

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 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 

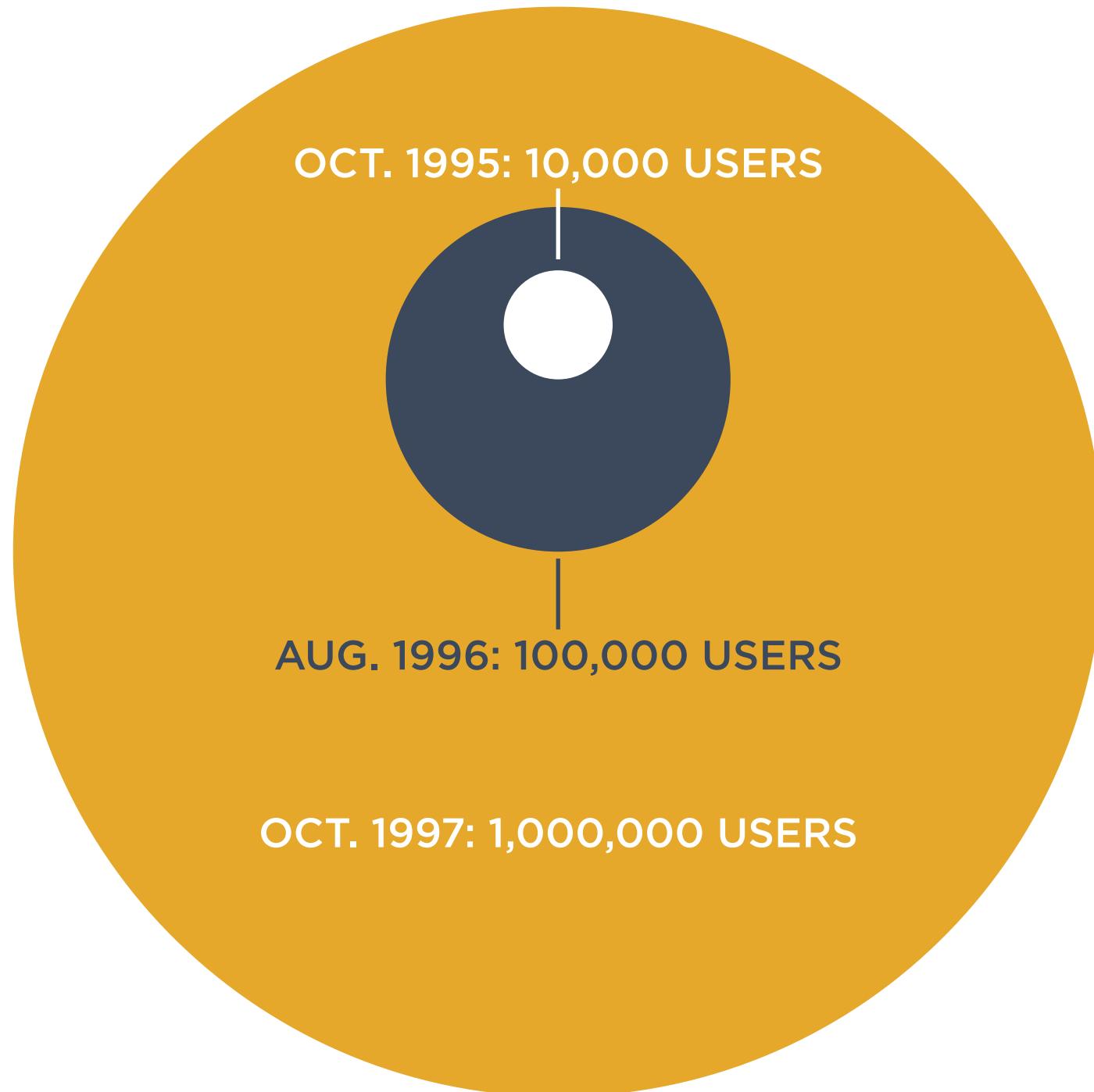
since June 5, 1996

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:



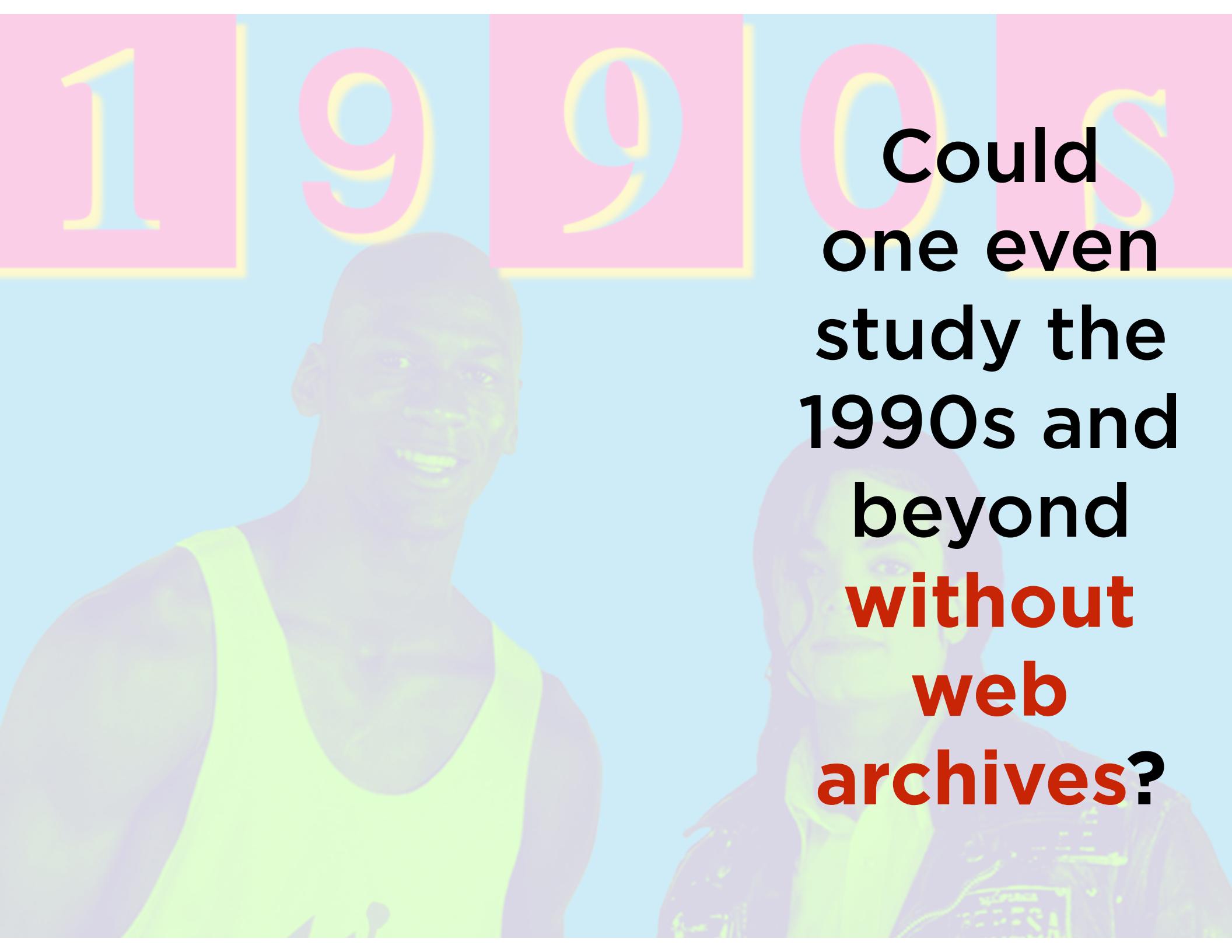
**This is a scale that
boggles the mind -
compare it to the Old
Bailey (197,745 trials
between 1674 and 1913)**

