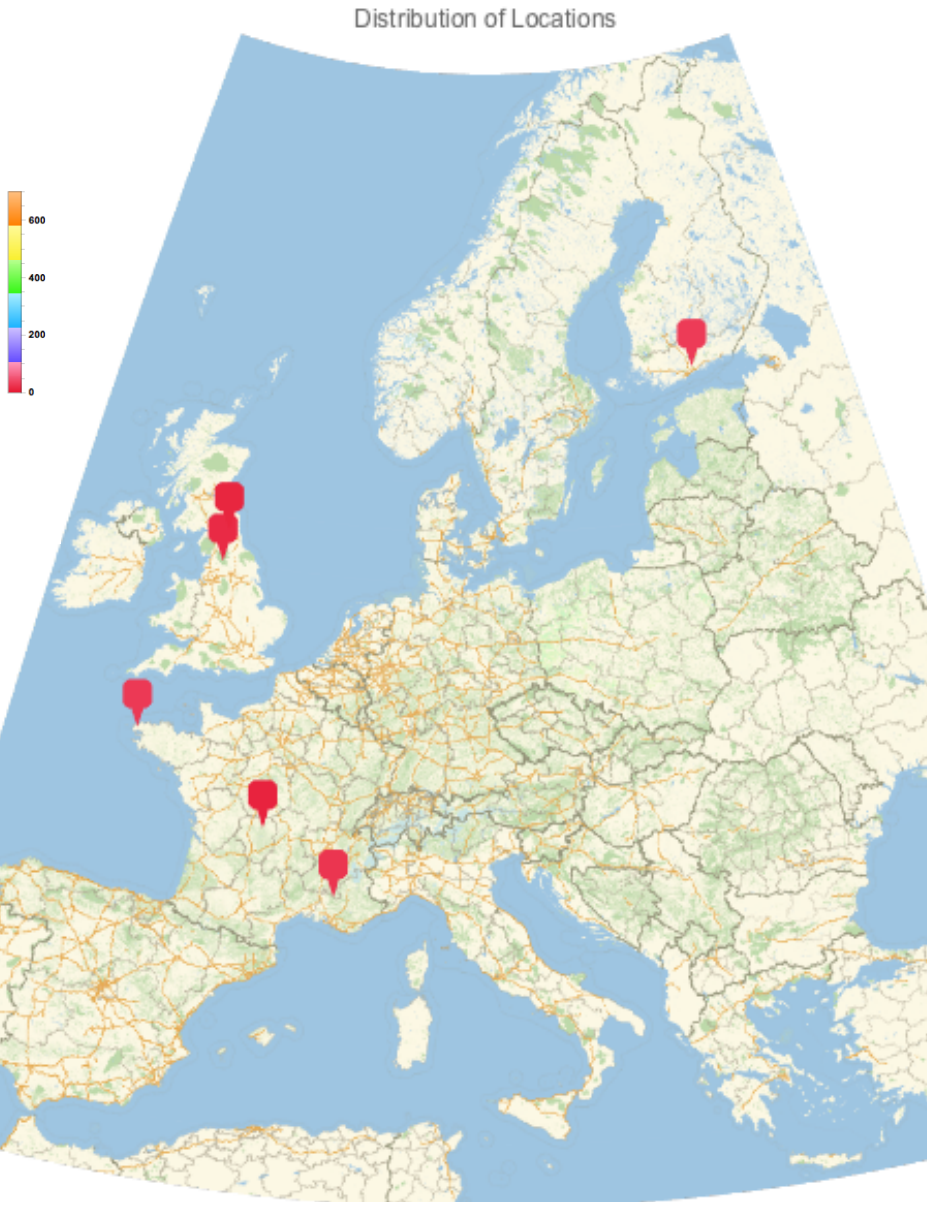


Extract Entities

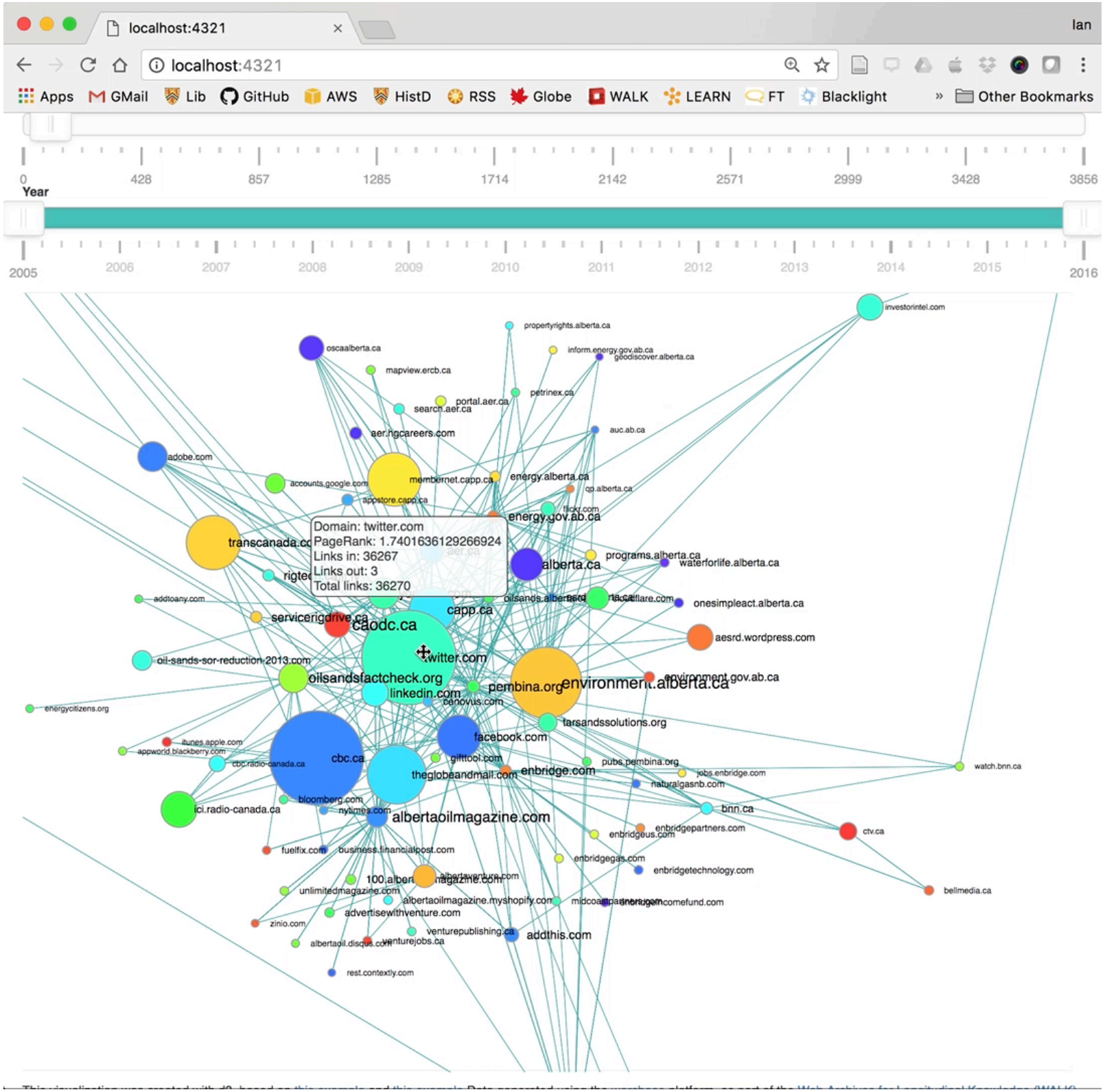


Extract Entities

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,
```


localhost:4321
Ian
localhost:4321
Apps
GMail
Lib
GitHub
AWS
HistD
RSS
Globe
WALK
LEARN
FT
Blacklight
Other Bookmarks
0
428
857
1285
1714
2142
2571
2999
3428
3856
Year
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Domain: twitter.com
PageRank: 1.7401636129266924
Links in: 36267
Links out: 3
Total links: 36270
investorintel.com
propertyrights.alberta.ca
oscaalberta.ca
inform.energy.gov.ab.ca
geodiscover.alberta.ca
