Archives Unleashed:

Unlocking Born-Digital Sources through Interdisciplinary Collaboration

---

Ian Milligan
Assistant Professor
@ianmilligan1

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Thanks

Today's Talk

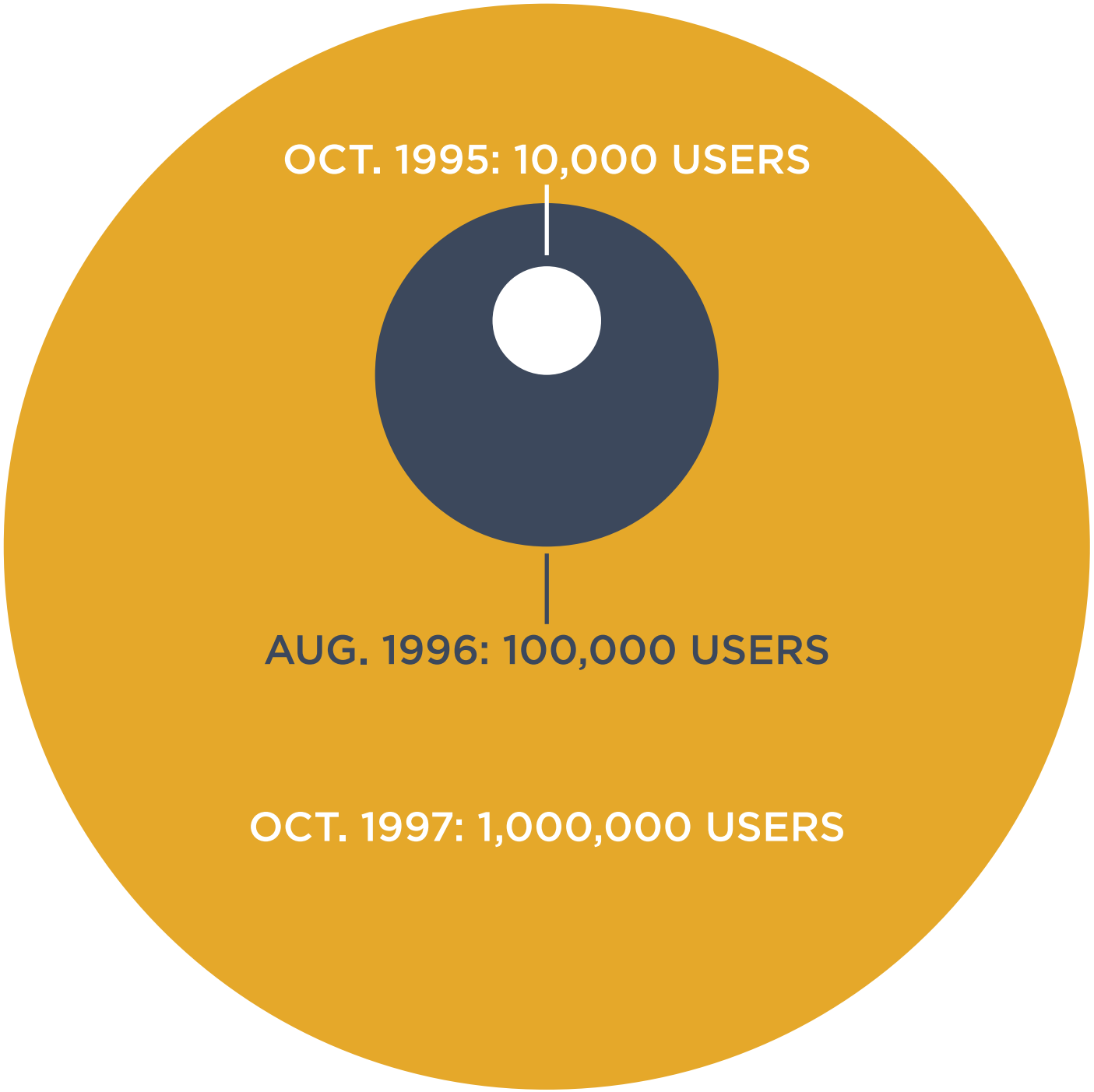
- The Problem
- The Fears
- The Team-Based Solutions
  - Computer Science
  - Community
  - Library

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:



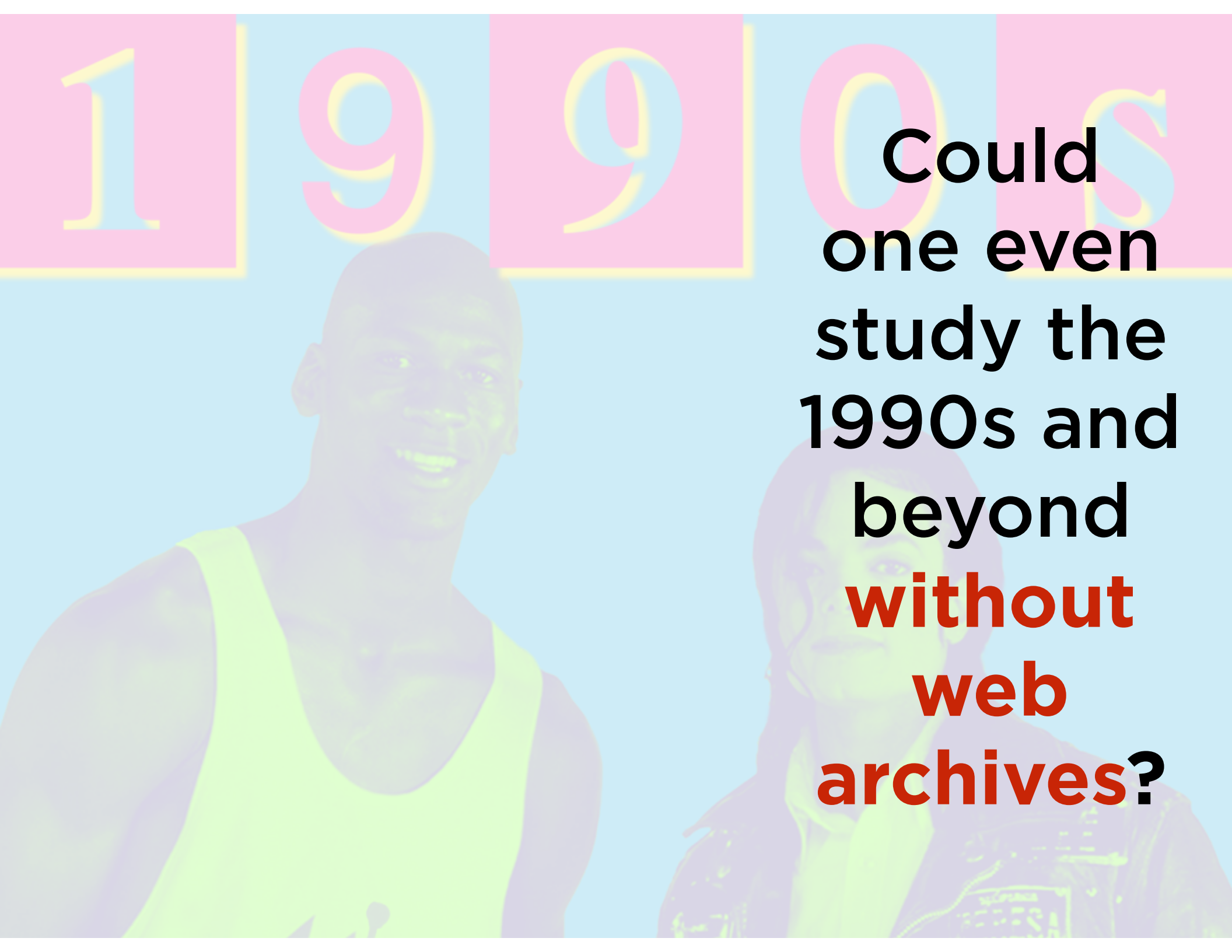
This is a scale that boggles the mind - compare it to the Old Bailey (197,745 trials between 1674 and 1913)

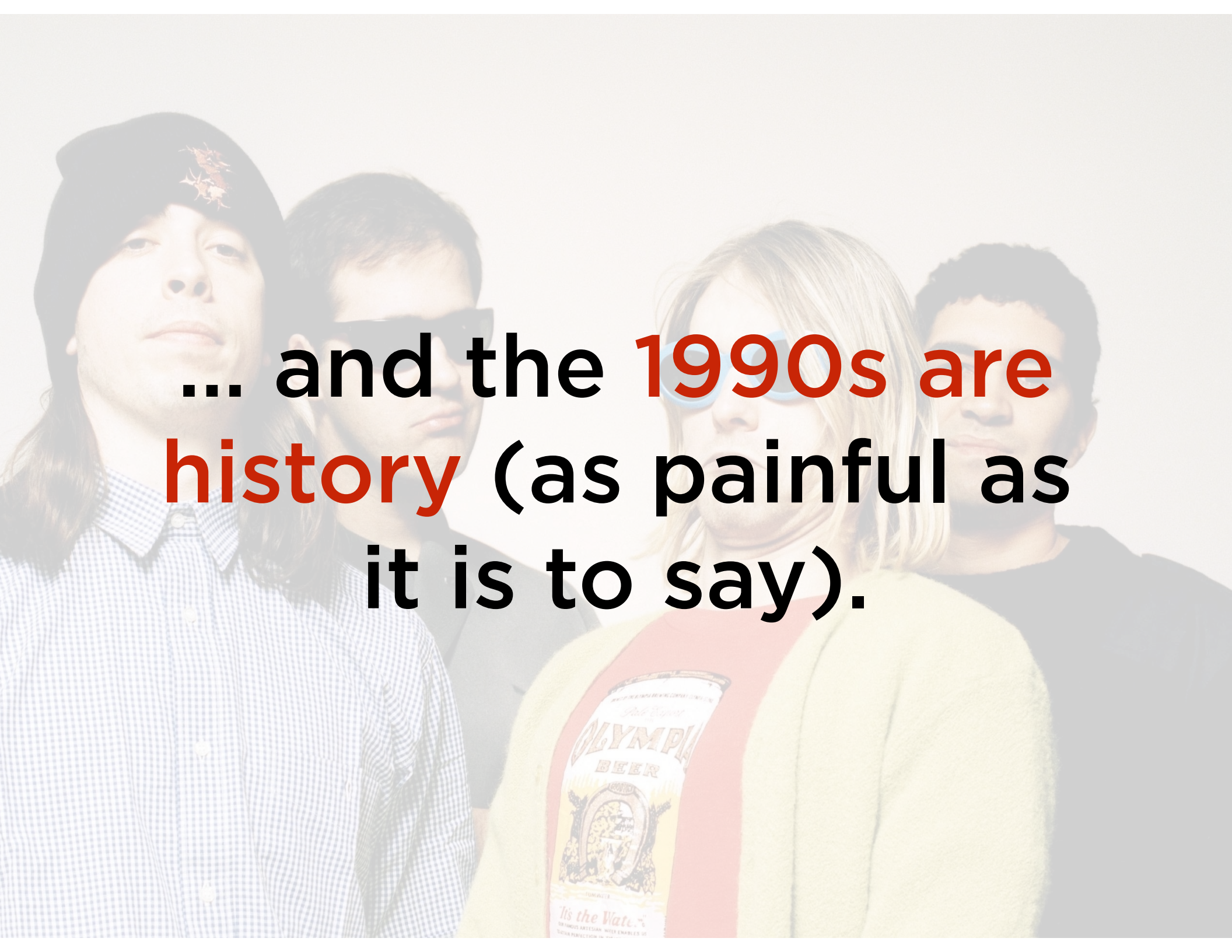

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
374

Scarcity

~~Scarcity~~

Abundance


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

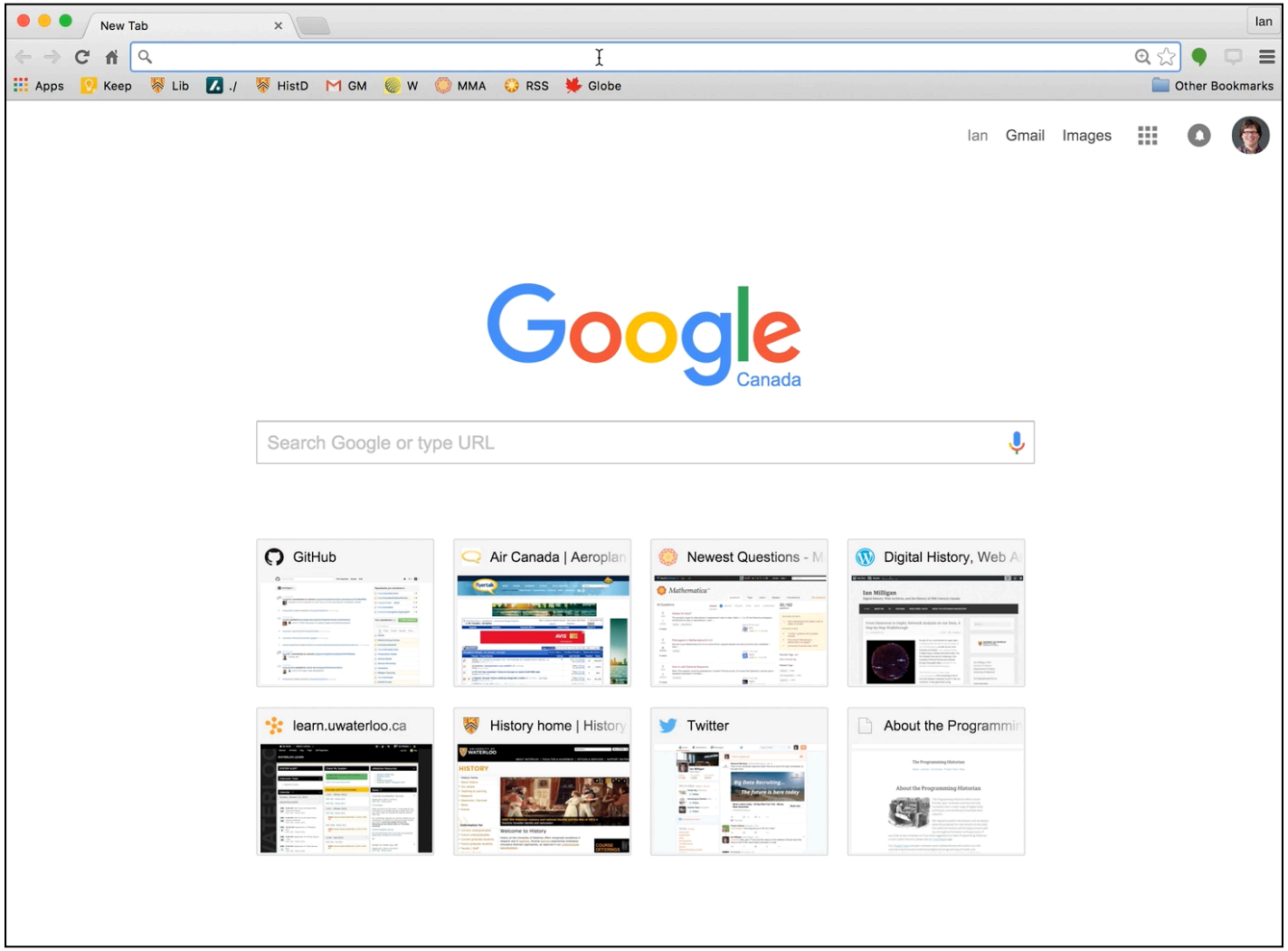


... and the 1990s are history (as painful as it is to say).

And I have fears

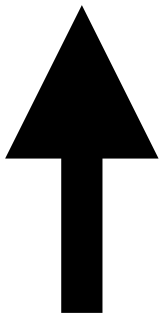
The decisions we make today will lay the foundations for how we work with born-digital cultural heritage.

Nightmare Scenario



This won't be enough!

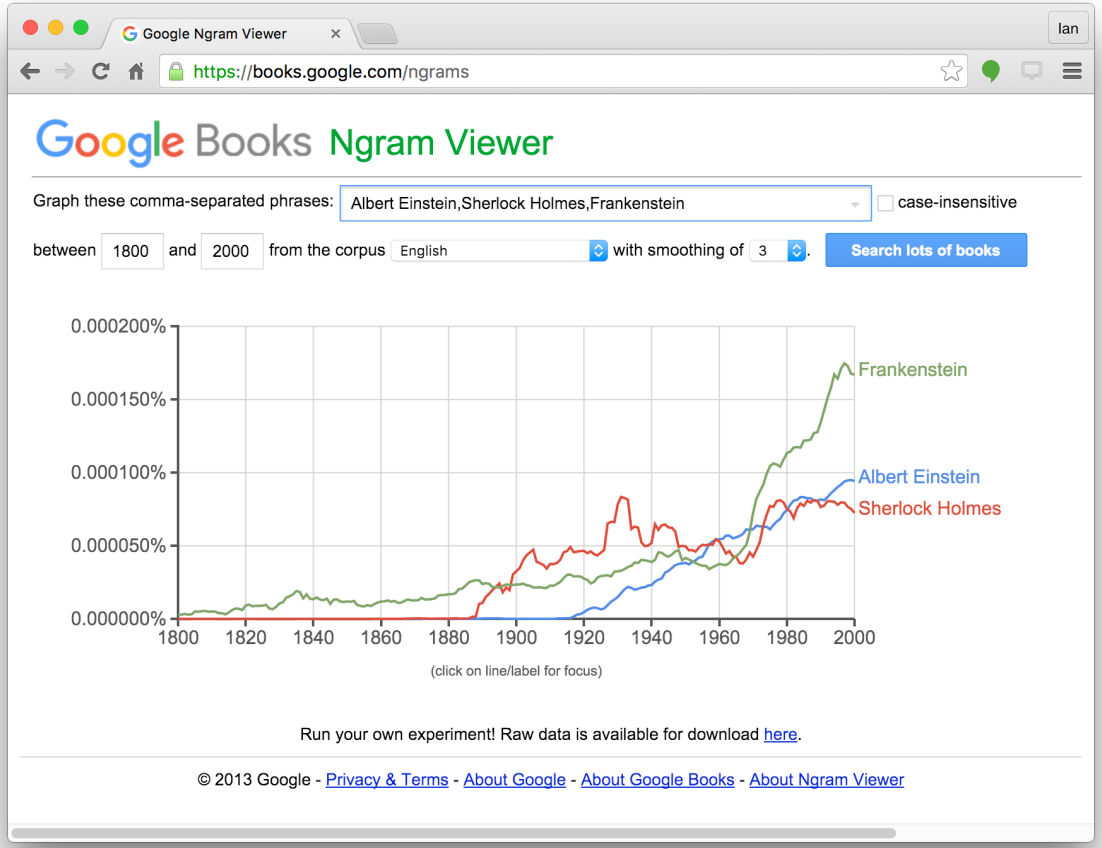
… but what will our search engines look like?