Scarcity



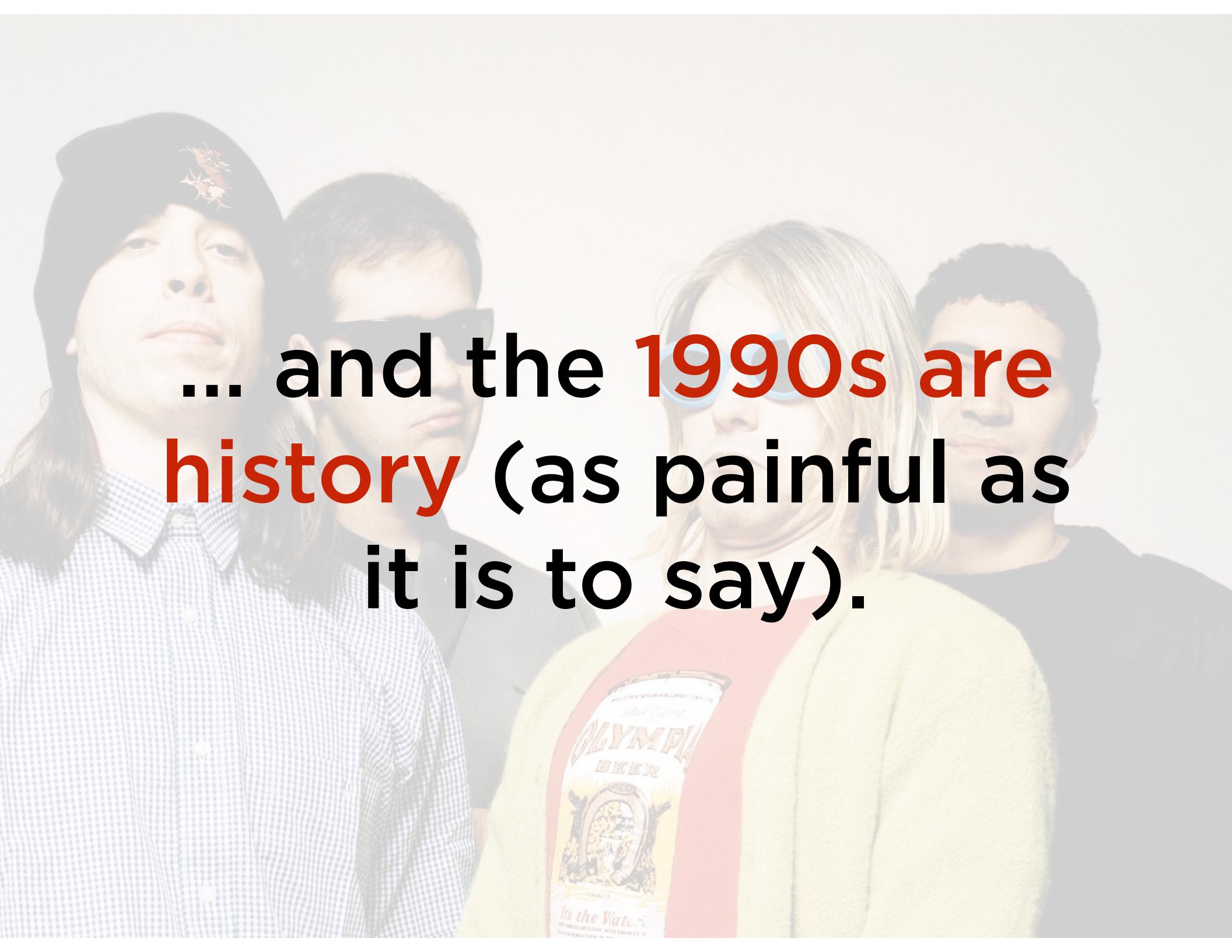
Scarcity
Abundance





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

1990s

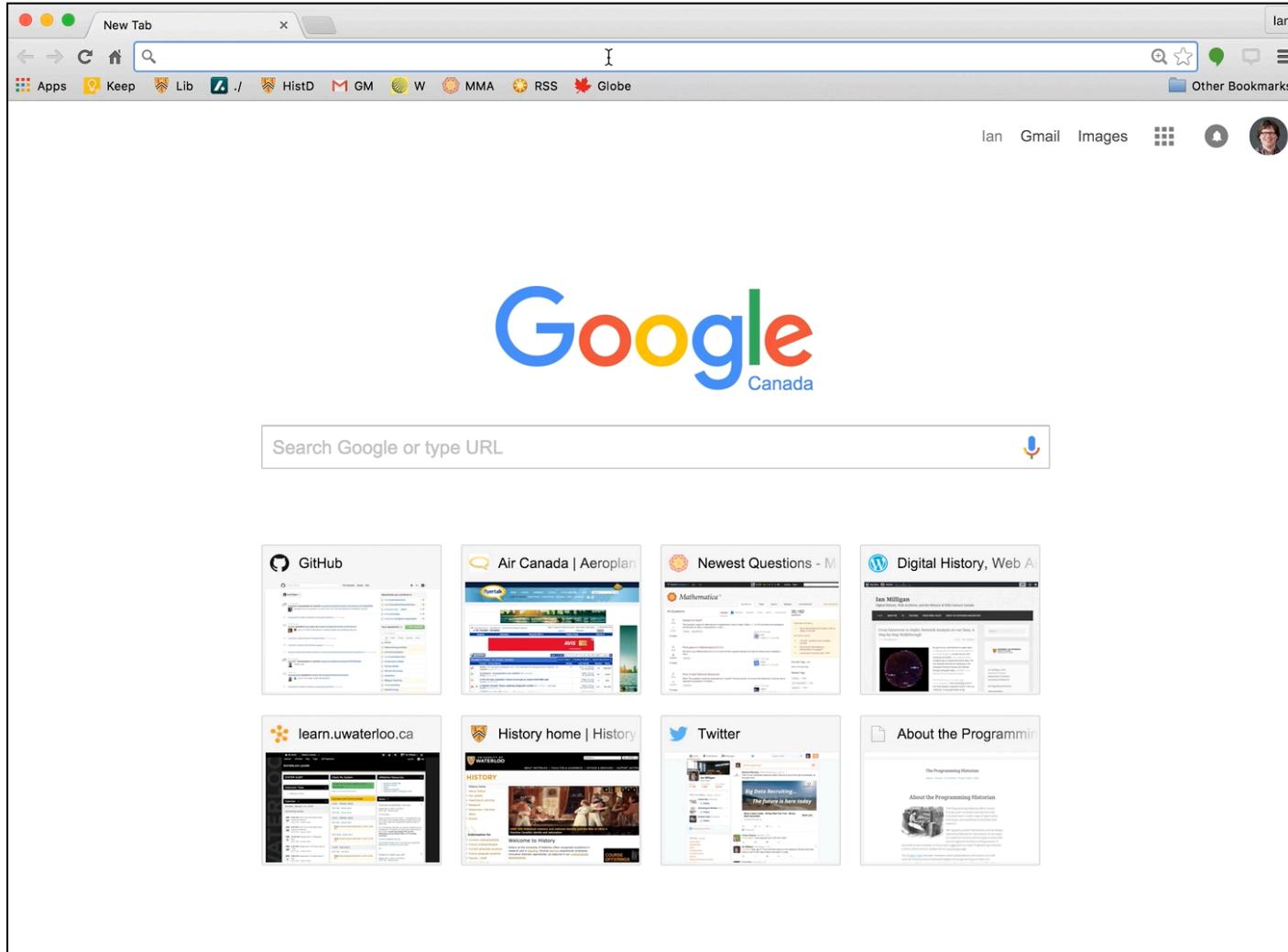


... and the 1990s are
history (as painful as
it is to say).