mapview.ercb.ca
petrinex.ca
portal.aer.ca
search.aer.ca
aer.hgcareers.com
auc.ab.ca
adobe.com
accounts.google.com
membernet.capp.ca
energy.alberta.ca
cp.alberta.ca
appstore.capp.ca
flickr.com
energy.gov.ab.ca
transcanada.cc
programs.alberta.ca
alberta.ca
waterforlife.alberta.ca
addtoany.com
onesimpleact.alberta.ca
capp.ca
servicerigdrive.ca
caodc.ca
aesrd.wordpress.com
twitter.com
oil-sands-sor-reduction-2013.com
oilsandsfactcheck.org
environment.gov.ab.ca
linkedin.com
pembina.org
cenovus.com
energycitizens.org
tarsandssolutions.org
facebook.com
itunes.apple.com
appworld.blackberry.com
cbc.radio-canada.ca
cbc.ca
gifttool.com
pubs.pembina.org
watch.bnn.ca
theglobeandmail.com
enbridge.com
jobs.enbridge.com
naturalgasnb.com
bloomberg.com
ici.radio-canada.ca
nytimes.com
albertaoilmagazine.com
bnn.ca
enbridgepartners.com
ctv.ca
enbridgeus.com
fuelfix.com
business.financialpost.com
enbridgegas.com
enbridgetechnology.com
albertaventure.com
unlimitedmagazine.com
albertaoilmagazine.myshopify.com
bellmedia.ca
advertisewithventure.com
zinio.com
venturepublishing.ca
addthis.com
venturejobs.ca
albertaoil.disqus.com
rest.contextly.com

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.