We can’t let THIS write our histories

Nightmare Scenario

- Historians rely uncritically on **date-ordered or algorithmically-ranked keyword search results**, putting them at mercy of search algorithms they do not understand



My deepest fear:

Historians are completely left out of post-1996 research, letting everybody else do the work (a la Culturomics project/ Science magazine article);

Our profession gets left behind...

The historians who came to the meeting were intelligent, kind, and encouraging. But they didn't seem to have a good sense of how to wield quantitative data to answer questions, didn't have relevant computational skills, and didn't seem to have the time to dedicate to a big multiauthor collaboration. It's not their fault: these things don't appear to be taught or encouraged in history departments right now.

- Erez Leiberman Aiden and Jean-Baptiste Michel

What can we do to access this information and avoid my nightmare?

I can’t do it alone

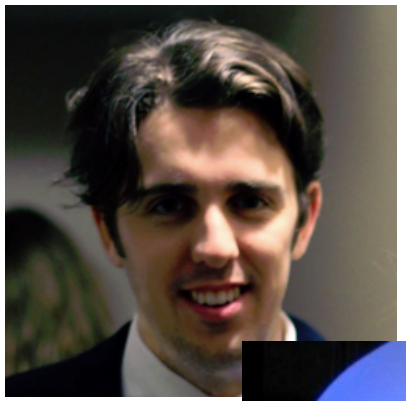
Need to bring web archives into conversation with the broader scholarly community

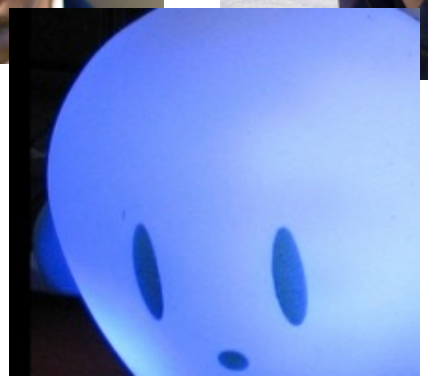

Teamwork

Web Archives for Historical Research

Historians

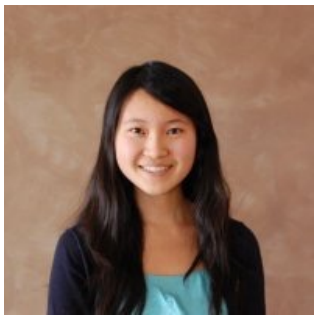






Computer Scientists





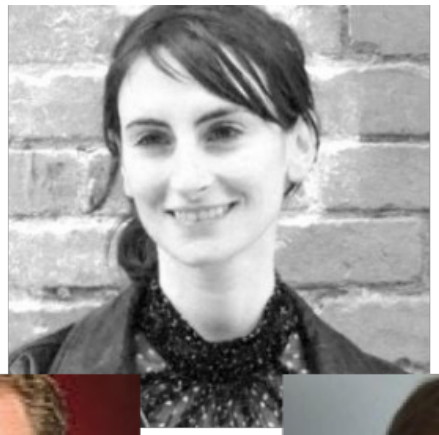
Librarians

Networks

Team Datahon

Historian

Communication



Computer Science

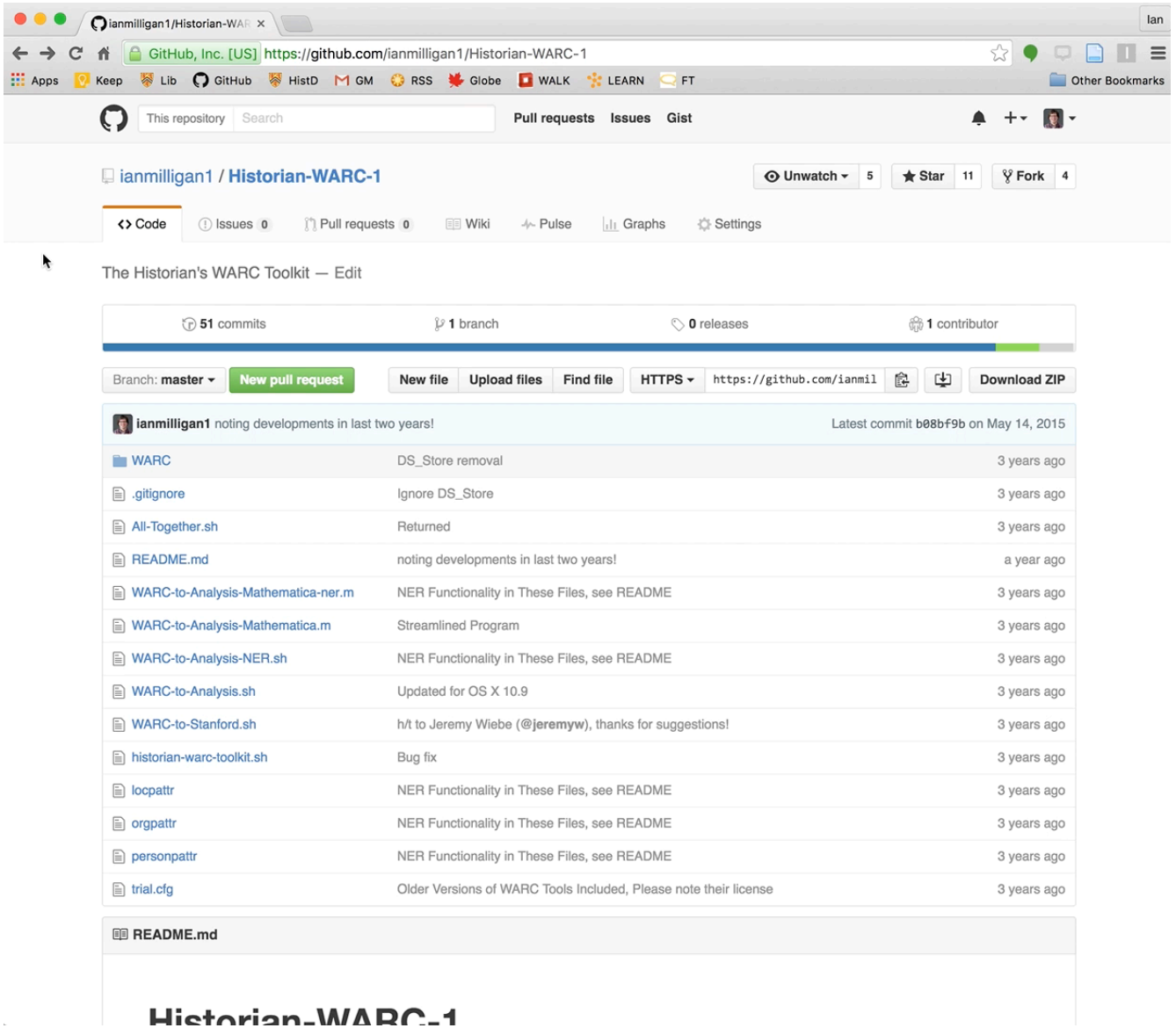
Communication

Library Services

Why can’t we do it alone?

- Varied skillsets
  - Some specialize in sustainability;
  - Others user interfaces;
  - Others security;
  - Others web archives;
  - And some various combinations of the above…

What happens when you do it alone?



Going it alone?

- Some values
  - Learning great skills!
  - Having the chops to talk to other people.
- But..
  - Shoddy code that’s not sustainable;
  - Not optimized for large datasets;
  - Missing the diversity of perspectives

Making Teams Work

- **Everybody needs to be happy**
  - Students need to be paid and represented on publications;
  - 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

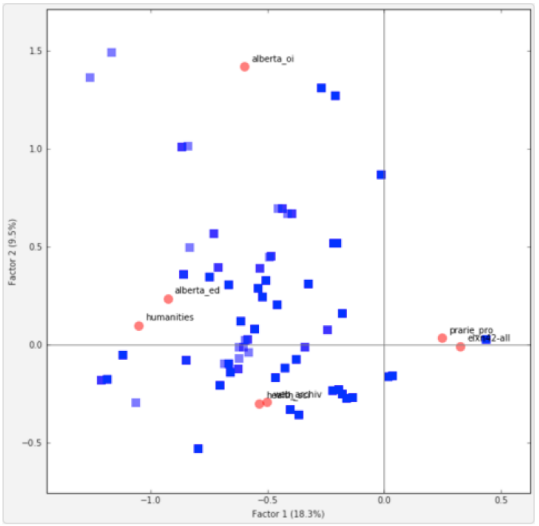
+ Invite People

#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.

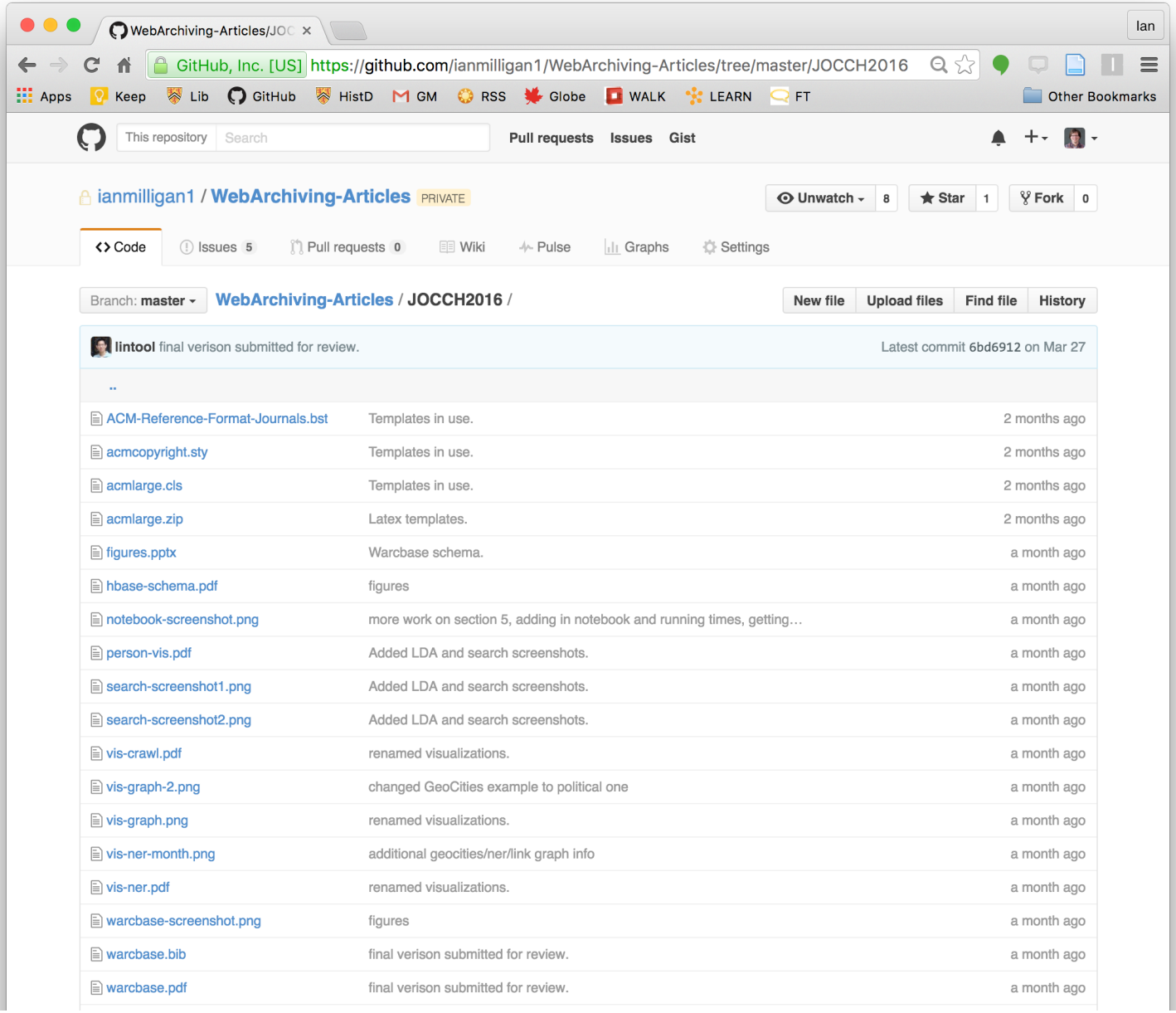
nruest 7:04 PM
@ianmilligan1: @ianmilligan1 `/data/cpp`
all copied over
1

ryandeschamps 9:44 PM
Excellent! Thanks so much!