And I have fears

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

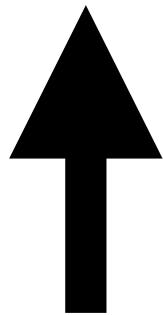
Nightmare Scenario



This won't be enough!



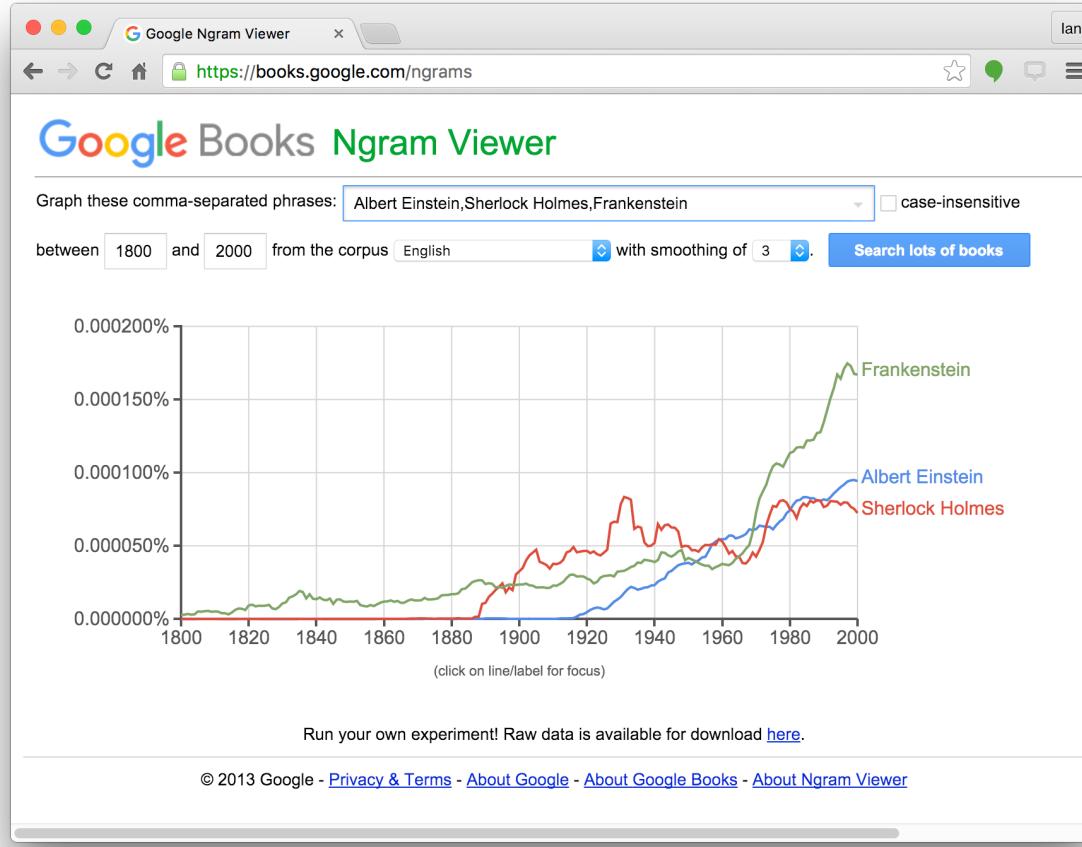
**... but what will our
search engines look
like?**



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

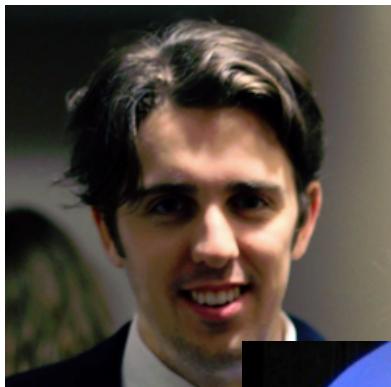
Teamwork

Web Archives for Historical Research

Historians



Computer Scientists



Librarians



Networks

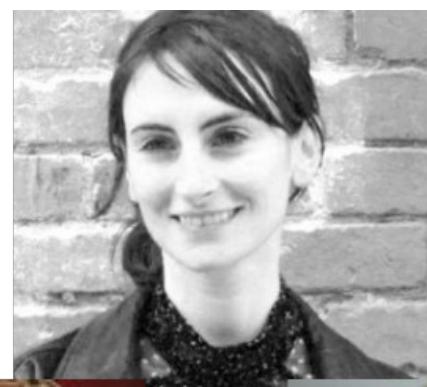


Team Datahon

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?

The screenshot shows a GitHub repository page for 'ianmilligan1/Historian-WARC-1'. The repository has 51 commits, 1 branch, 0 releases, and 1 contributor (the user). The latest commit was made 3 years ago. The commit history lists various files like WARC, .gitignore, All-Together.sh, README.md, etc., with descriptions such as 'DS_Store removal' and 'Ignore DS_Store'. The repository also contains a README.md file.

File	Description	Time Ago
WARC	DS_Store removal	3 years ago
.gitignore	Ignore DS_Store	3 years ago
All-Together.sh	Returned	3 years ago
README.md	noting developments in last two years!	a year ago
WARC-to-Analysis-Mathematica-ner.m	NER Functionality in These Files, see README	3 years ago
WARC-to-Analysis-Mathematica.m	Streamlined Program	3 years ago
WARC-to-Analysis-NER.sh	NER Functionality in These Files, see README	3 years ago
WARC-to-Analysis.sh	Updated for OS X 10.9	3 years ago
WARC-to-Stanford.sh	h/t to Jeremy Wiebe (@jeremyw), thanks for suggestions!	3 years ago
historian-warc-toolkit.sh	Bug fix	3 years ago
locpatr	NER Functionality in These Files, see README	3 years ago
orgpatr	NER Functionality in These Files, see README	3 years ago
personpatr	NER Functionality in These Files, see README	3 years ago
trial.cfg	Older Versions of WARC Tools Included, Please note their license	3 years ago

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

#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

Constant Communication

A screenshot of a GitHub repository page titled "ianmilligan1 / WebArchiving-Articles". The repository is private and has 5 issues, 0 pull requests, and 1 star. It was last updated on March 27. The commit history shows frequent activity from a user named "lintool" over the past month, with many commits related to "WebArchiving-Articles / JOCCH2016". The commits include updates to LaTeX templates, Warcbase schema, and visualization files, along with several "final verison submitted for review" messages.