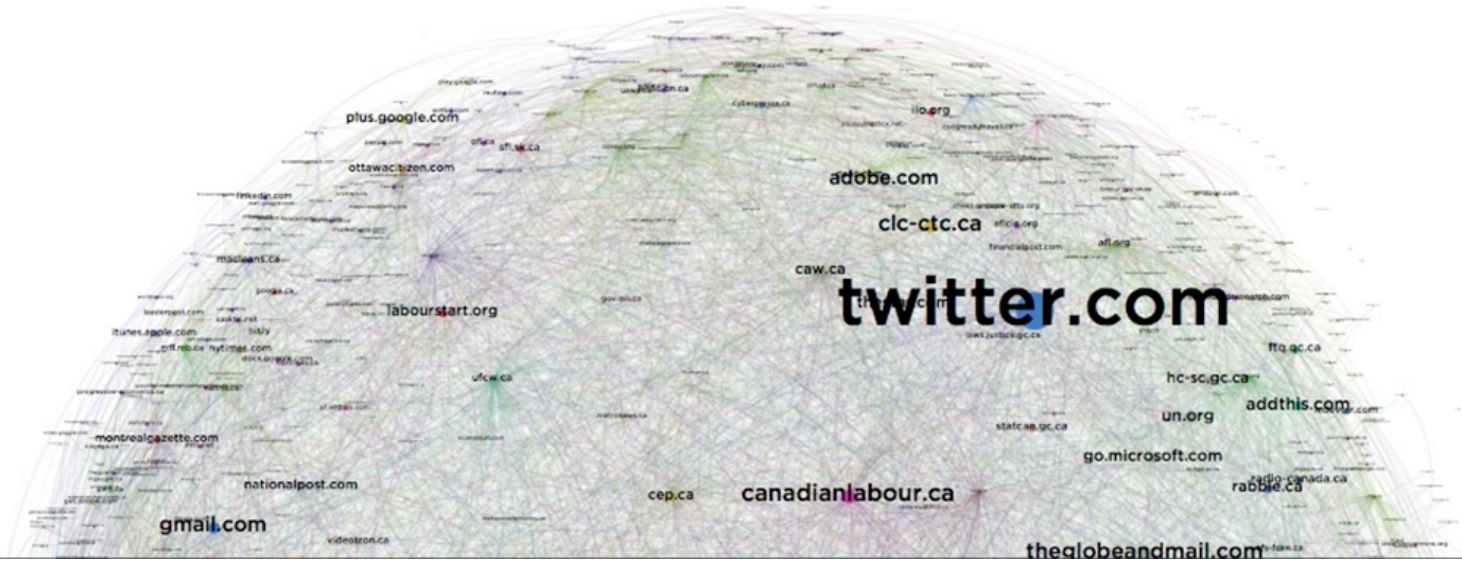
Currently, this is a prototype site providing access to one such archive, the University of Toronto's Canadian Political Parties and Political Interest Groups collection. This website allows you to search content from 50 political parties and political interest groups, from October 2005 to March 2015.

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:

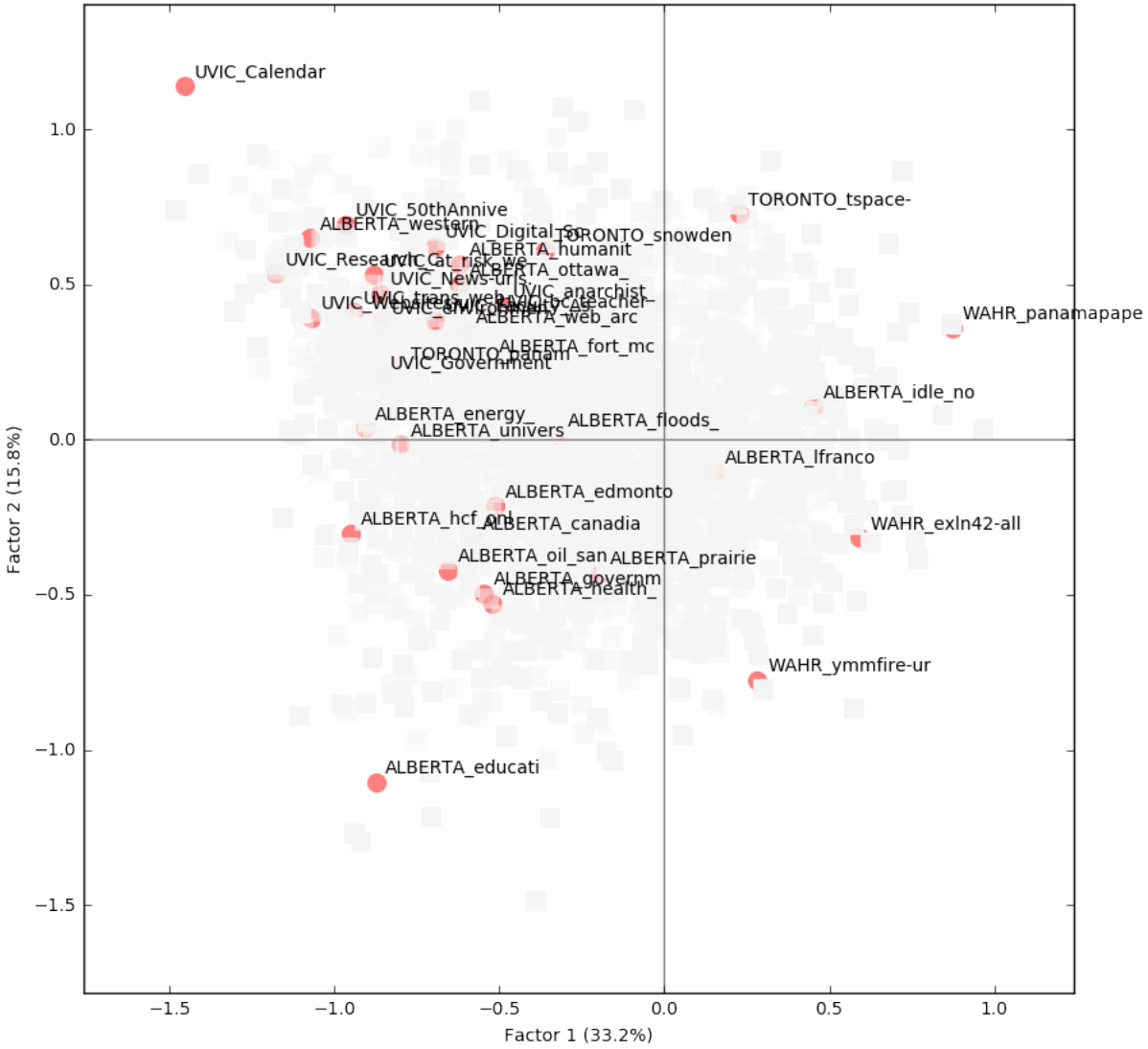
- **Basic keyword searching** [Example: "Rob Ford", only Liberal.ca]
- **Graphing trends over time** [Example: Liberal Opposition Leaders, 2005-2015]