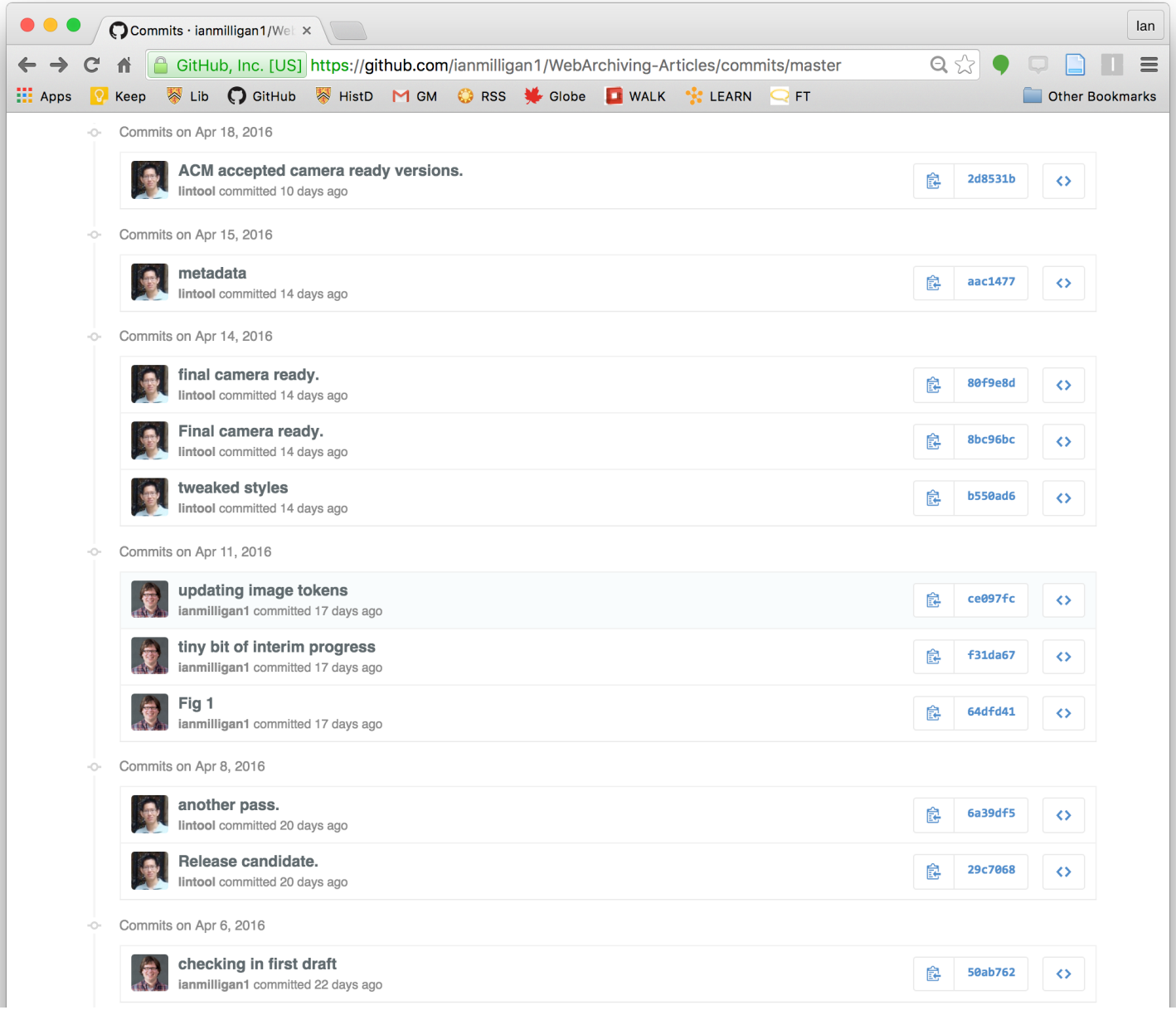
May 20th

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



Constant Communication

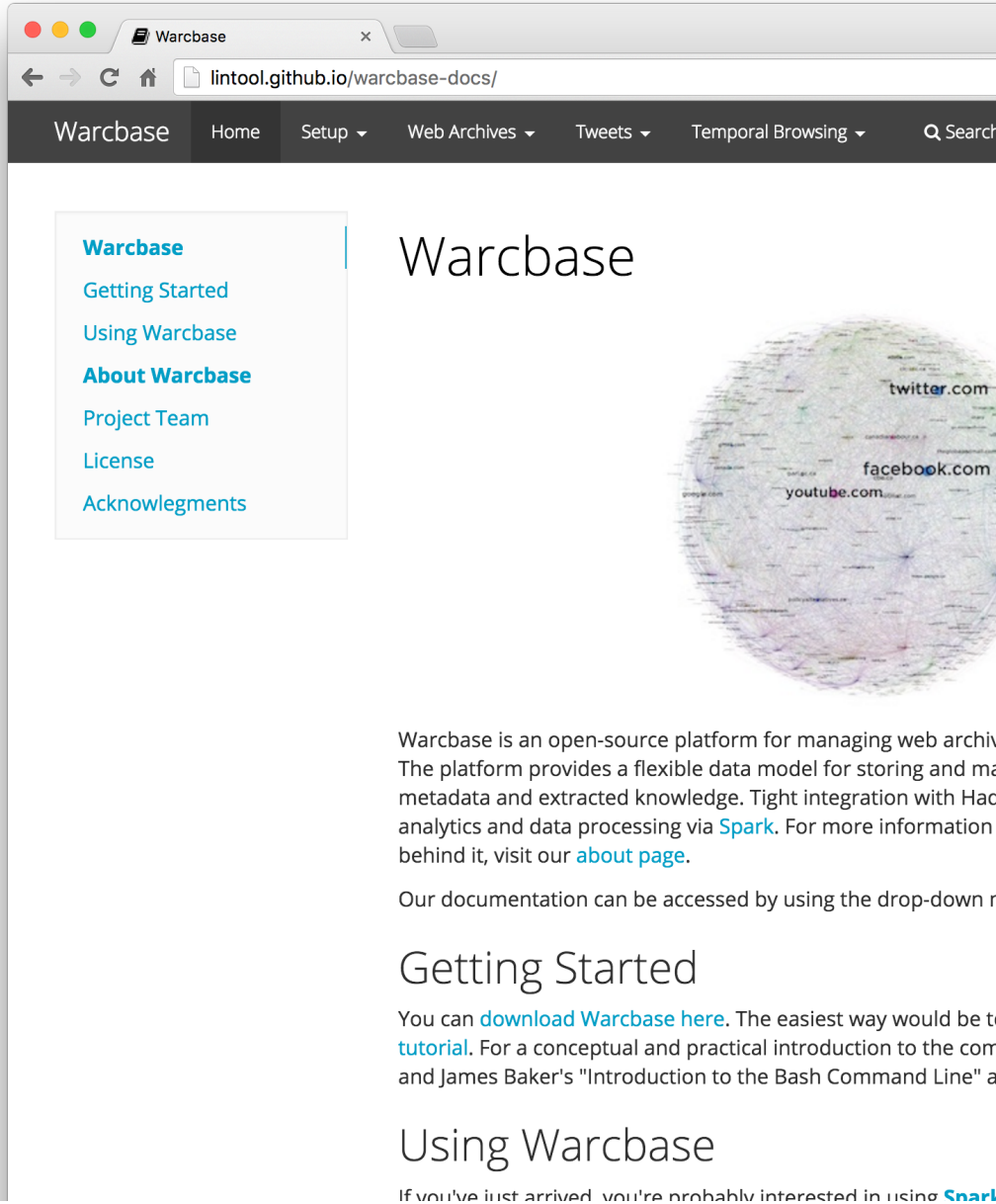


Three Projects

- **Computer Science-Driven**
- **Community-Driven**
- **Library-Driven**

Case One: Warcbase

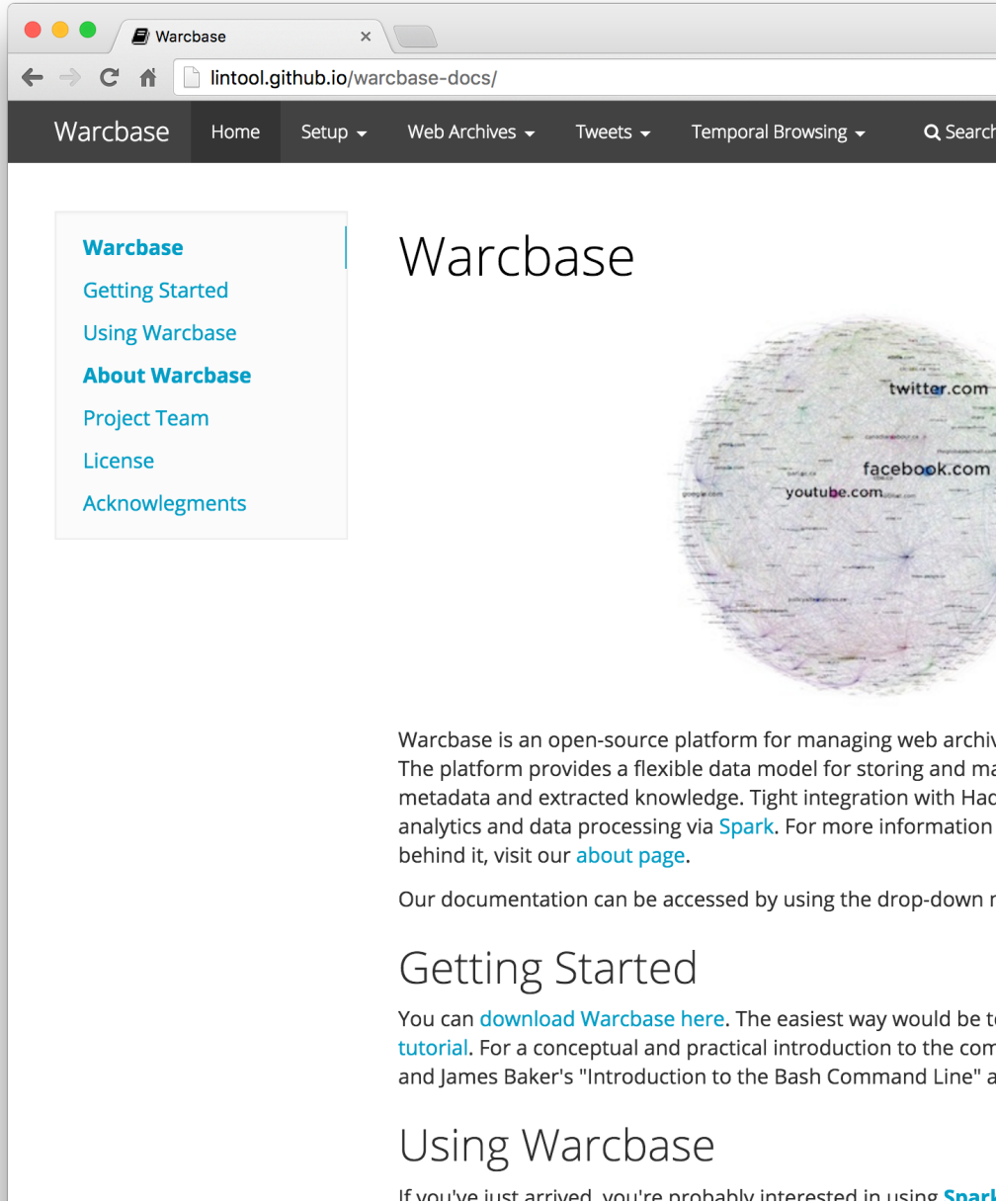
- **Web Archive Analytics**
- **Making Web Archives play with Digital Humanists**



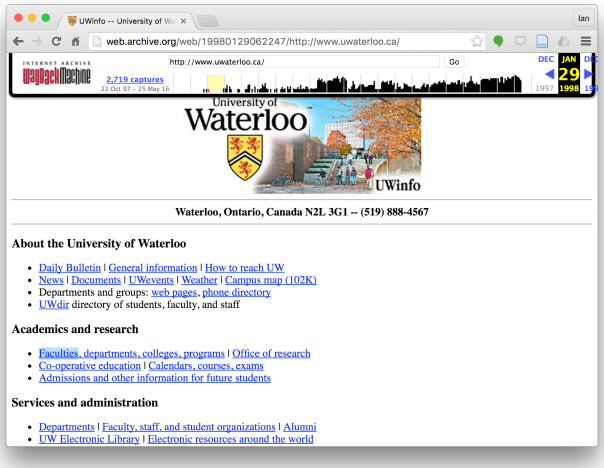
Case One: Warcbase

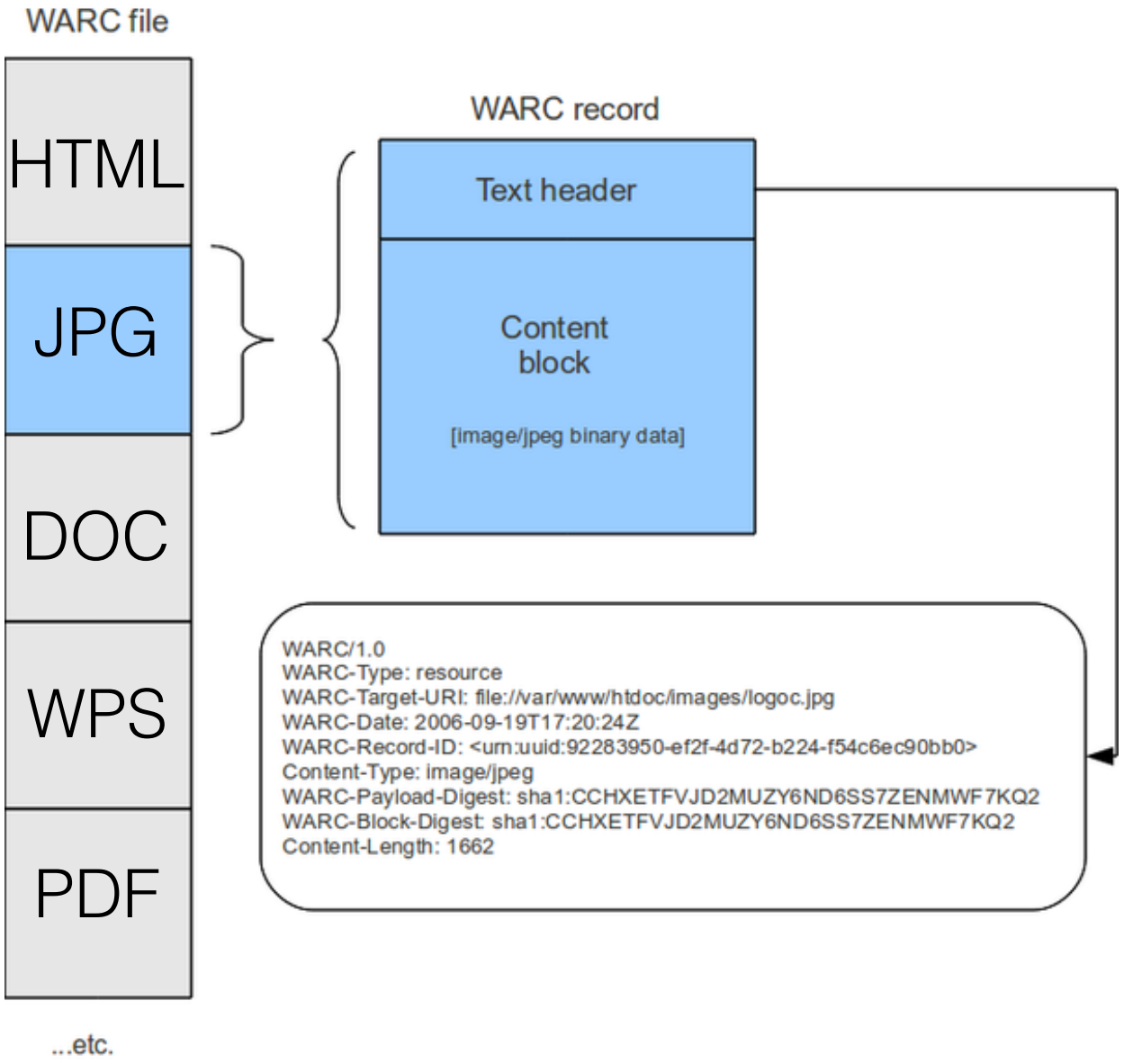
- **Jimmy Lin** (main developer, CS/lead), **Ian Milligan** (co-lead, history), **Jeremy Wiebe** (history/PhD), **Alice Zhou** (computer science, undergrad), **Youngbin Kim** (computer science, undergrad), **Nick Ruest** (librarian @ York)

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



Why Warcbase?

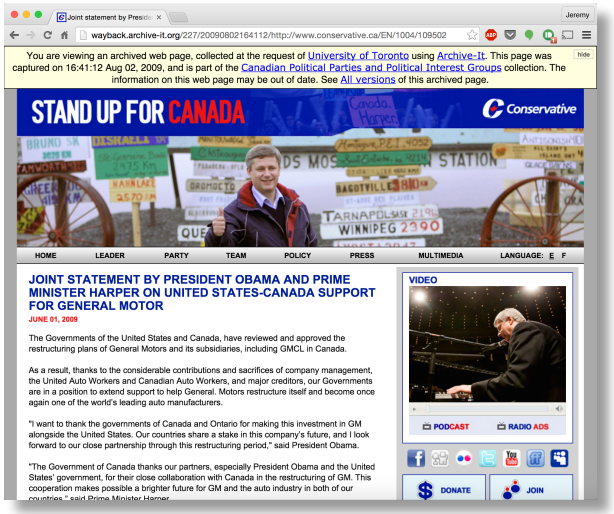




Warcbase

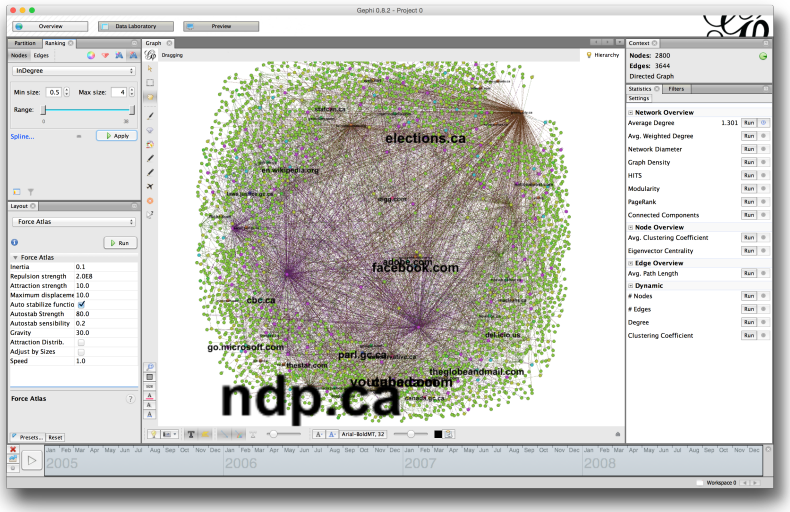
An open-source platform for managing web archives

http://warcbase.org



Two main facets

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

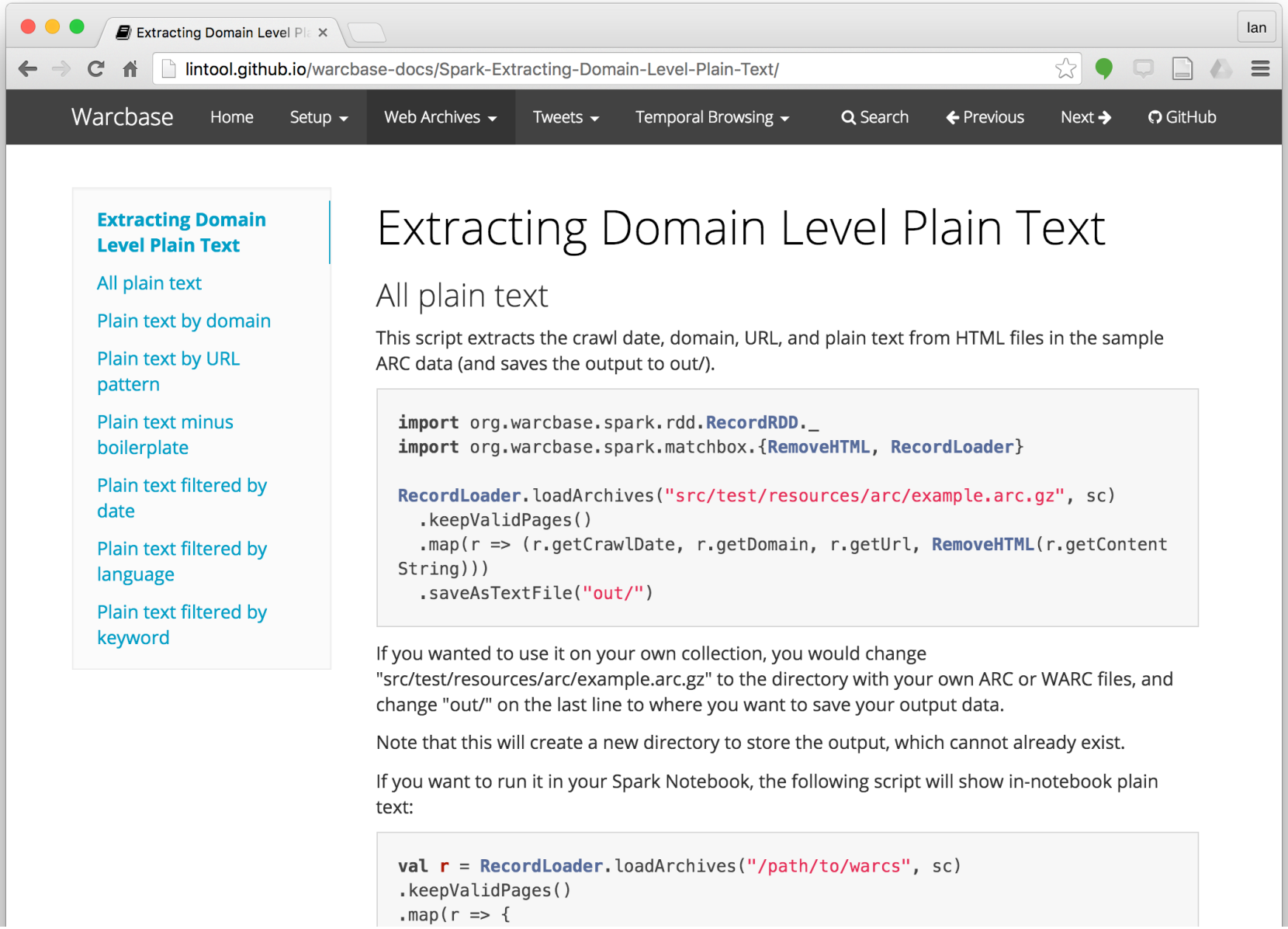


Funded by Mellon, SSHRC, NSERC, and Government of Ontario.