File / Commit Message	Description	Time Ago
ACM-Reference-Format-Journals.bst	Templates in use.	2 months ago
acmcopyright.sty	Templates in use.	2 months ago
acmlarge.cls	Templates in use.	2 months ago
acmlarge.zip	Latex templates.	2 months ago
figures.pptx	Warcbase schema.	a month ago
hbase-schema.pdf	figures	a month ago
notebook-screenshot.png	more work on section 5, adding in notebook and running times, getting...	a month ago
person-vis.pdf	Added LDA and search screenshots.	a month ago
search-screenshot1.png	Added LDA and search screenshots.	a month ago
search-screenshot2.png	Added LDA and search screenshots.	a month ago
vis-crawl.pdf	renamed visualizations.	a month ago
vis-graph-2.png	changed GeoCities example to political one	a month ago
vis-graph.png	renamed visualizations.	a month ago
vis-ner-month.png	additional geocities/her/link graph info	a month ago
vis-ner.pdf	renamed visualizations.	a month ago
warcbase-screenshot.png	figures	a month ago
warcbase.bib	final verison submitted for review.	a month ago
warcbase.pdf	final verison submitted for review.	a month ago

Constant Communication

The screenshot shows a GitHub commit history for a repository named 'ianmilligan1/WebArchiving-Articles'. The commits are listed by date, starting from April 6, 2016, and ending with a commit on April 18, 2016. The commits are grouped by date, and each commit includes the author's profile picture, the commit message, the date it was committed, and a copy icon, a commit hash, and a diff icon.

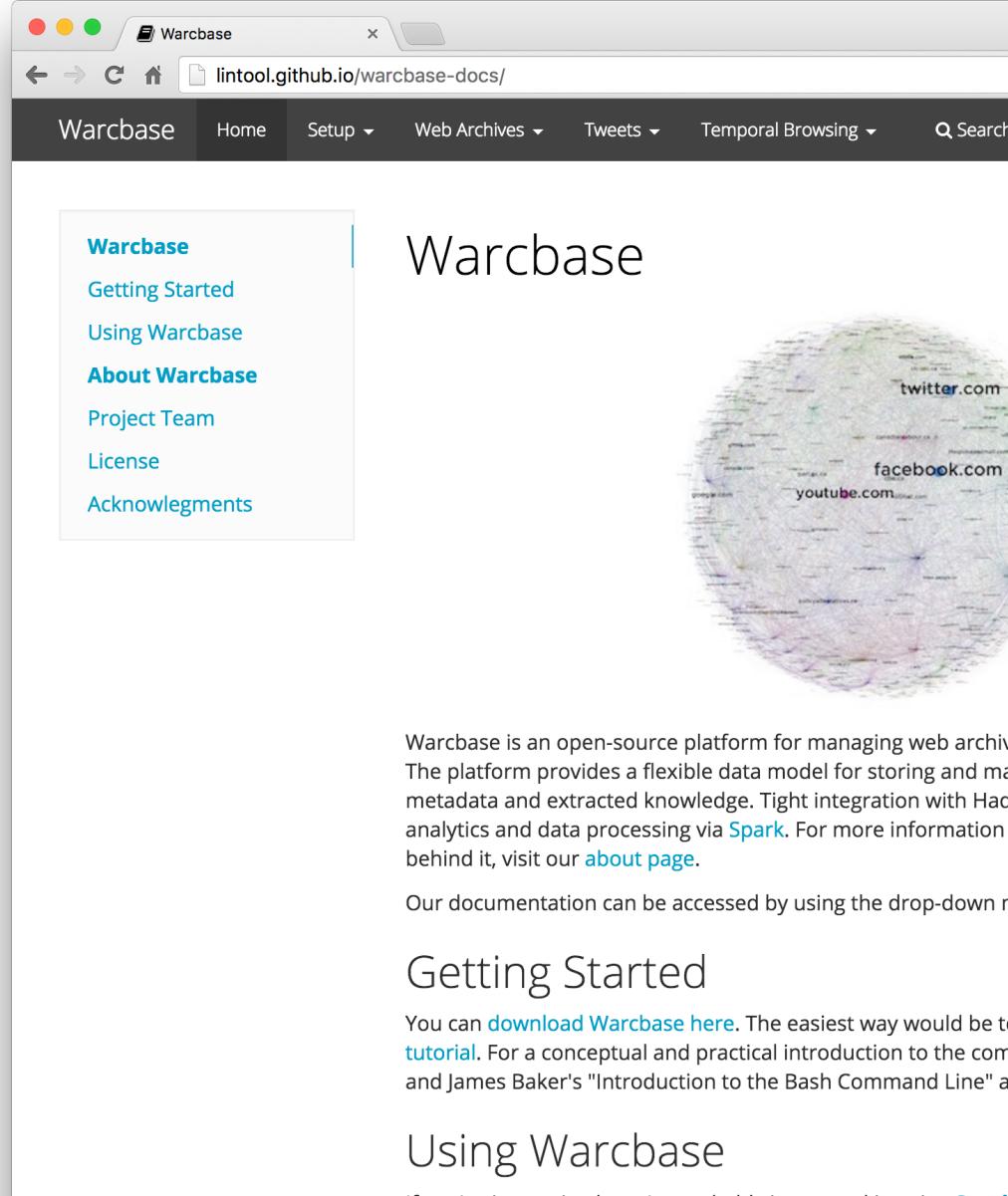
- Commits on April 18, 2016:
 - ACM accepted camera ready versions. (lintool, 10 days ago)
- Commits on April 15, 2016:
 - metadata (lintool, 14 days ago)
- Commits on April 14, 2016:
 - final camera ready. (lintool, 14 days ago)
 - Final camera ready. (lintool, 14 days ago)
 - tweaked styles (lintool, 14 days ago)
- Commits on April 11, 2016:
 - updating image tokens (ianmilligan1, 17 days ago)
 - tiny bit of interim progress (ianmilligan1, 17 days ago)
 - Fig 1 (ianmilligan1, 17 days ago)
- Commits on April 8, 2016:
 - another pass. (lintool, 20 days ago)
 - Release candidate. (lintool, 20 days ago)
- Commits on April 6, 2016:
 - checking in first draft (ianmilligan1, 22 days ago)

Three Projects

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

Case One: Warcbase

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



The screenshot shows a web browser window displaying the Warcbase documentation at lintool.github.io/warcbase-docs/. The page has a dark-themed header with the Warcbase logo and a navigation bar with links to Home, Setup, Web Archives, Tweets, Temporal Browsing, and Search. A sidebar on the left contains links to 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 representing websites like twitter.com, facebook.com, and youtube.com. The text in the main content area describes Warcbase as an open-source platform for managing web archives, using Hadoop for analytics, and integrating with Spark for data processing. It also mentions the availability of documentation and a tutorial.

Warcbase

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 platform and its features, visit our [about page](#).

Our documentation can be accessed by using the drop-down menu in the top navigation bar.