- **Advanced search, including words in proximity to each other** [Example: environmental and tax within 25 words of each other]

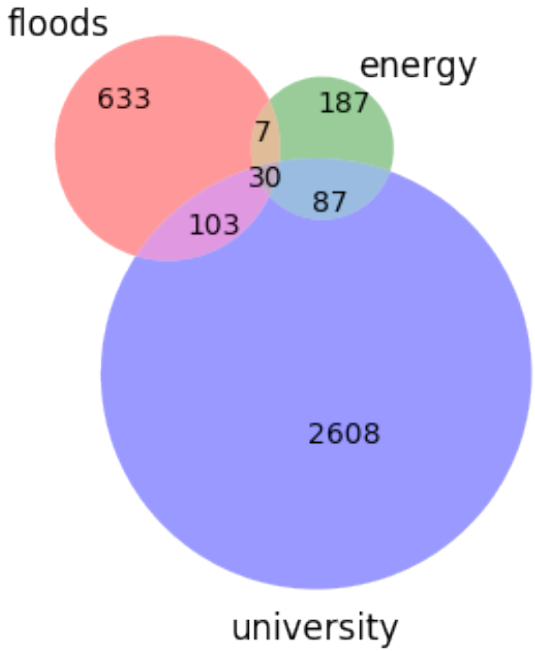
Below, here are all of the links for the entire time period, visualized below.



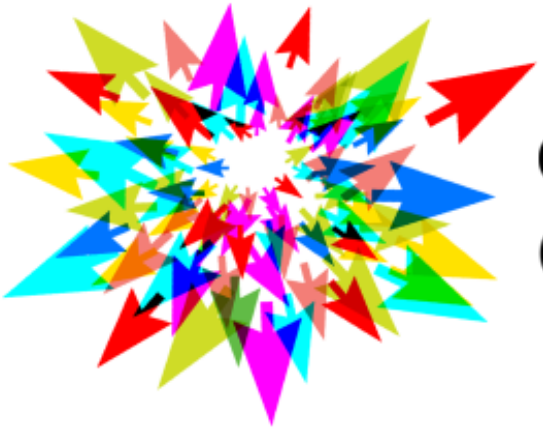
Bringing together it all into an interface like this, central hub for Web Archives in Canada.

Exploring collection coverage and curatorial models





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



Thanks!

Ian Milligan
Assistant Professor
@ianmilligan1

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History