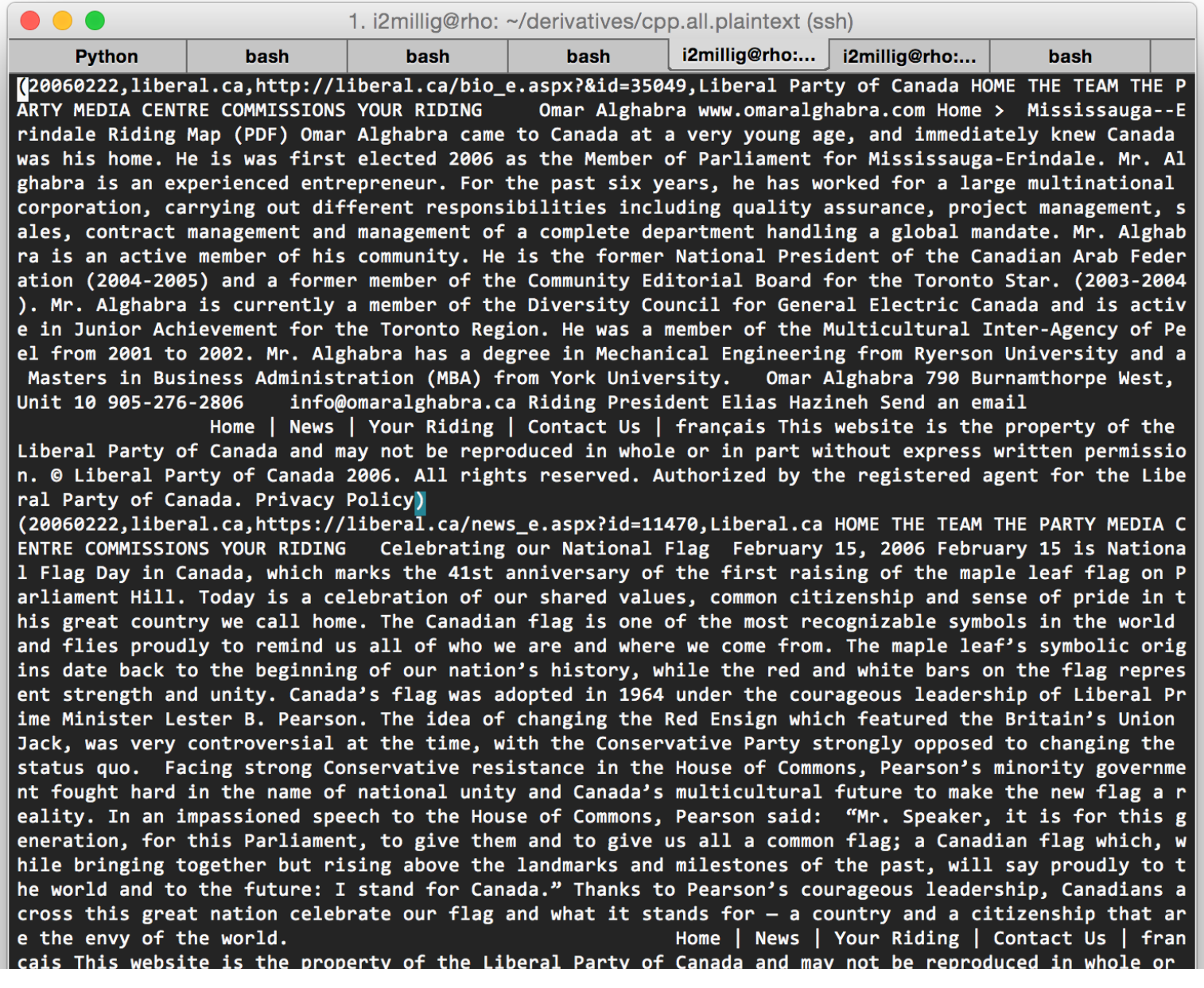
Warcbase

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

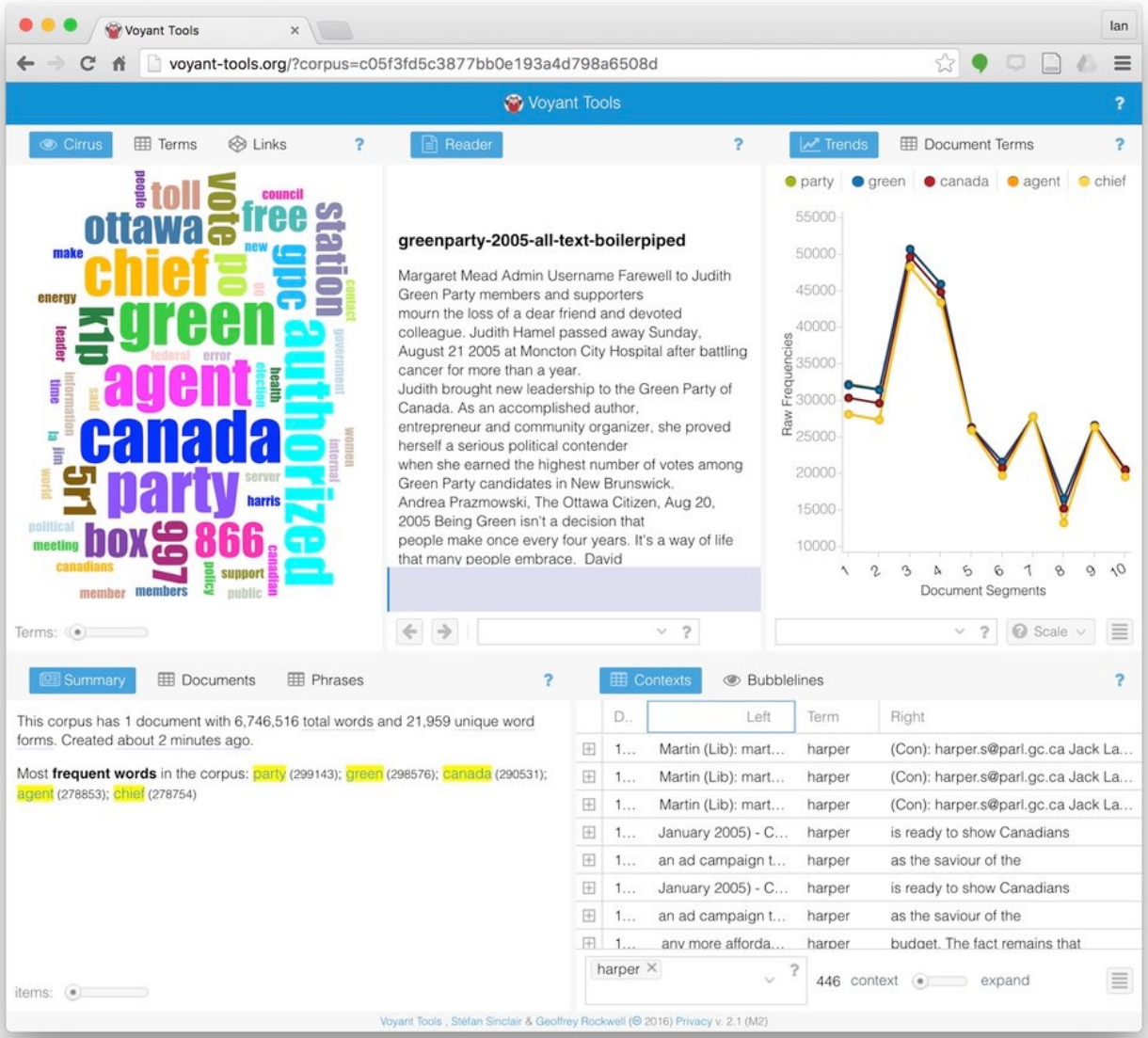
docs.warcbase.org



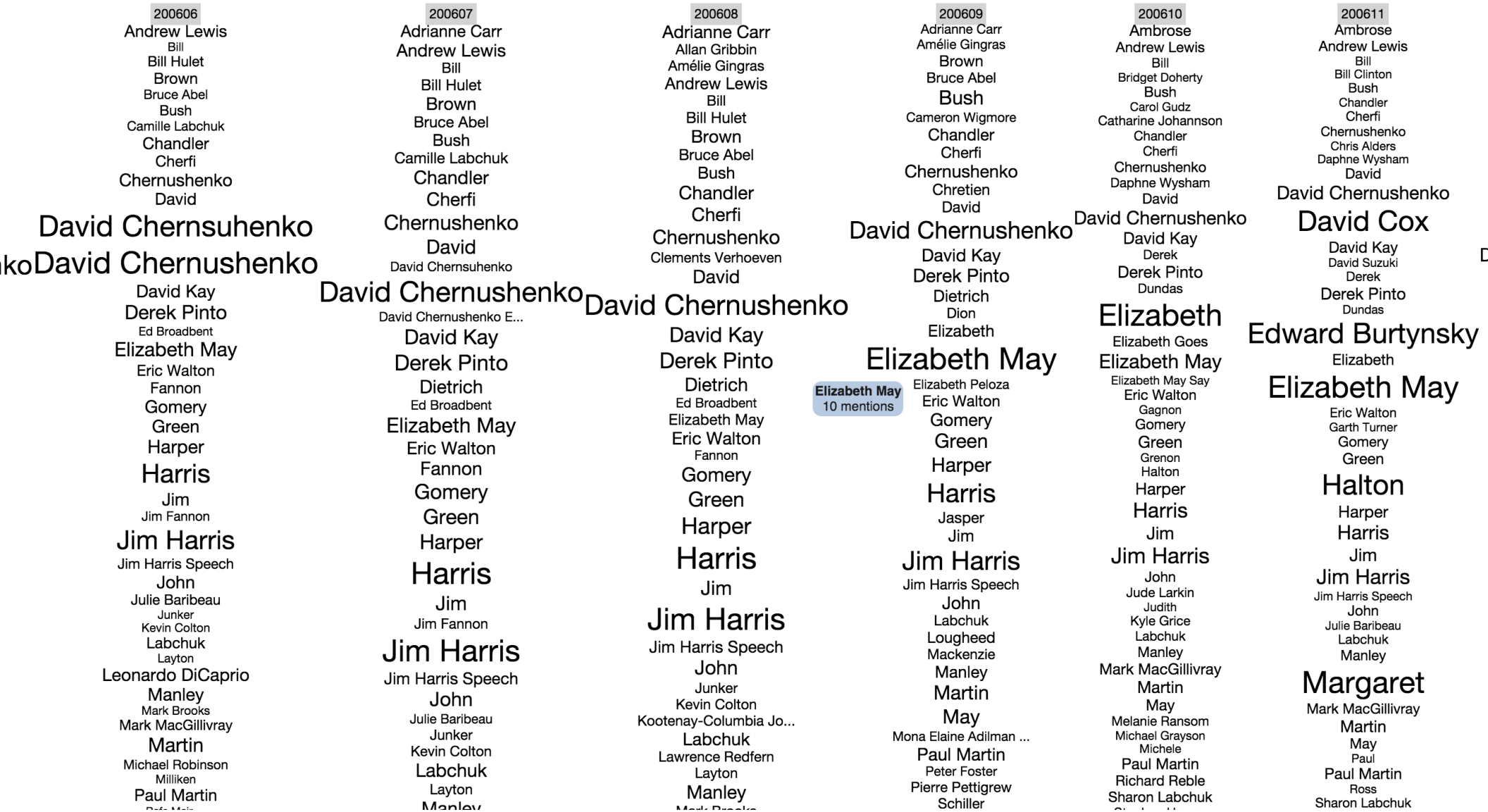
Extract all Text



Extract all Text

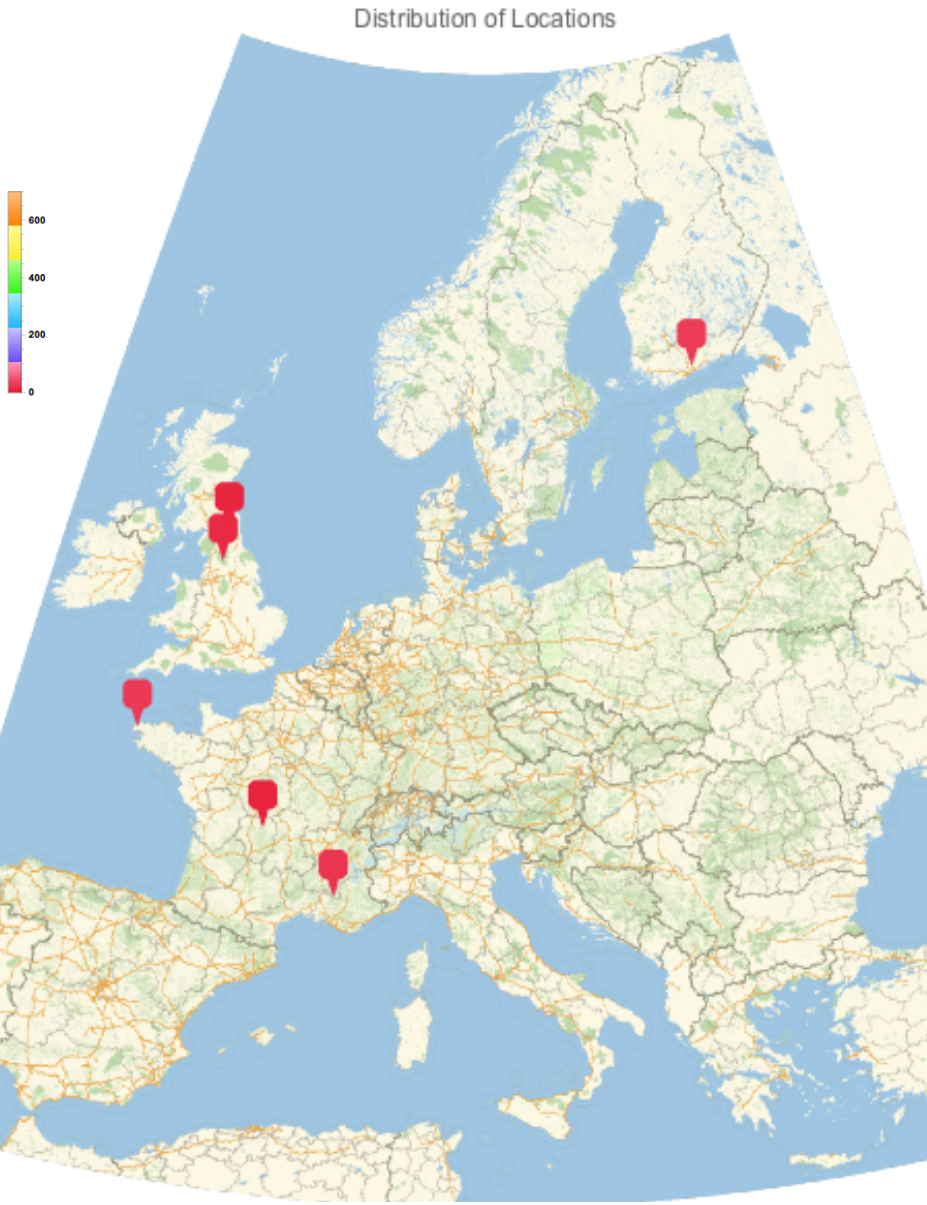


Extract Entities



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,
```

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

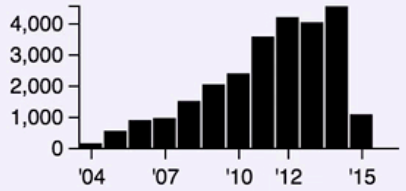
Ian Milligan

729 collaborators

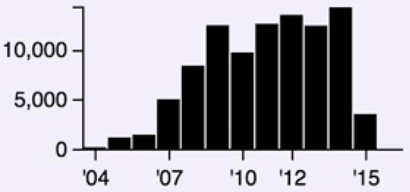
121,854 emails

My Stats | Top Collaborators

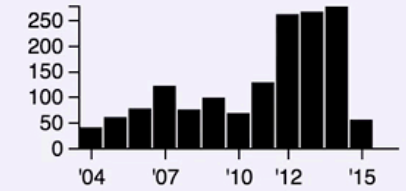
Emails Sent

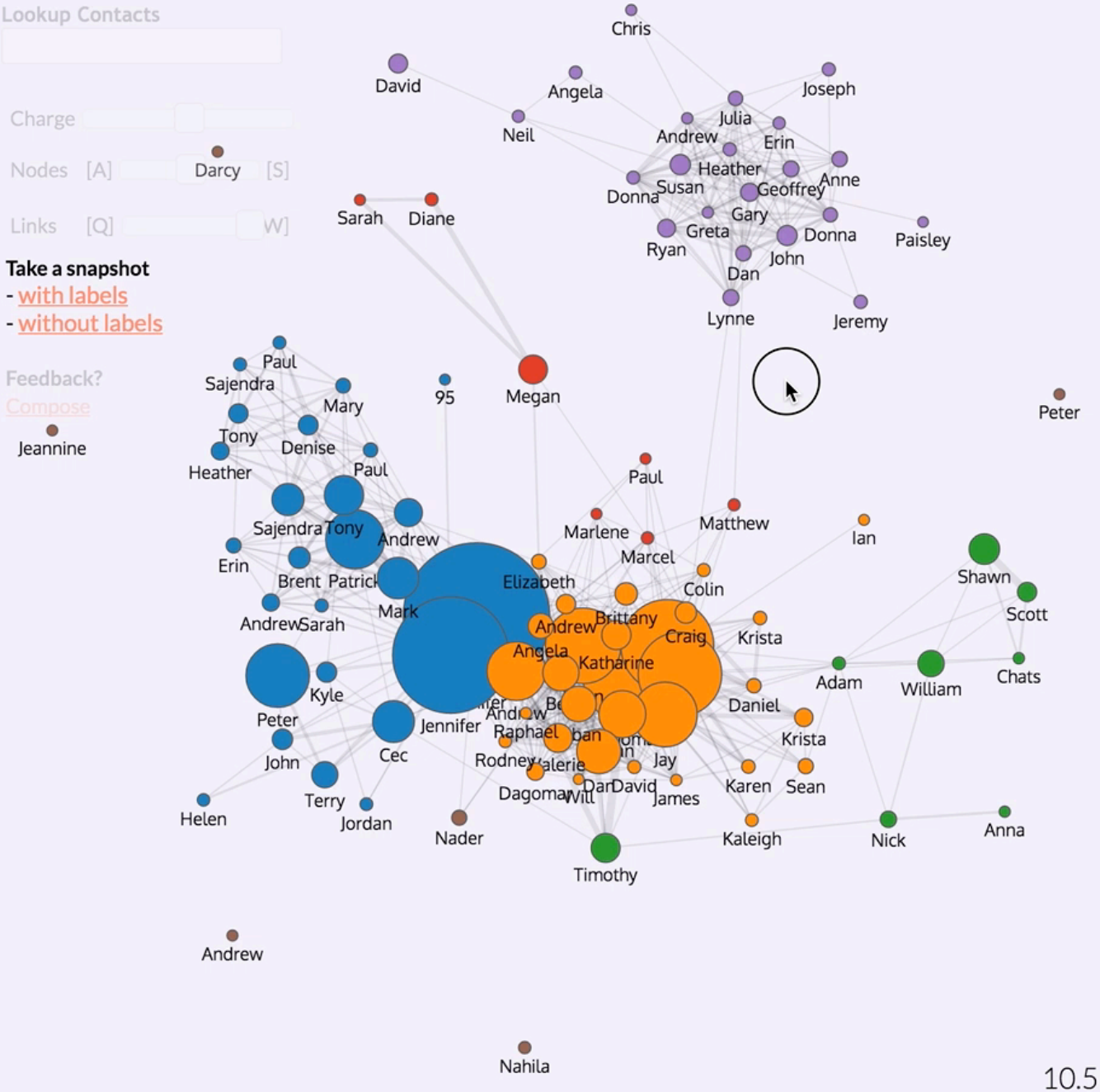


Emails Received



New Collaborators





10.5 years

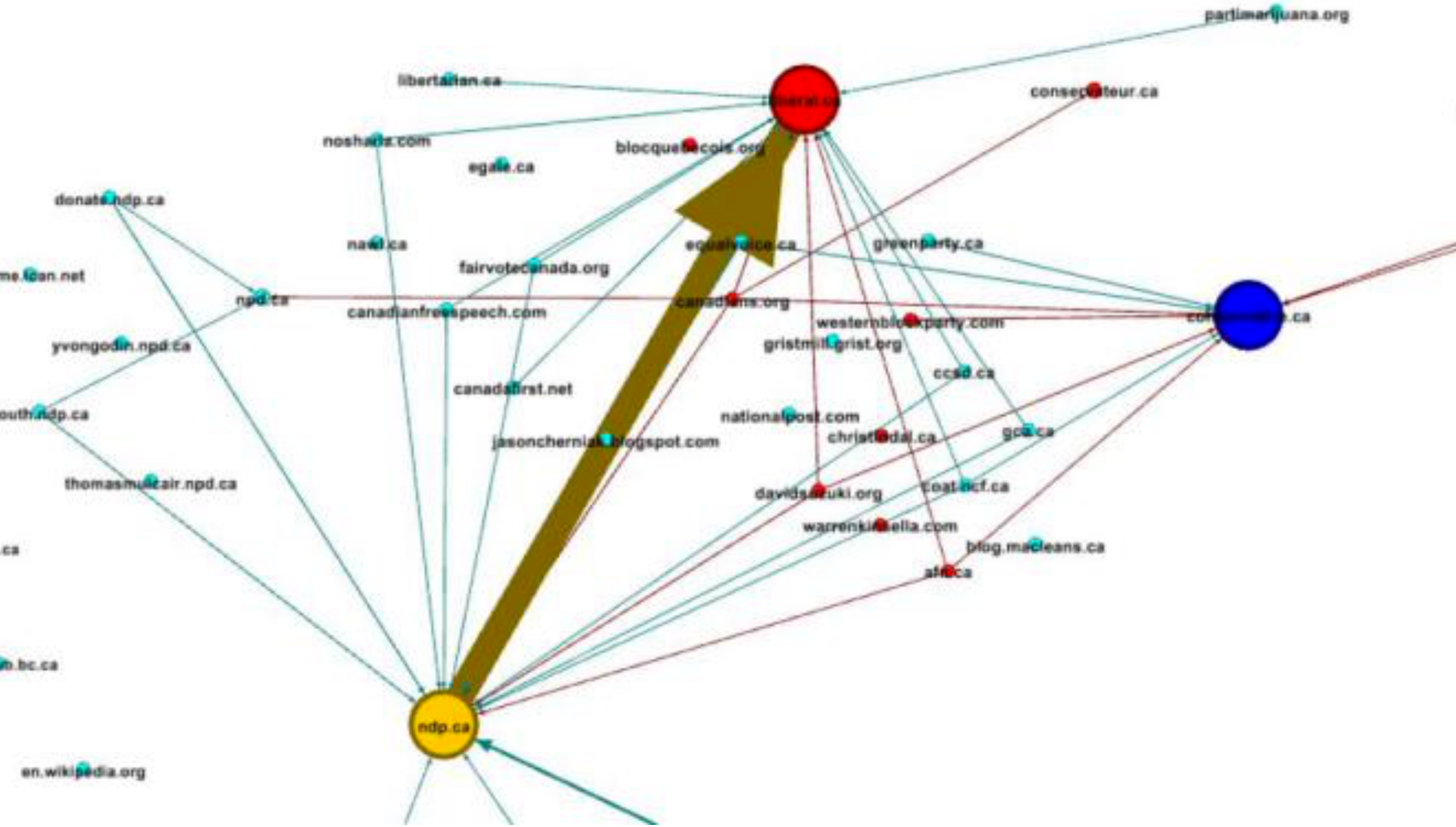
26 Sep 2004 - 12 Mar 2015

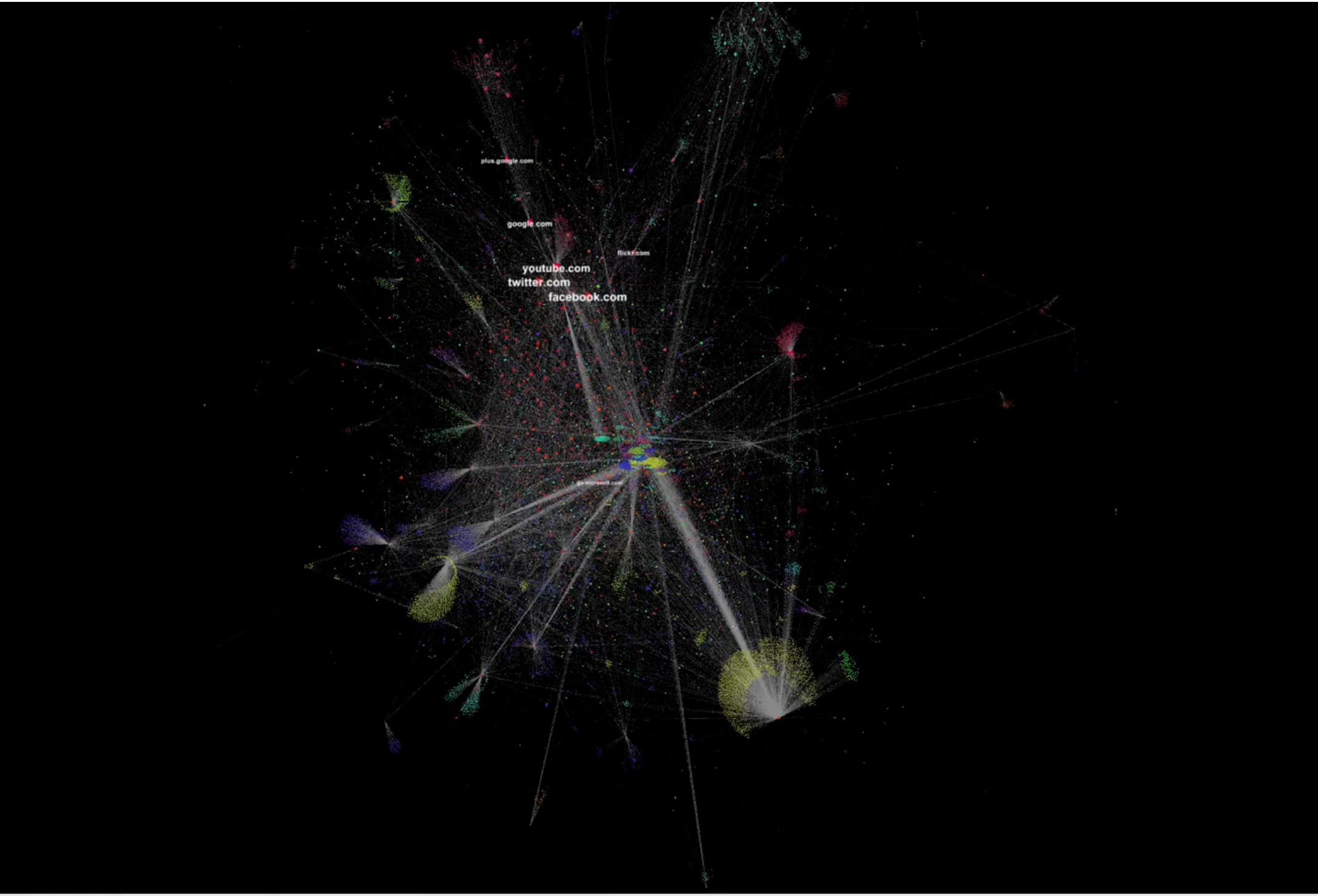
All • Past Year • Past Month • Past Week

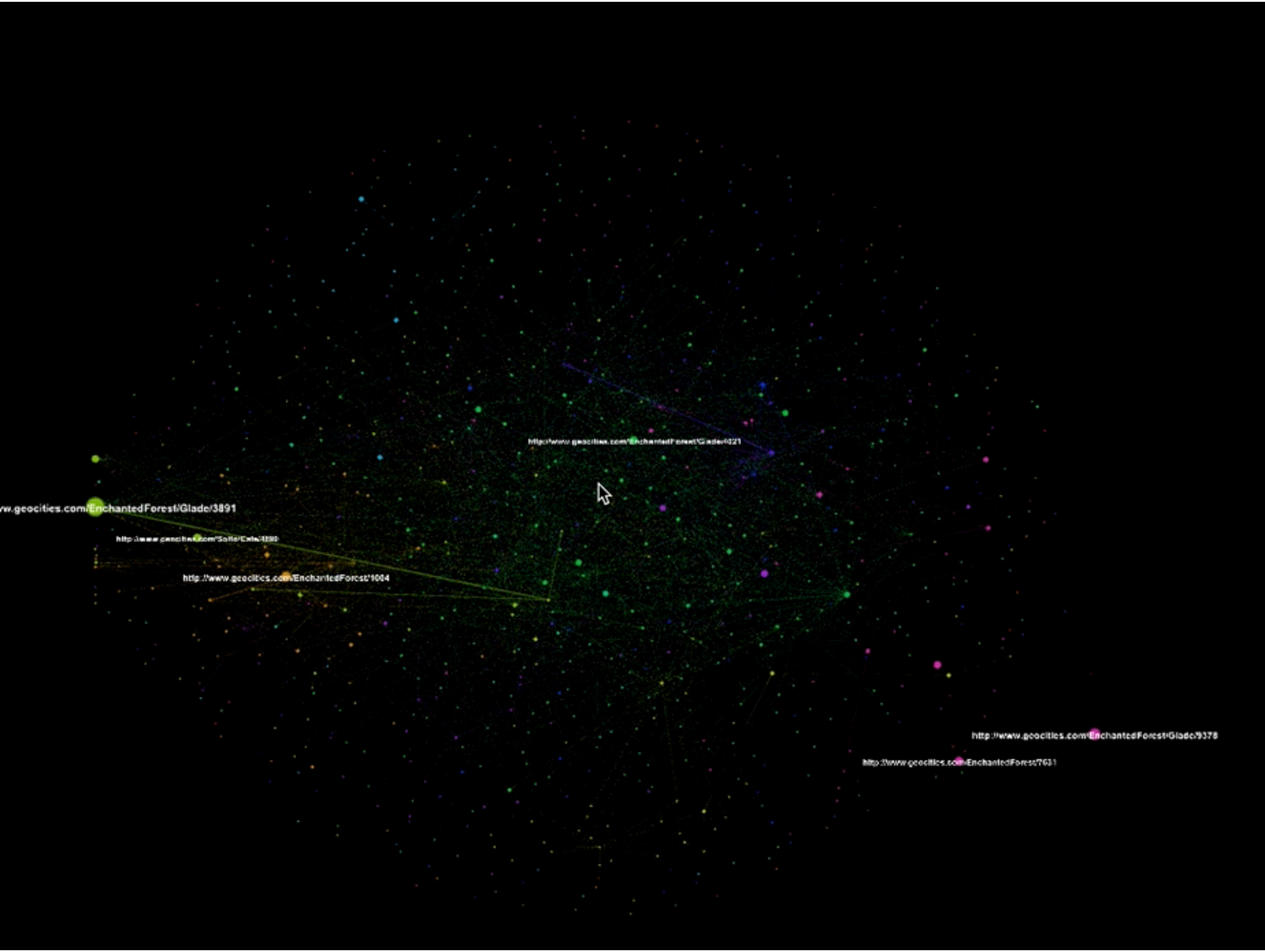

The **National Security Agency** cares more about metadata than content.

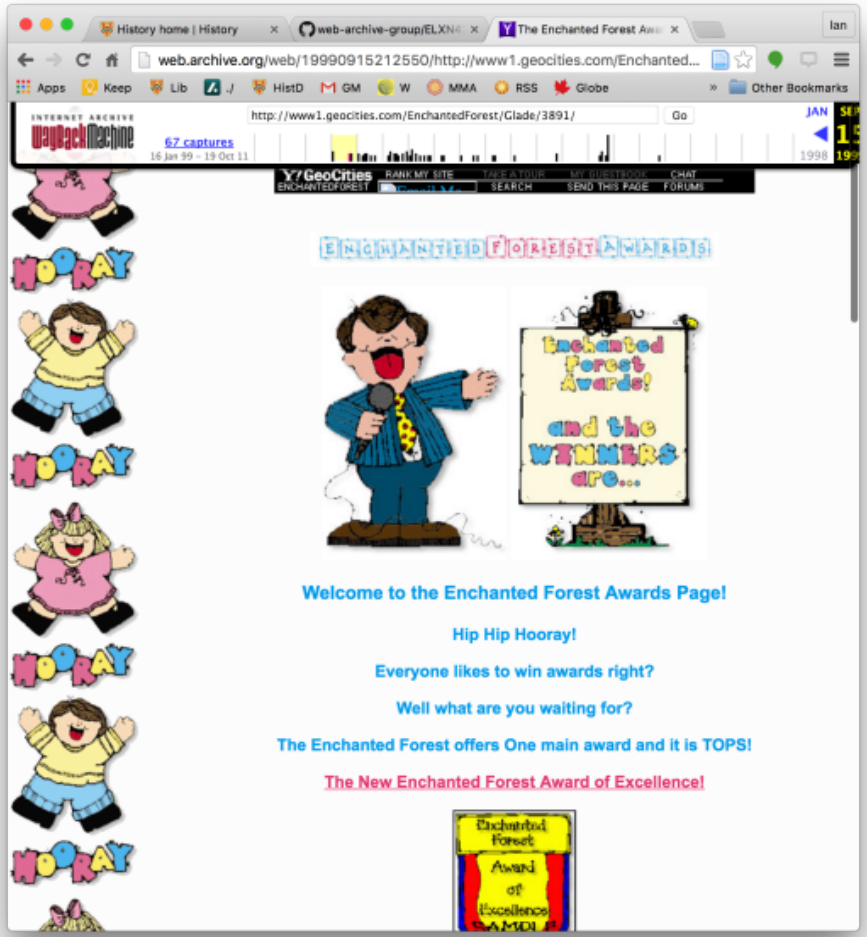
(this also speaks to the ethical concerns at play with this material, another reason to have us here!)

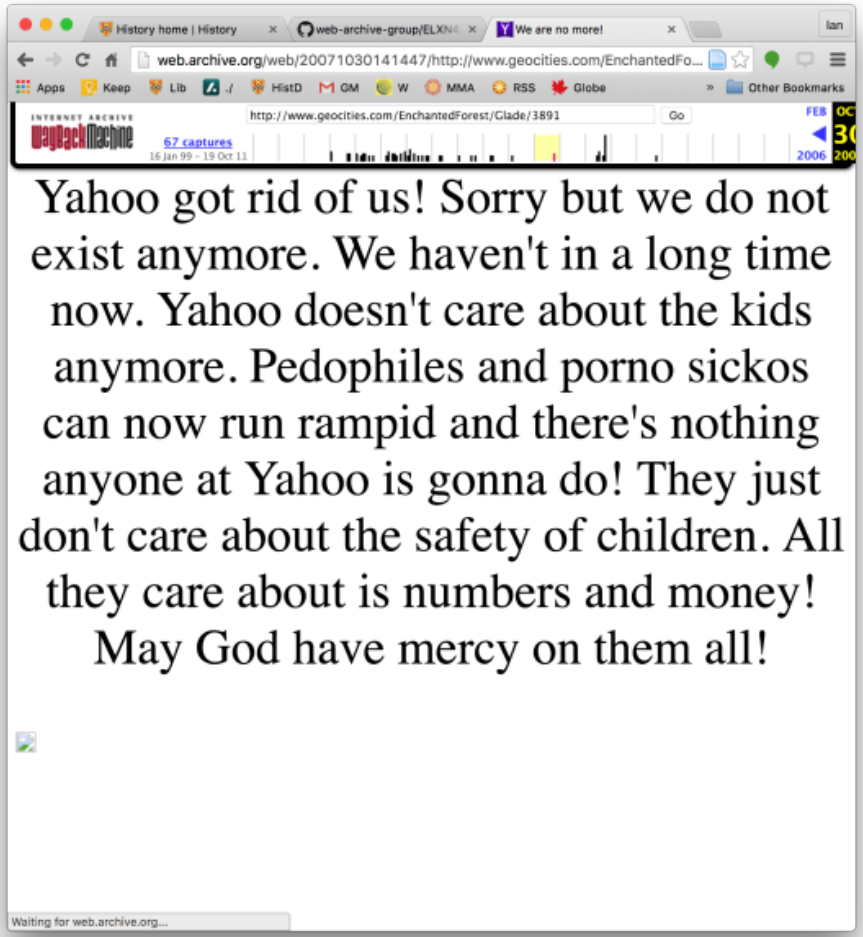
2005 Canadian Federal Election




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


http://www.geocities.com/EnchantedForest/Glade/9221
ww.geocities.com/EnchantedForest/Glade/3891
http://www.geocities.com/EnchantedForest/1004
http://www.geocities.com/EnchantedForest/Glade/9378
http://www.geocities.com/EnchantedForest/7631


web.archive.org/web/19990915212550/http://www1.geocities.com/Enchanted...
http://www1.geocities.com/EnchantedForest/Glade/3891/
67 captures
16 Jan 99 - 19 Oct 11