Getting Started

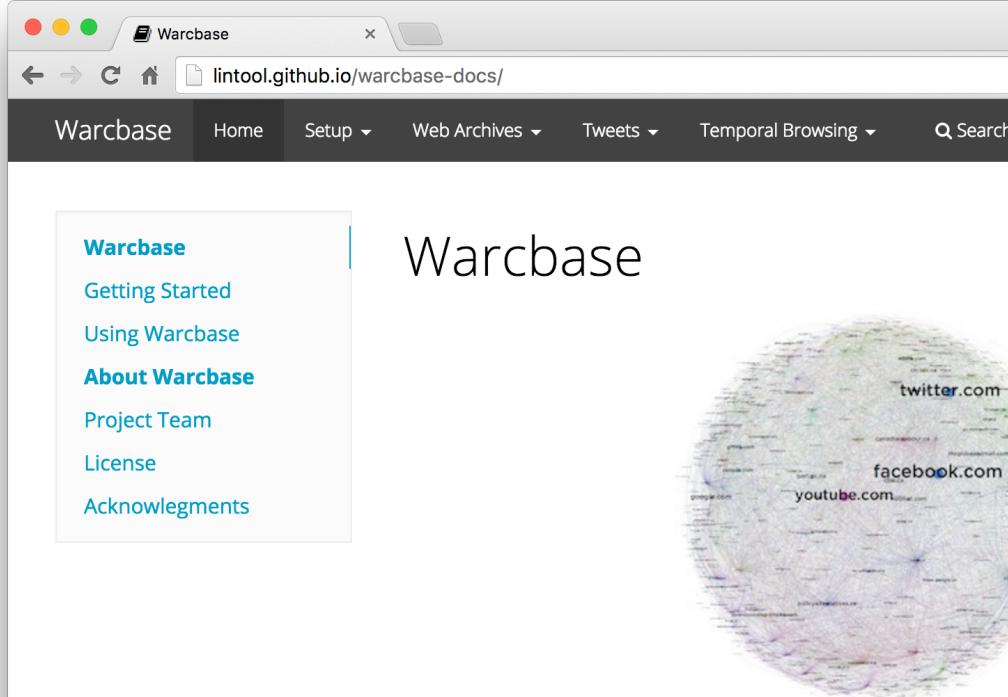
You can [download Warcbase here](#). The easiest way would be to follow the [tutorial](#). For a conceptual and practical introduction to the command-line interface, see [Hadoop](#) and James Baker's "Introduction to the Bash Command Line" at [GitHub](#).

Using Warcbase

If you've just arrived, you're probably interested in using [Spark](#) or [Apache Hadoop](#) to process your data. You can also explore the [API](#) or [CLI](#) to interact with Warcbase directly.

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.

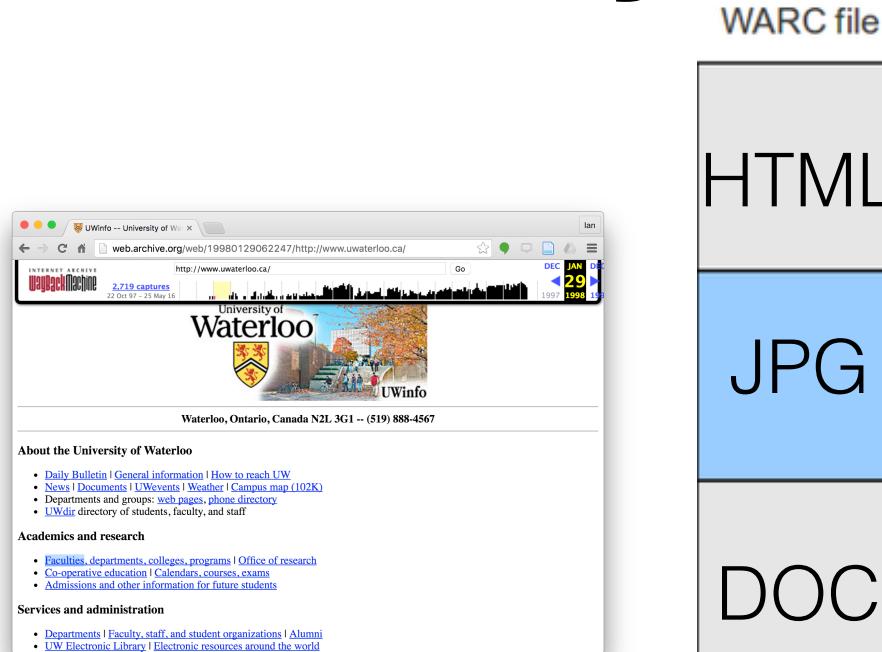
Getting Started

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.

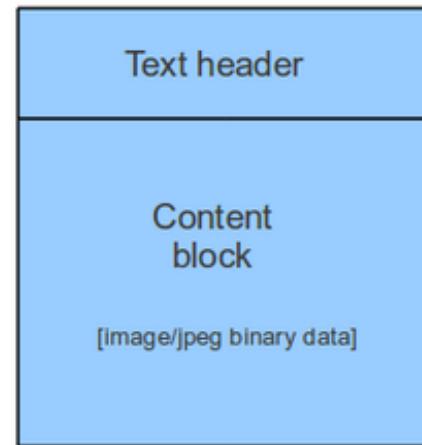
Why Warcbase?



WARC file



WARC record



```
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-f54c6ec90bb0>
Content-Type: image/jpeg
WARC-Payload-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
WARC-Block-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
Content-Length: 1662
```

Warcbase

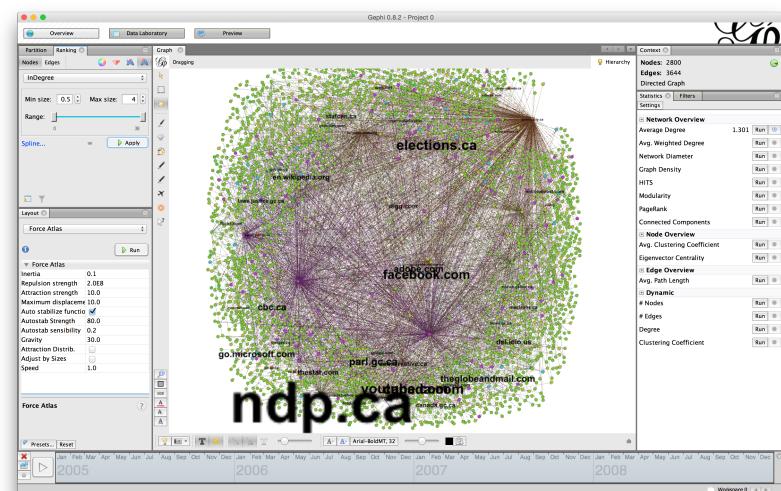
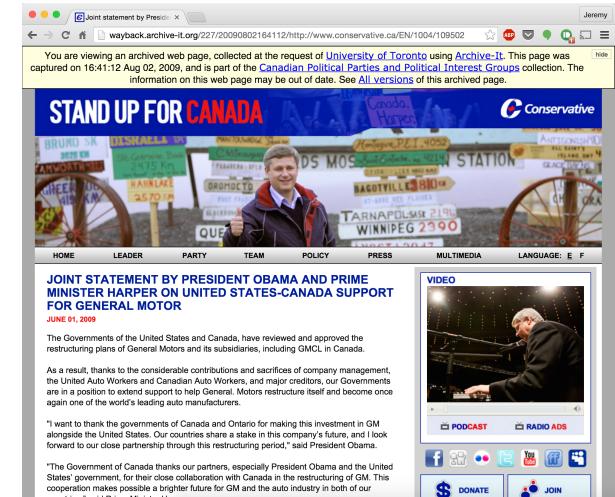
An open-source platform for managing web archives

<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

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.

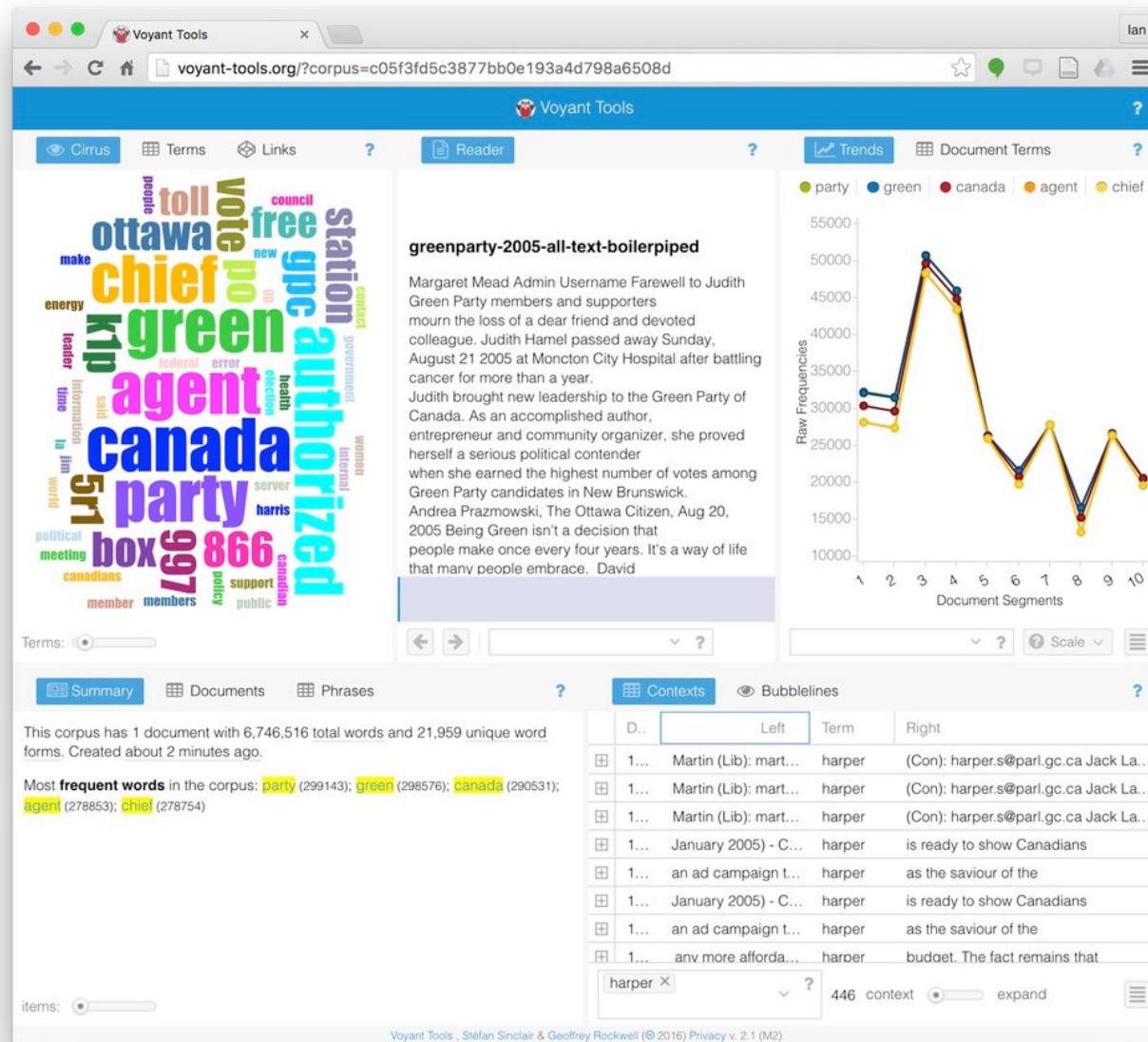
If you want to run it in your Spark Notebook, the following script will show in-notebook plain text:

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

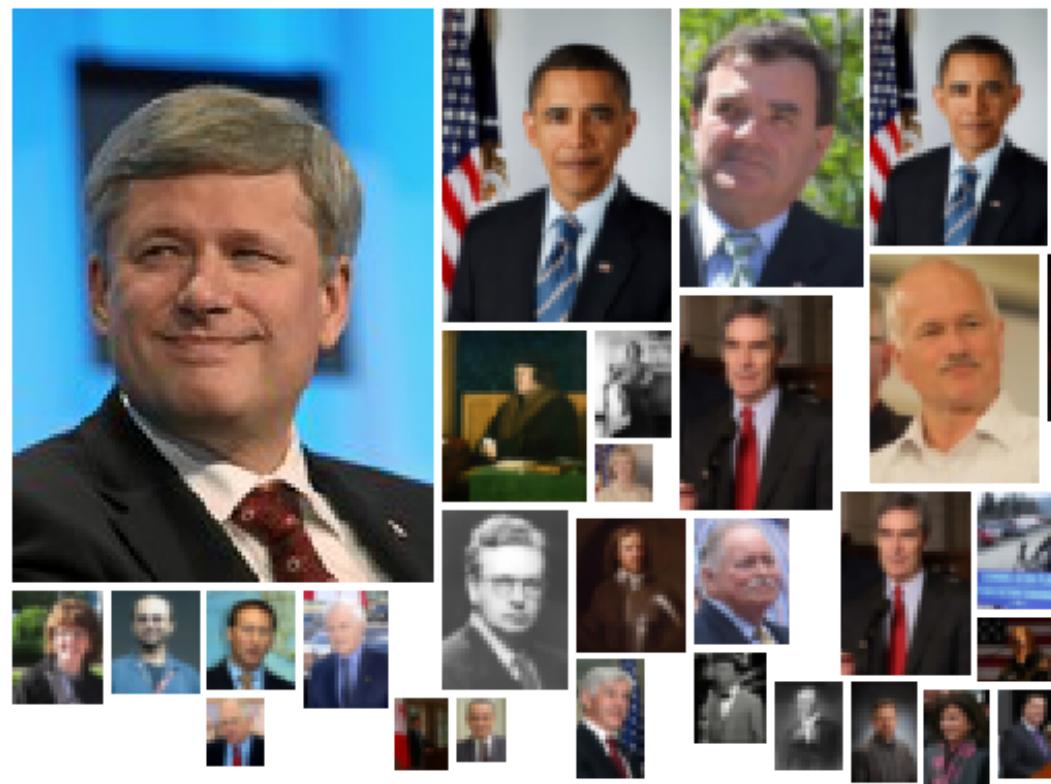
Extract all Text

```
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



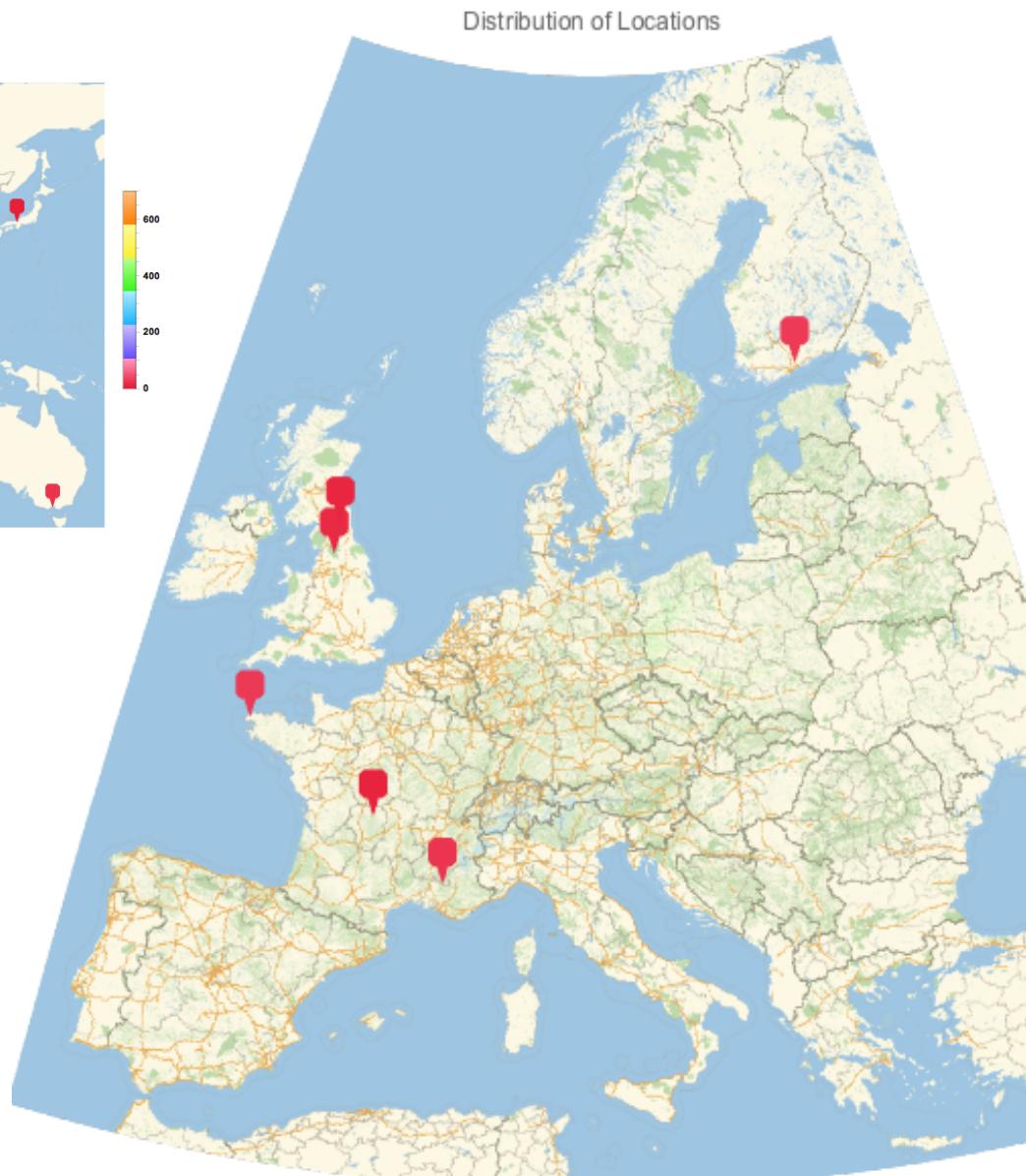
Extract Entities



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,
```