GeoCities
RANK MY SITE
SEARCH
SEND THIS PAGE
FORUMS
ENCHANTED FOREST AWARDS
HOORAY
HOORAY
HOORAY
HOORAY
enchanted forest awards!
and the WINNERS are...
Welcome to the Enchanted Forest Awards Page!
Hip Hip Hooray!
Everyone likes to win awards right?
Well what are you waiting for?
The Enchanted Forest offers One main award and it is TOPS!
The New Enchanted Forest Award of Excellence!
Enchanted Forest
Award
of
Excellence


web.archive.org/web/20071030141447/http://www.geocities.com/EnchantedFo...
http://www.geocities.com/EnchantedForest/Glade/3891
67 captures
16 Jan 99 - 19 Oct 11
Yahoo got rid of us! Sorry but we do not exist anymore. We haven't in a long time now. Yahoo doesn't care about the kids anymore. Pedophiles and porno sickos can now run rampid and there's nothing anyone at Yahoo is gonna do! They just don't care about the safety of children. All they care about is numbers and money! May God have mercy on them all!
Waiting for web.archive.org...

Suddenly web archives can talk to digital humanists

… but to really build interest in this, we need a community!

Case Two: Datathons

- March 2016, University of Toronto
- June 2016, Library of Congress
- 2017, Bay Area? Ottawa? [Insert your library name here]
- Hackathon/Datathon

Case Two: Datathons

- Organized by **Ian Milligan** (Waterloo), **Nathalie Casemajor** (UQO), **Jimmy Lin** (UW), **Matthew Weber** (Rutgers), **Nicholas Worby** (Toronto)
- SSHRC/NSF

The animating question

Web archives are great, but access and usage are a considerable problem.

KEEP CALM AND HACK AND YACK

Fort Book

Day One: Talking (or ‘Yakking’)

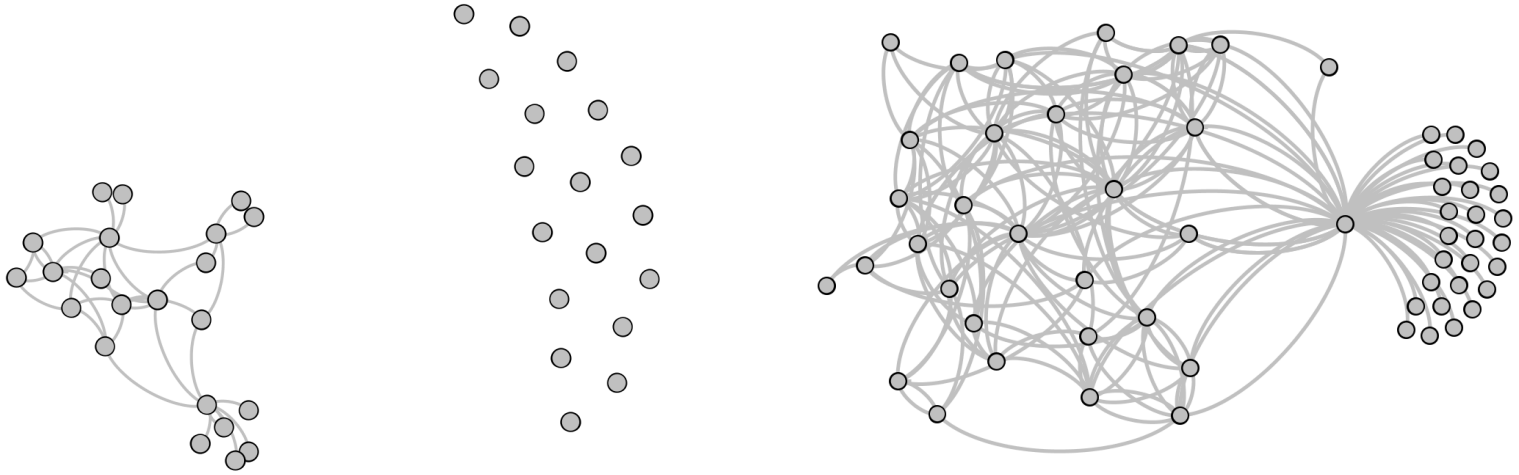
Day One: Team Forming

Video by Teis Moller Kristensen and Matthew Weber (Rutgers)

Day One: Socializing

Day Two - Three: Work

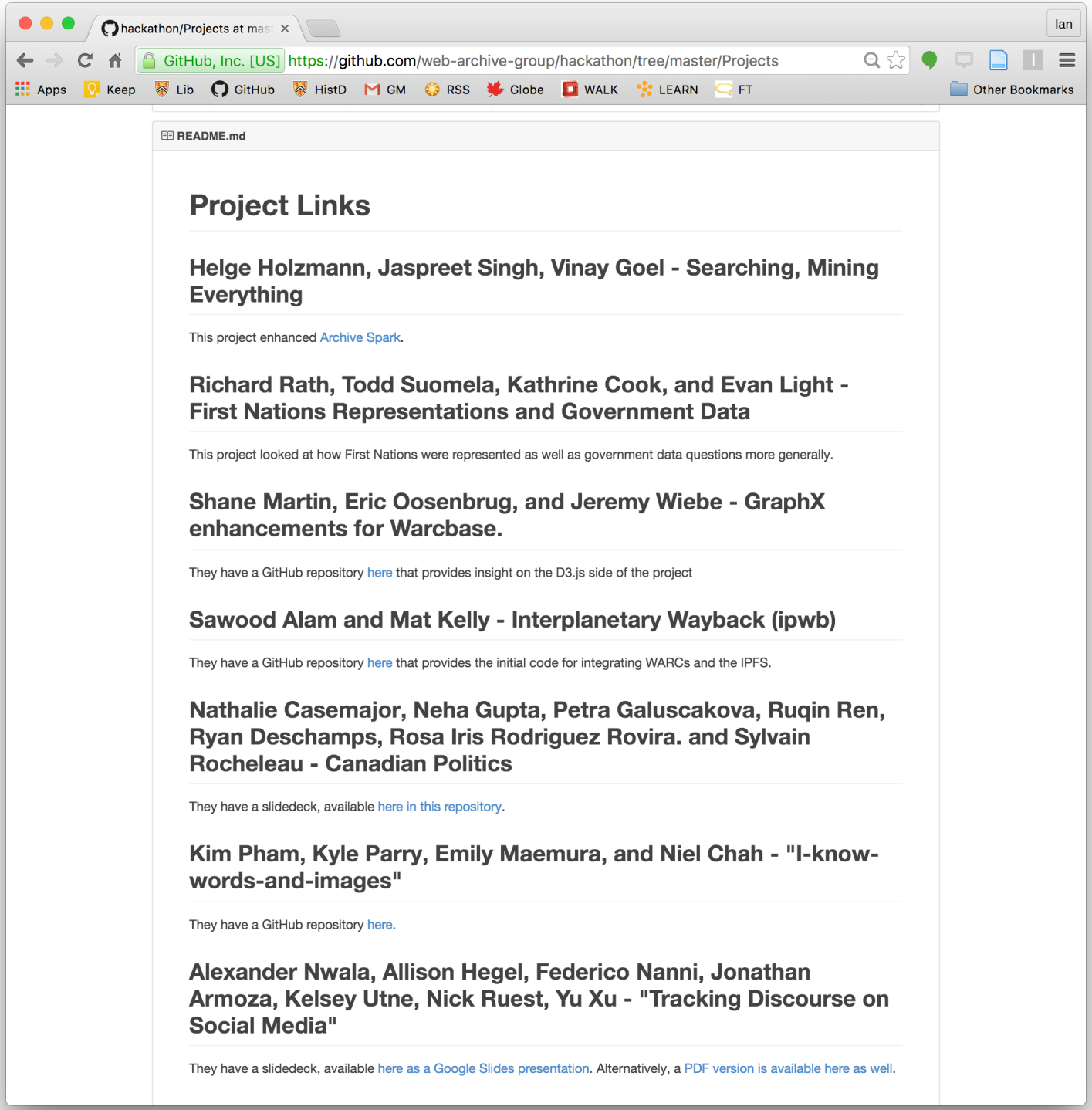
Forming Connections



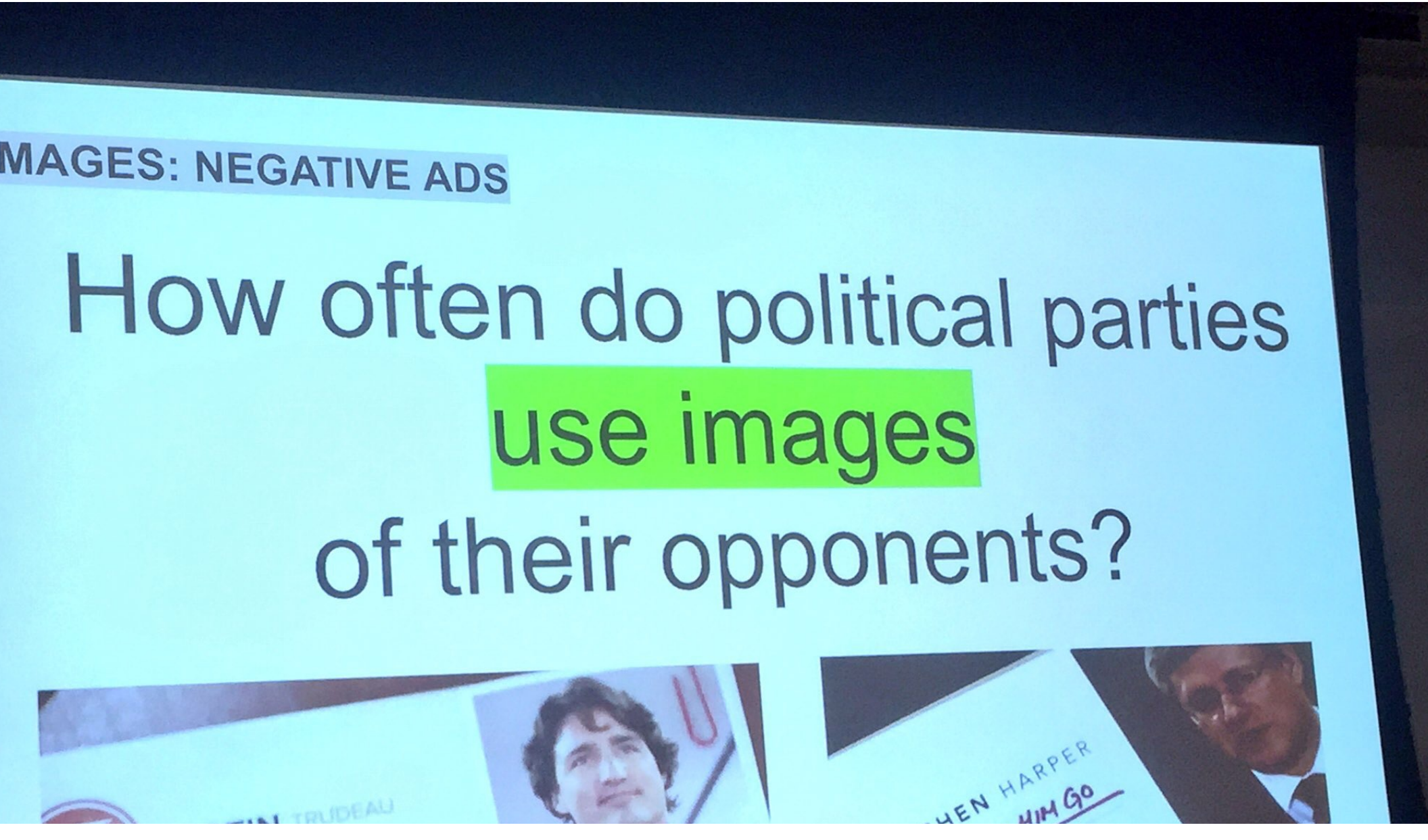
Exchanging ideas *before…* *and after.*

Matt Weber + Teis Moller Kristensen

The Projects

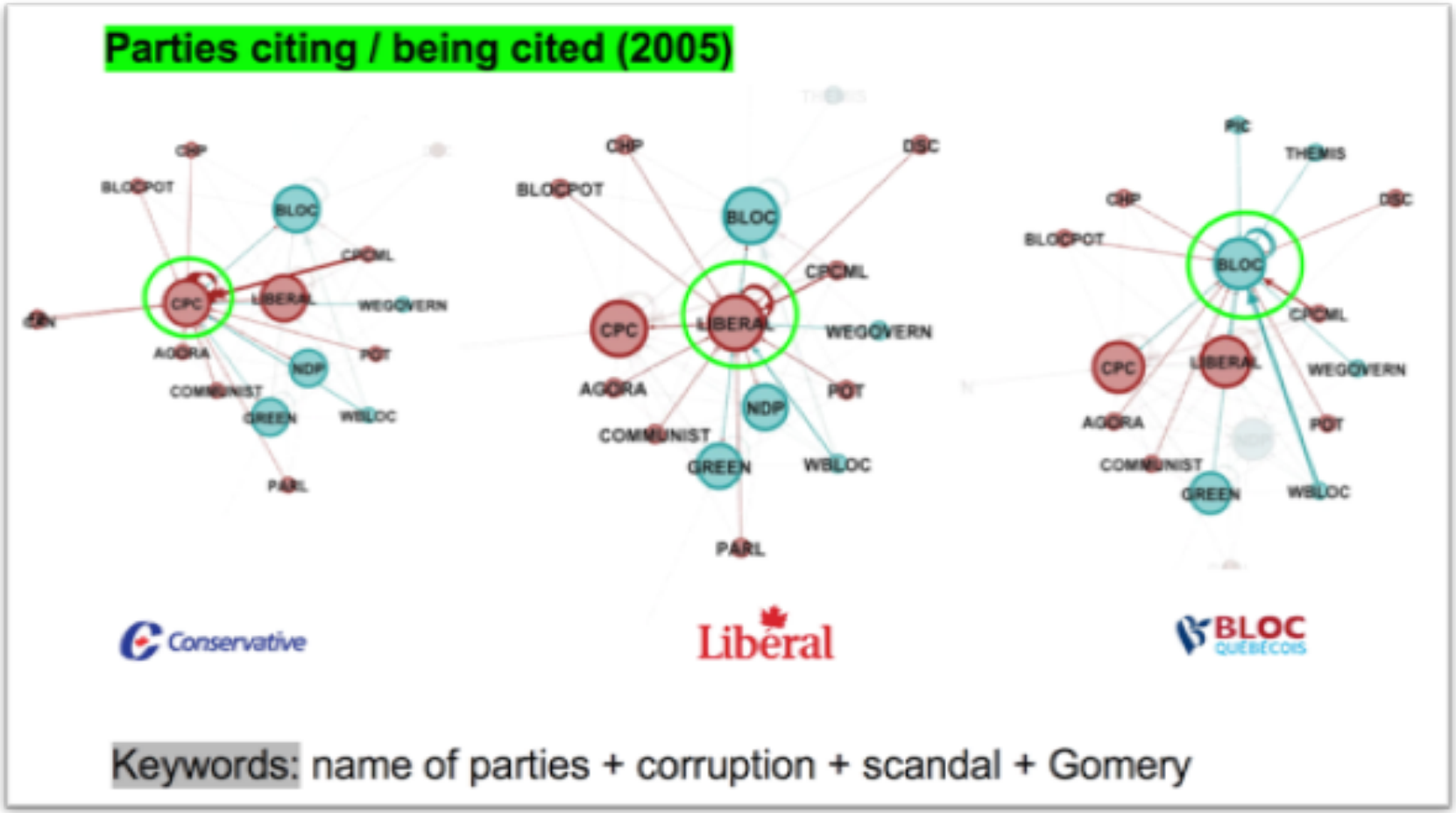


Day Three: Awards



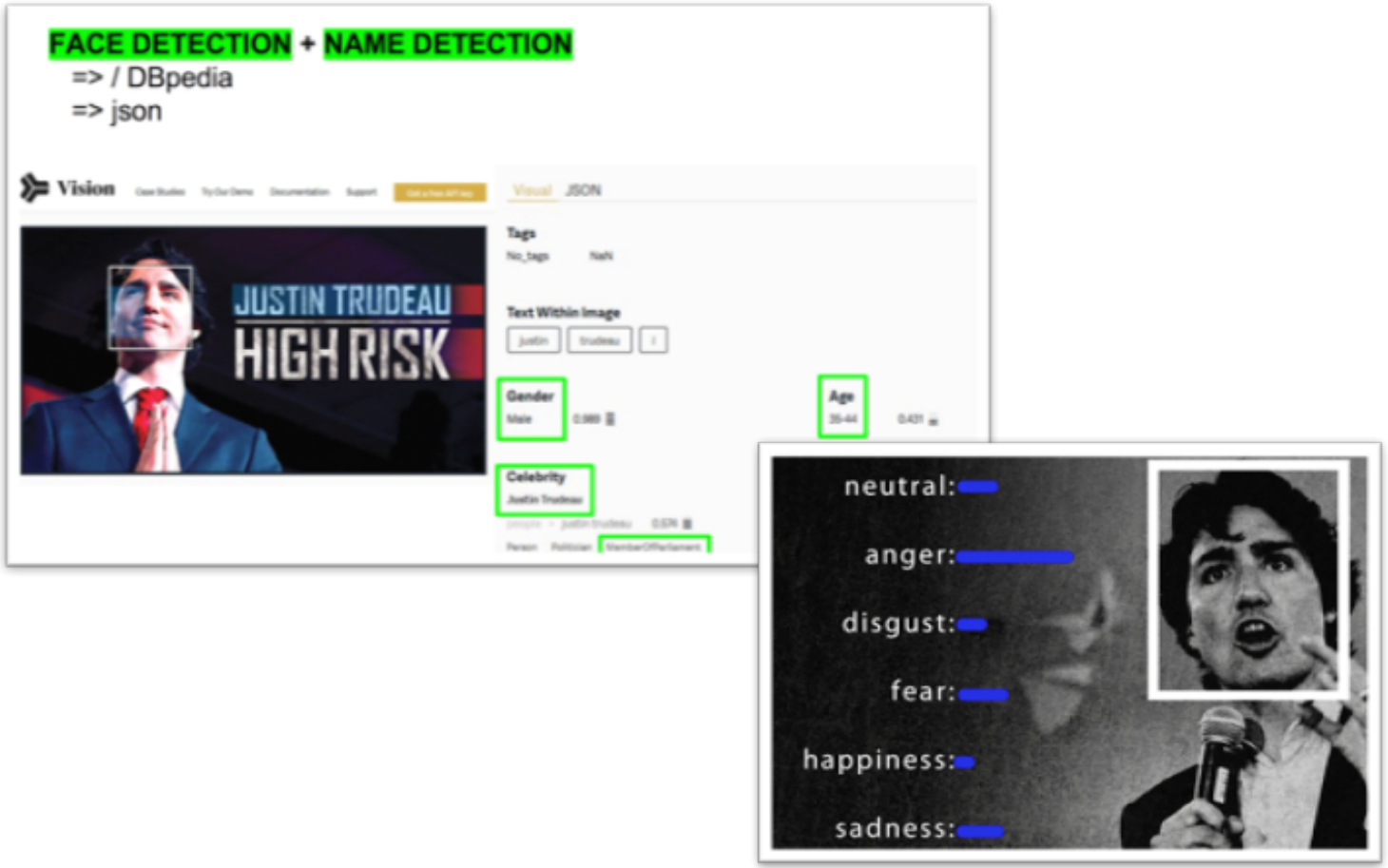
Sample Projects:

How do Canadian political parties represent their opponents?



Sample Projects:

How do Canadian political parties represent their opponents?



Sample Projects:
A diverse team

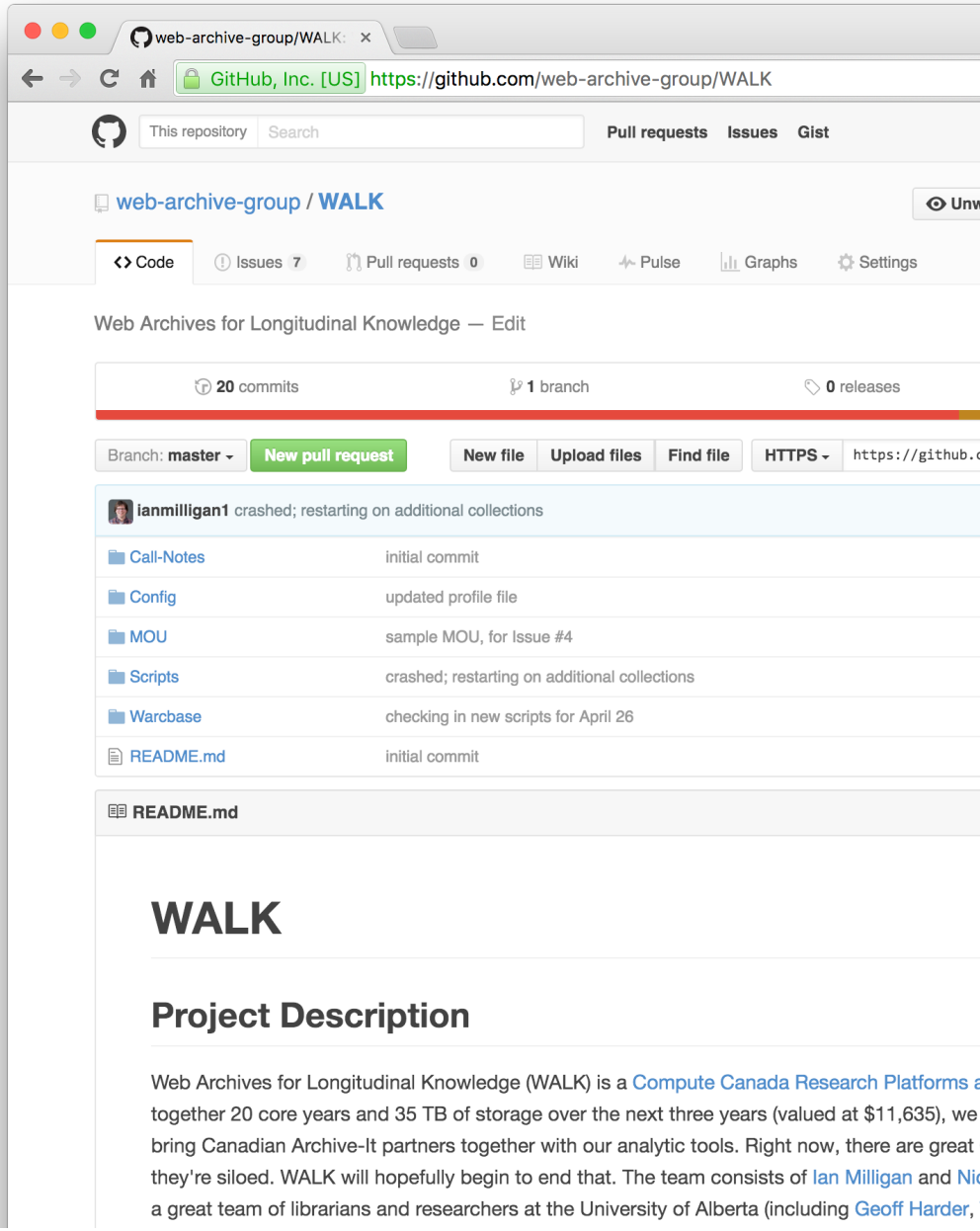
Nathalie Casemajor (Assistant Professor of Communication, Université du Québec en Outaouais)
Neha Gupta (Postdoctoral Fellow in Geography, Memorial University)
Petra Galuscakova (PhD Candidate in Linguistics, Charles University in Prague)
Ruqin Ren (PhD Candidate in Communication, University of Southern California)
Ryan Deschamps (PhD Candidate in Politics, University of Regina)
Rosa Iris Rodriguez Rovira (PhD Candidate in Communications, Université du Québec en Outaouais)
Sylvain Rocheleau (PhD Candidate in Cognitive Informatics, Université du Québec à Montréal)

… datathons can be ephemeral; what about infrastructure?

Case Three: Infrastructure

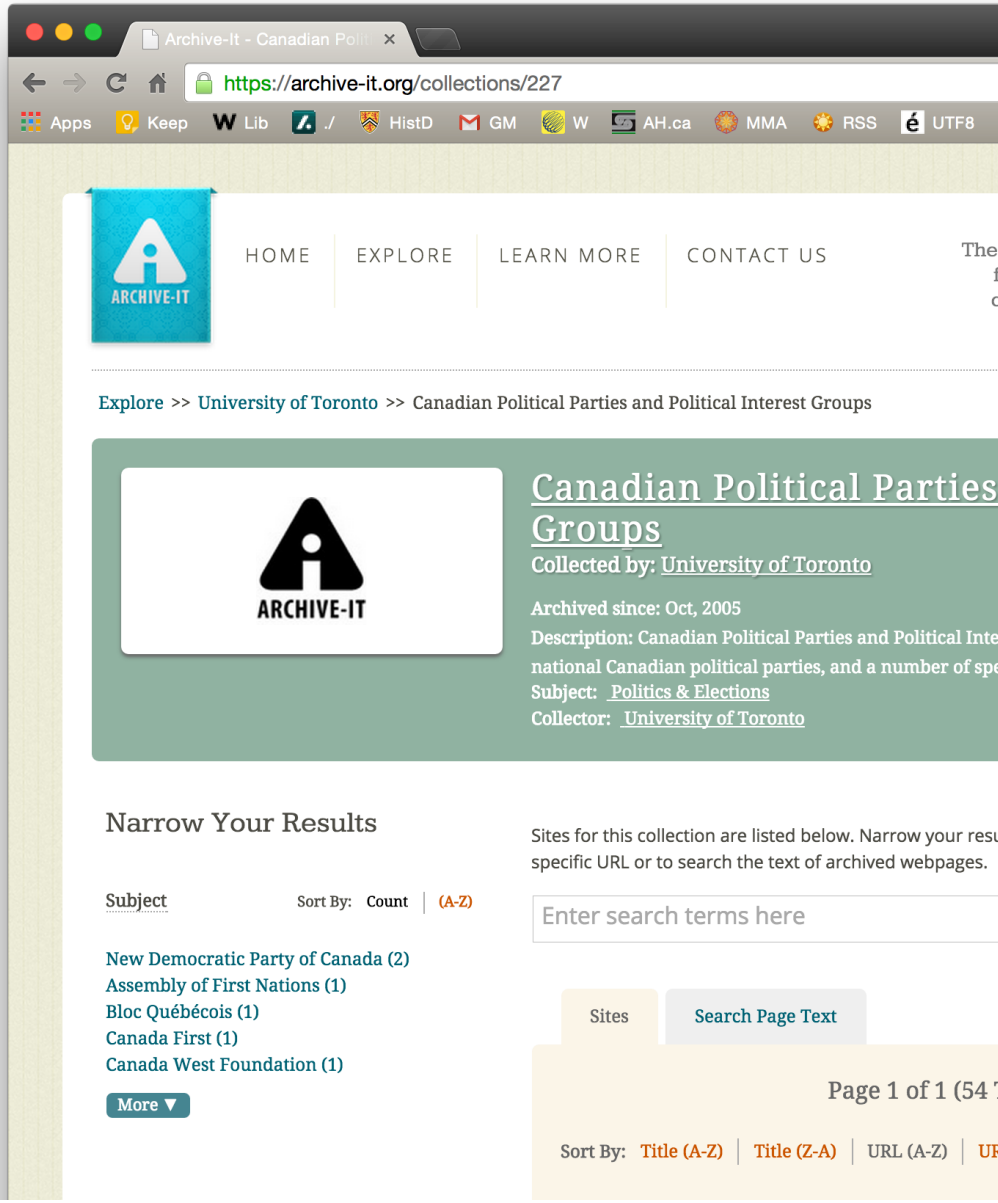
- **Web Archives for Longitudinal Knowledge (WALK)**

- **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).



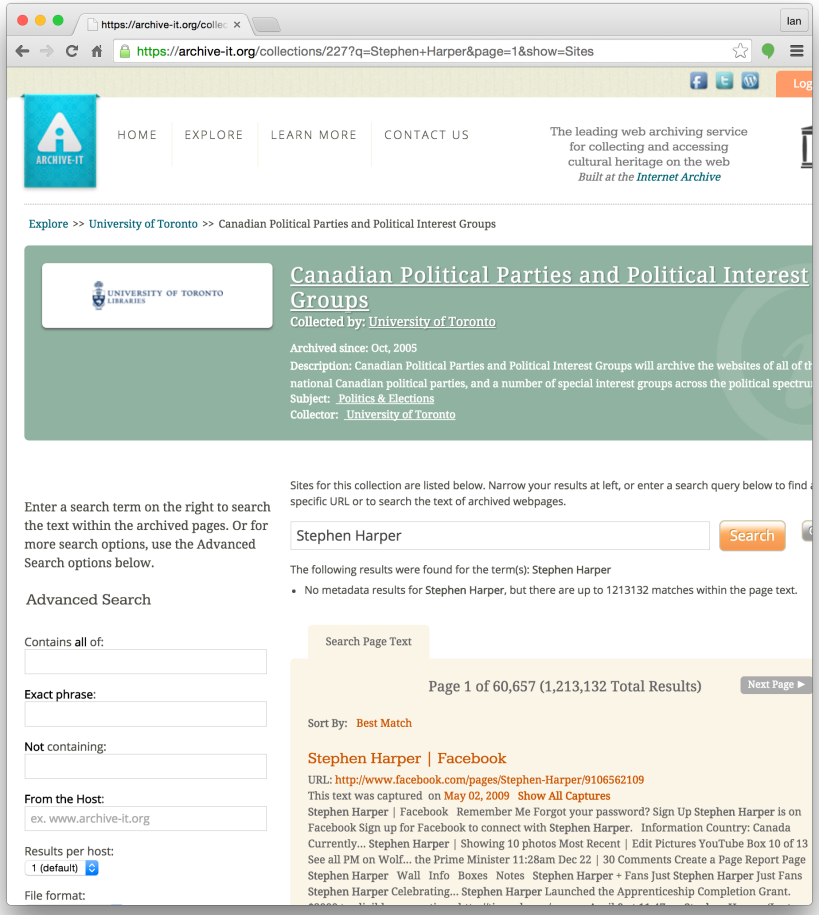
Archive-It Collections

- Currently ~ 25 Canadian partners, ~ 130 collections
- Back-end provider of web archiving services



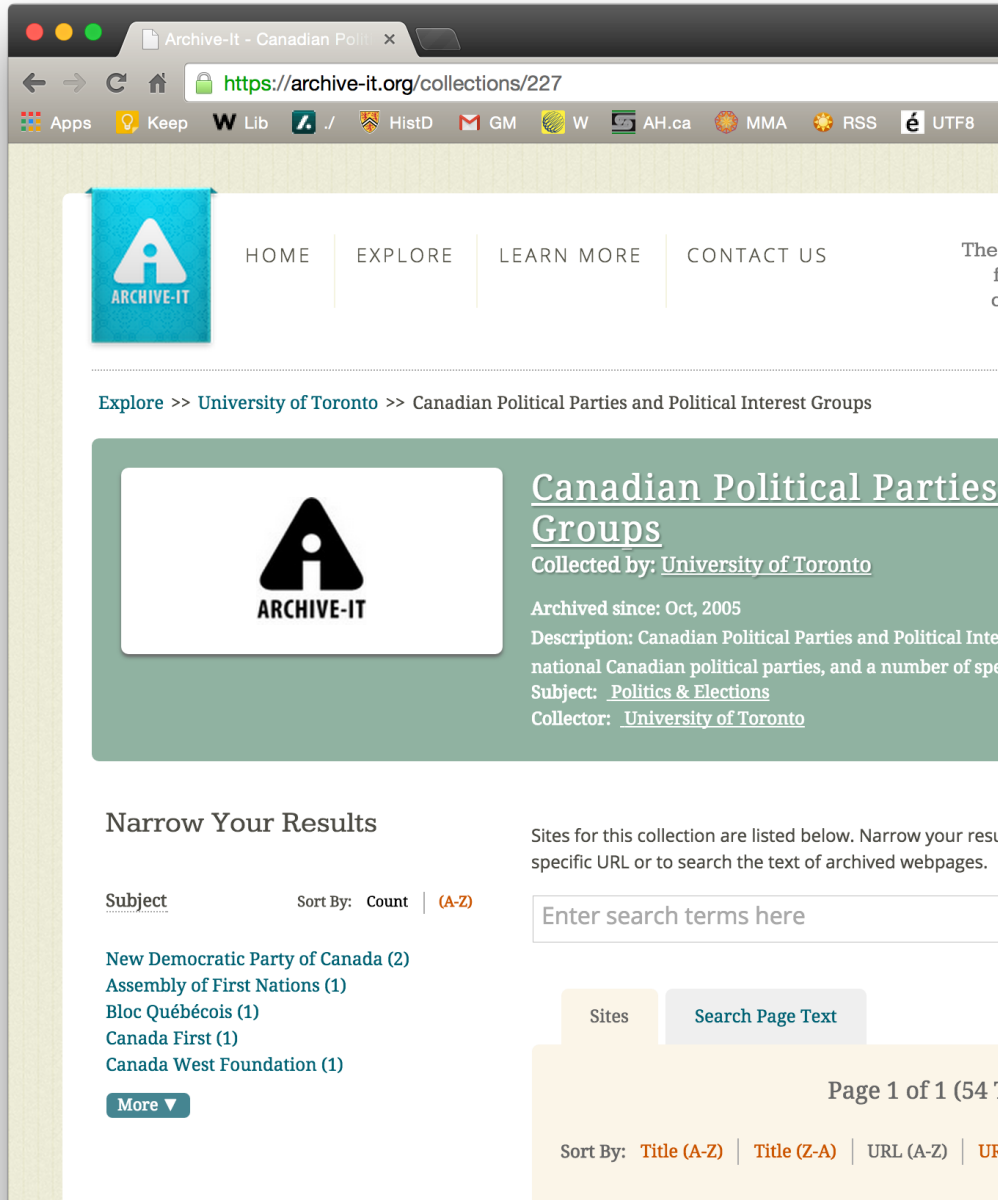
Current Interface