**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]

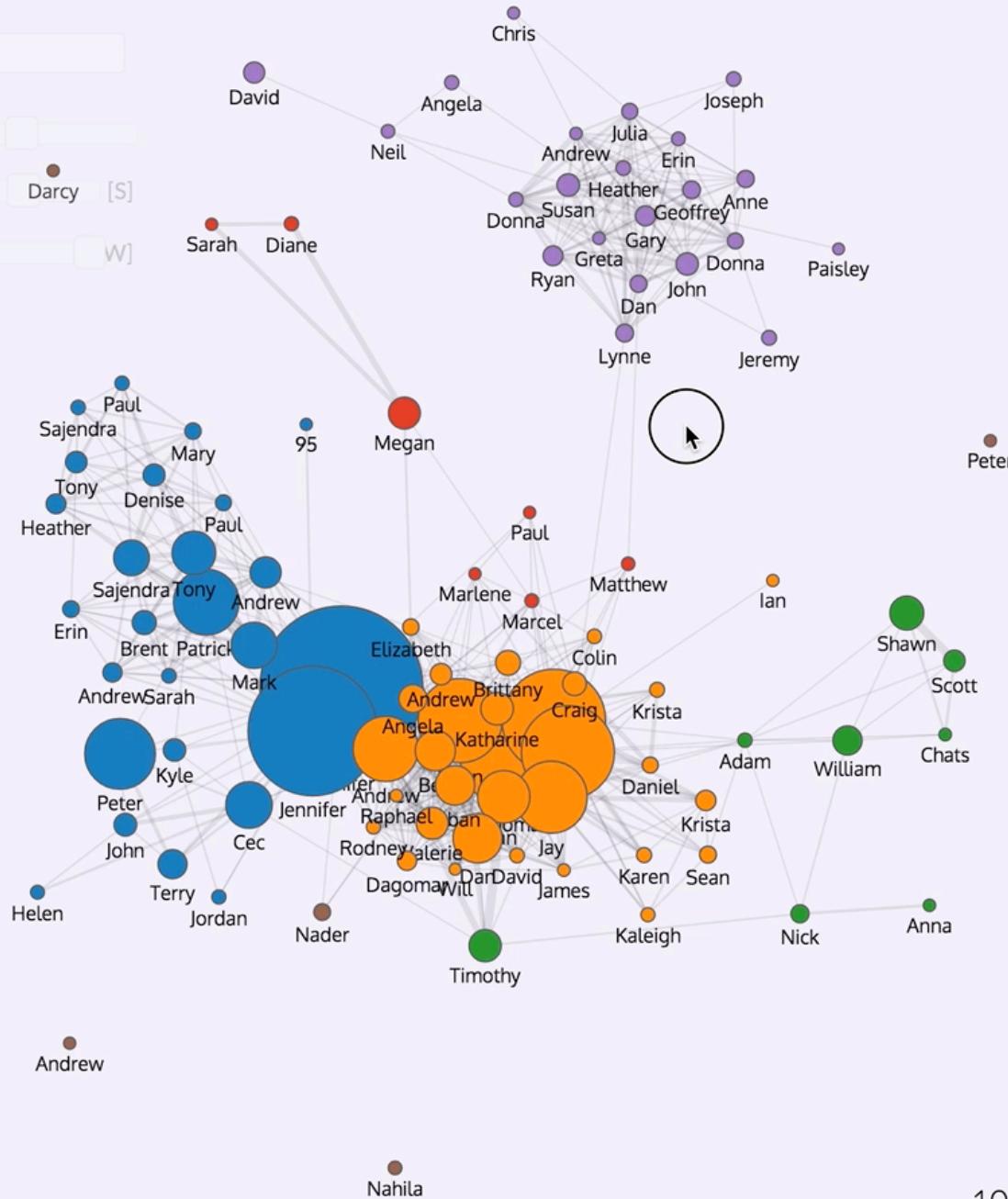
Take a snapshot

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

Feedback?

[Compose](#)

Jeannine



Nahila

10.5 years

26 Sep 2004 - 12 Mar 2015

[All](#) • [Past Year](#) • [Past Month](#) • [Past Week](#)

Ian Milligan

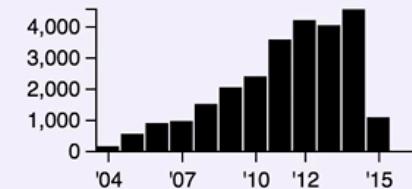


729 collaborators

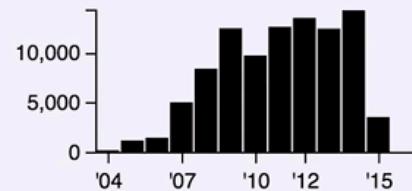
121,854 emails

My Stats[Top Collaborators](#)