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



Building Portals

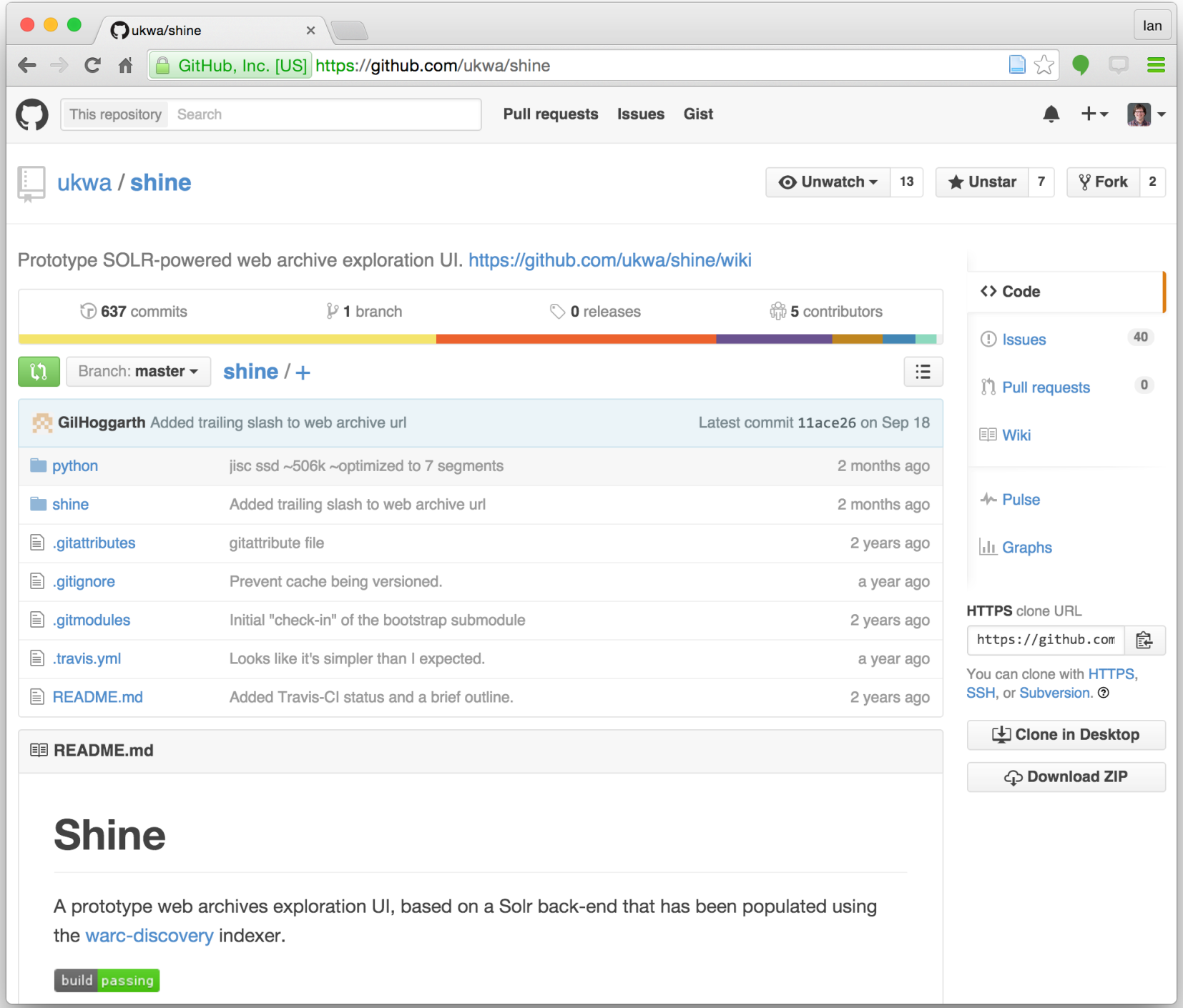
- Democratizing access so that people can use them.



Canadian Political Parties & Political Interest Group Collection

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

How could we build better access?


ukwa/shine
Ian
GitHub, Inc. [US] https://github.com/ukwa/shine
This repository Search
Pull requests Issues Gist
ukwa / shine
Unwatch 13 Unstar 7 Fork 2
Prototype SOLR-powered web archive exploration UI. https://github.com/ukwa/shine/wiki
637 commits 1 branch 0 releases 5 contributors
Branch: master shine / +
GilHoggarth Added trailing slash to web archive url
Latest commit 11ace26 on Sep 18
python jisc ssd ~506k ~optimized to 7 segments 2 months ago
shine Added trailing slash to web archive url 2 months ago
.gitattributes gitattribute file 2 years ago
.gitignore Prevent cache being versioned. a year ago
.gitmodules Initial "check-in" of the bootstrap submodule 2 years ago
.travis.yml Looks like it's simpler than I expected. a year ago
README.md Added Travis-CI status and a brief outline. 2 years ago
README.md
Shine
A prototype web archives exploration UI, based on a Solr back-end that has been populated using the warc-discovery indexer.
build passing
Code
Issues 40
Pull requests 0
Wiki
Pulse
Graphs
HTTPS clone URL
https://github.com
You can clone with HTTPS, SSH, or Subversion.
Clone in Desktop
Download ZIP

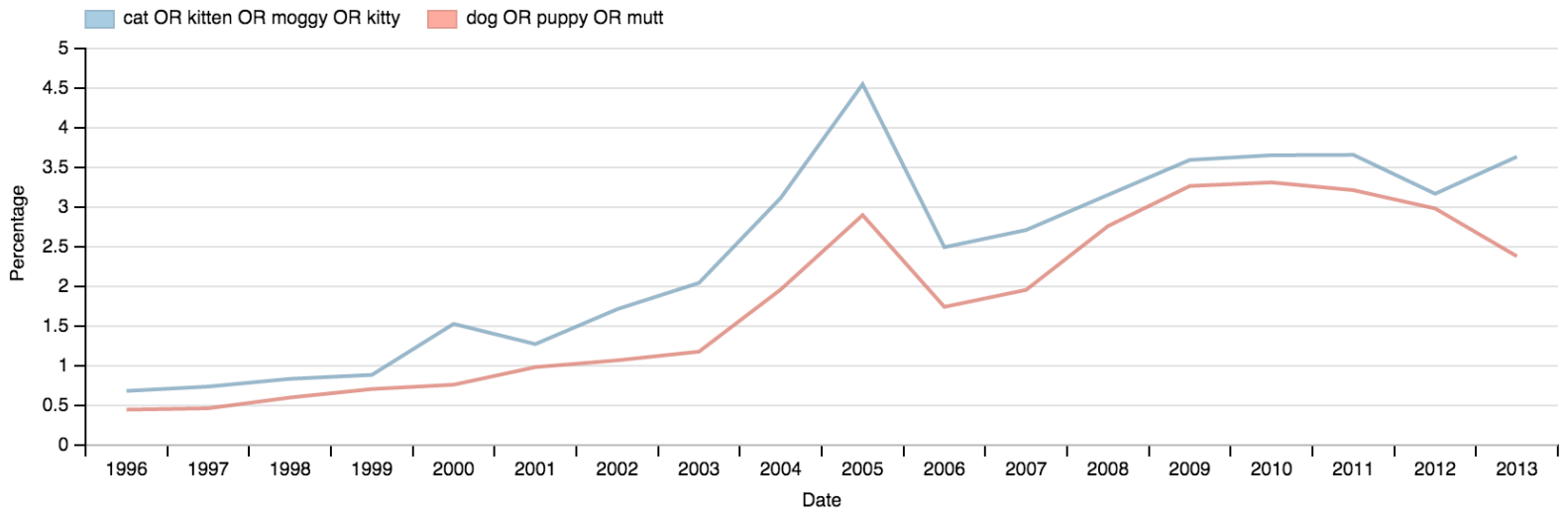
Trend results 1996-2013 fo

https://www.webarchive.org.uk/shine/graph?query=cat+OR+kitten+OR+moggy+OR+kitty%2C+dog+OR+pu...

UK Web Archive | Search | Trends | Login

Warning! This is a research prototype for a web archive search service, and may be taken down at any time.

cat OR kitten OR moggy OR kitty, dog OR | 1996 | 2013 | Update Graph | Reset All



Click on a point on the graph to show a random sample of the matching records from that year...

WebArchives.ca

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

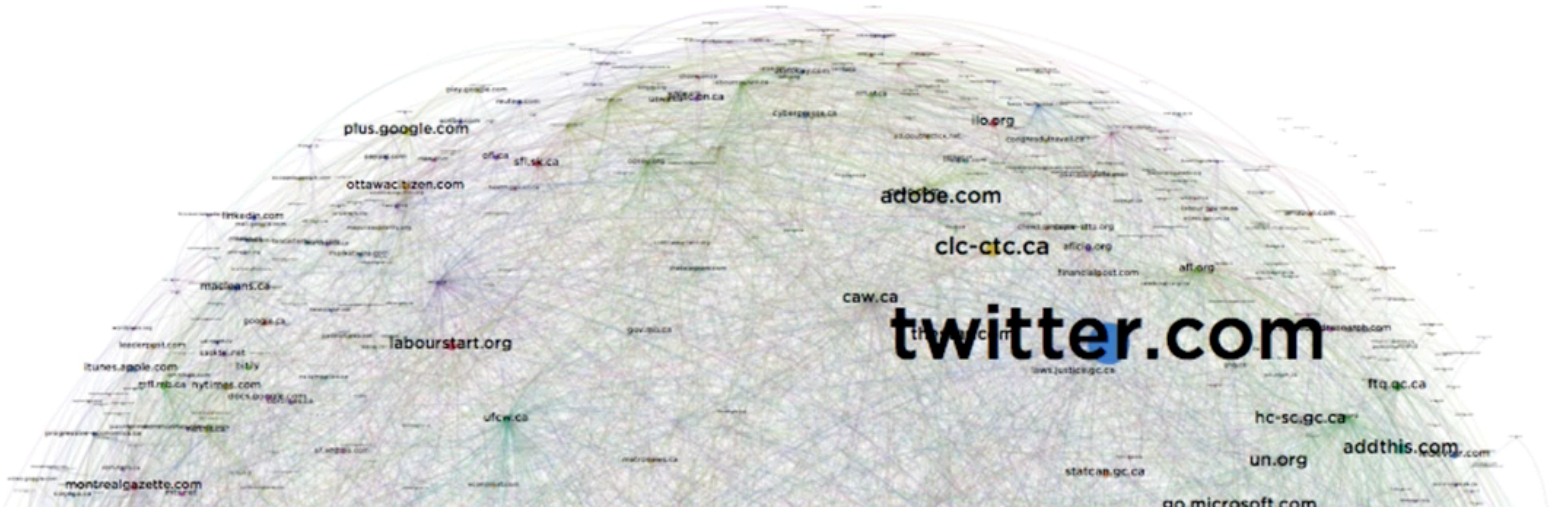
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

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]

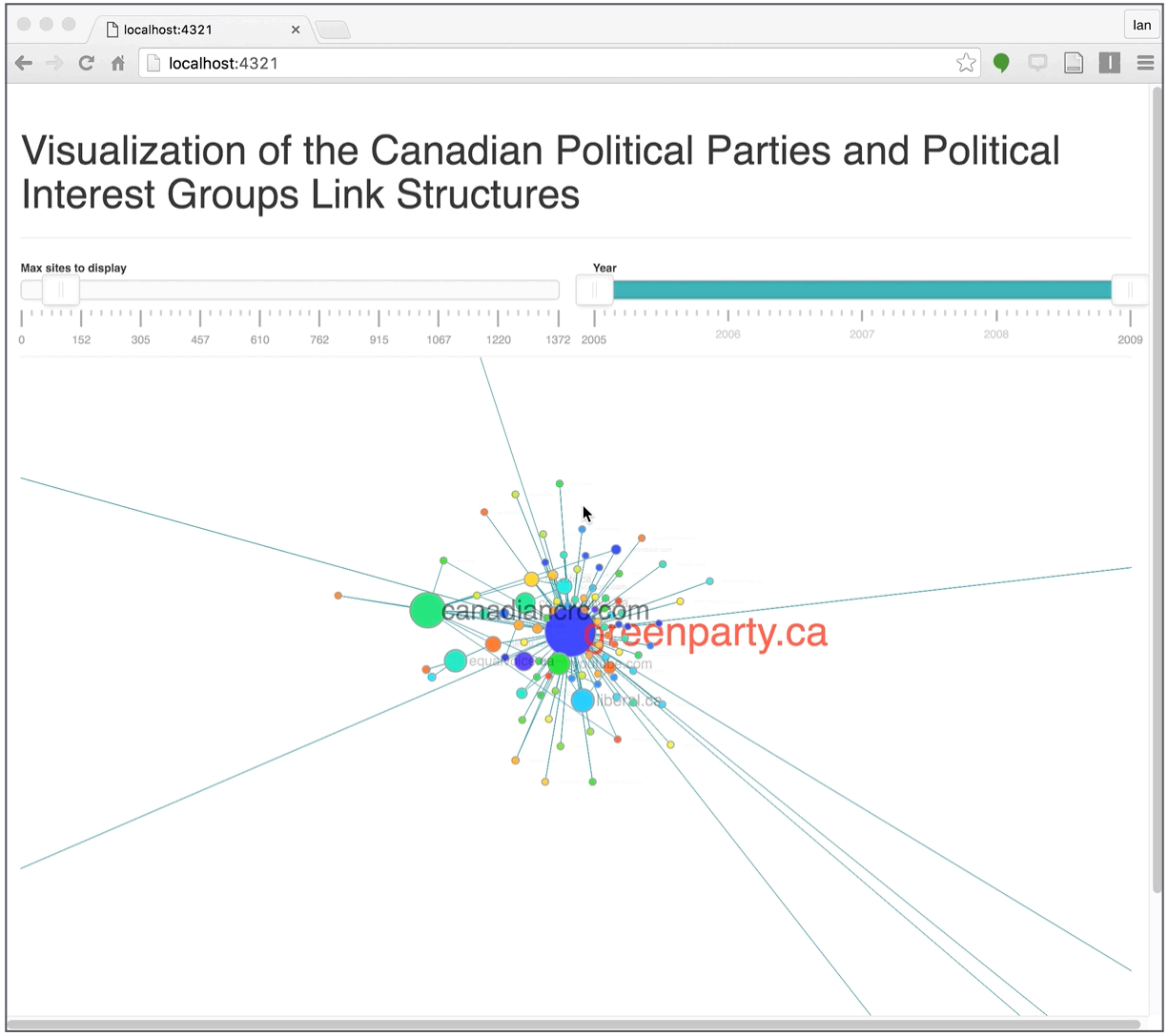
Below, here are all of the links for the entire time period, visualized below.



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


Political / activism
Industry / economic
French language
Grey / technical
???
alberta_oi
alberta_fl
energy_env
canadian_b
prarie_pro
elxn42-all
idle_no_mo
ottawa_sho
web_archiv
lfrancopho
humanities
health_sci
Factor 1
Factor 2
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Letting people play with metadata as well



What’s next?

Continuing to engage with born-digital cultural resources

Historians need to learn from all of you – our profession might be profoundly affected by the digital turn, but we're not ready.

Historians Need To Become Friends With:

- computers and algorithms;
- numbers;
- each other (teams);
- you/us (digital humanists);

Because, as I hope I have shown today..

it's worth it.

More voices, more people, the promise of social history achieved.

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Ontario

compute • calcul CANADA

UNIVERSITY OF WATERLOO

Thanks very much!

Questions?

Ian Milligan

Assistant Professor

@ianmilligan1

UNIVERSITY OF WATERLOO

FACULTY OF ARTS

Department of History