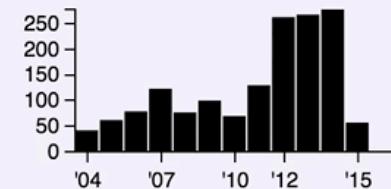
Emails Sent



Emails Received



New Collaborators

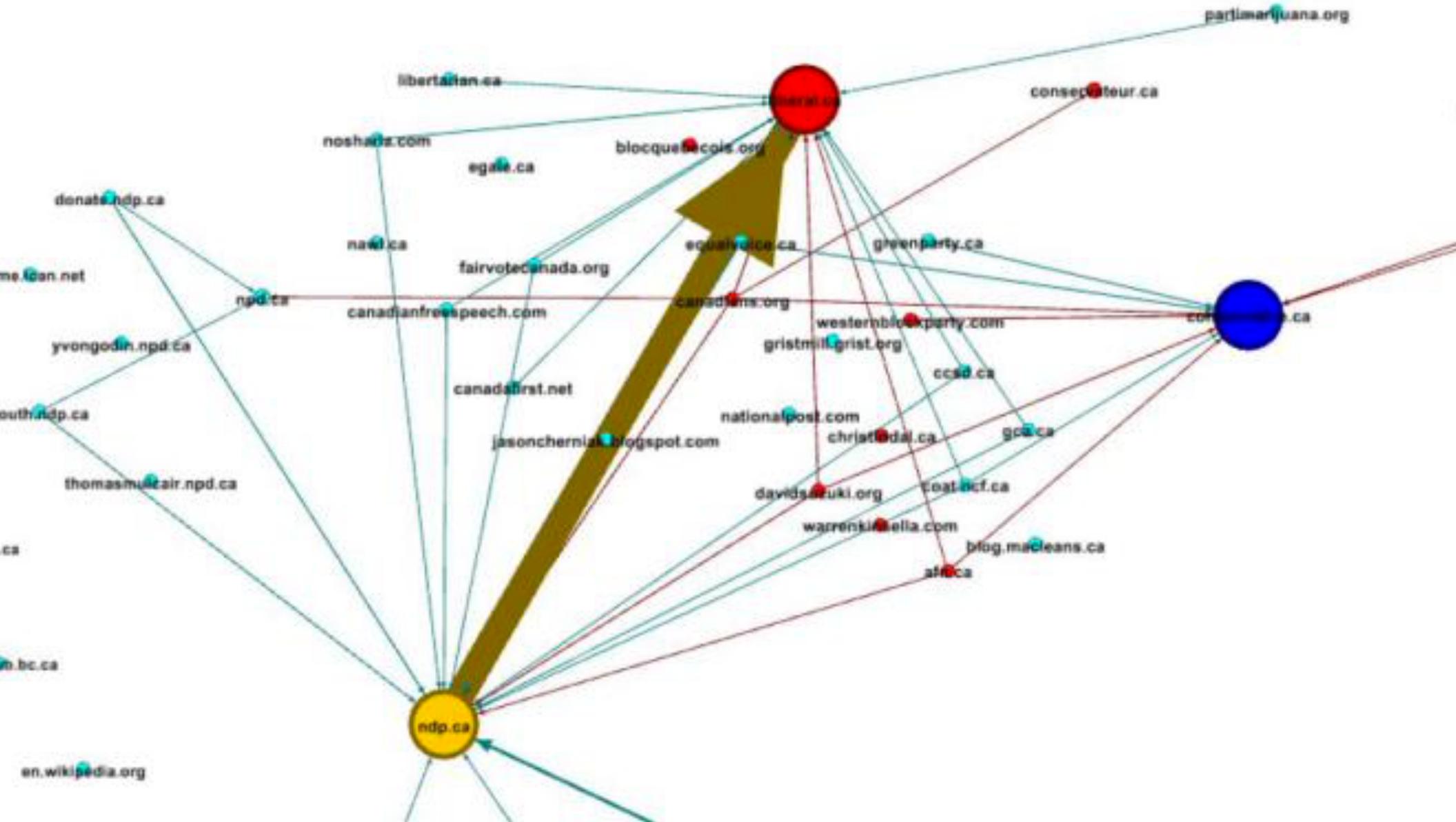


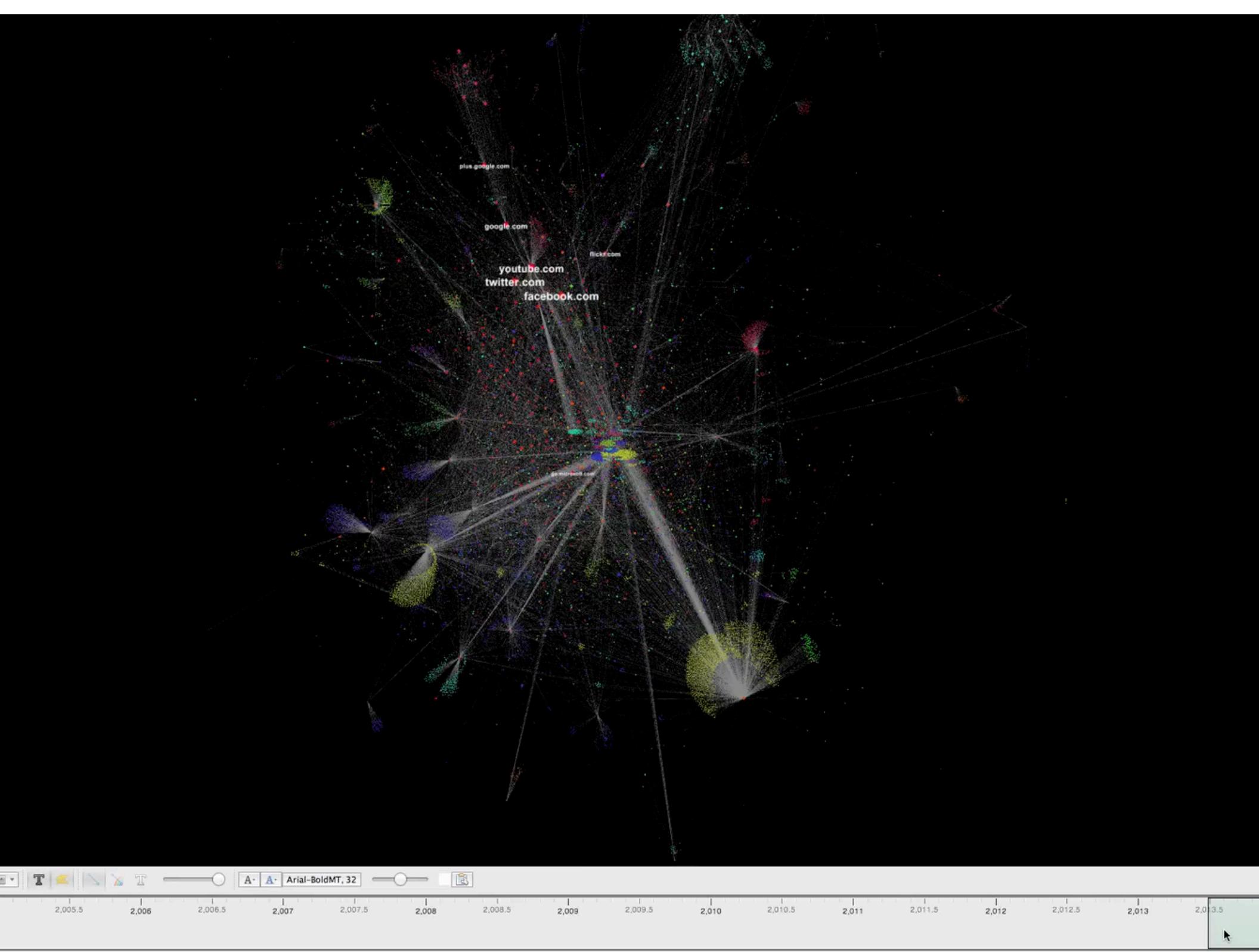


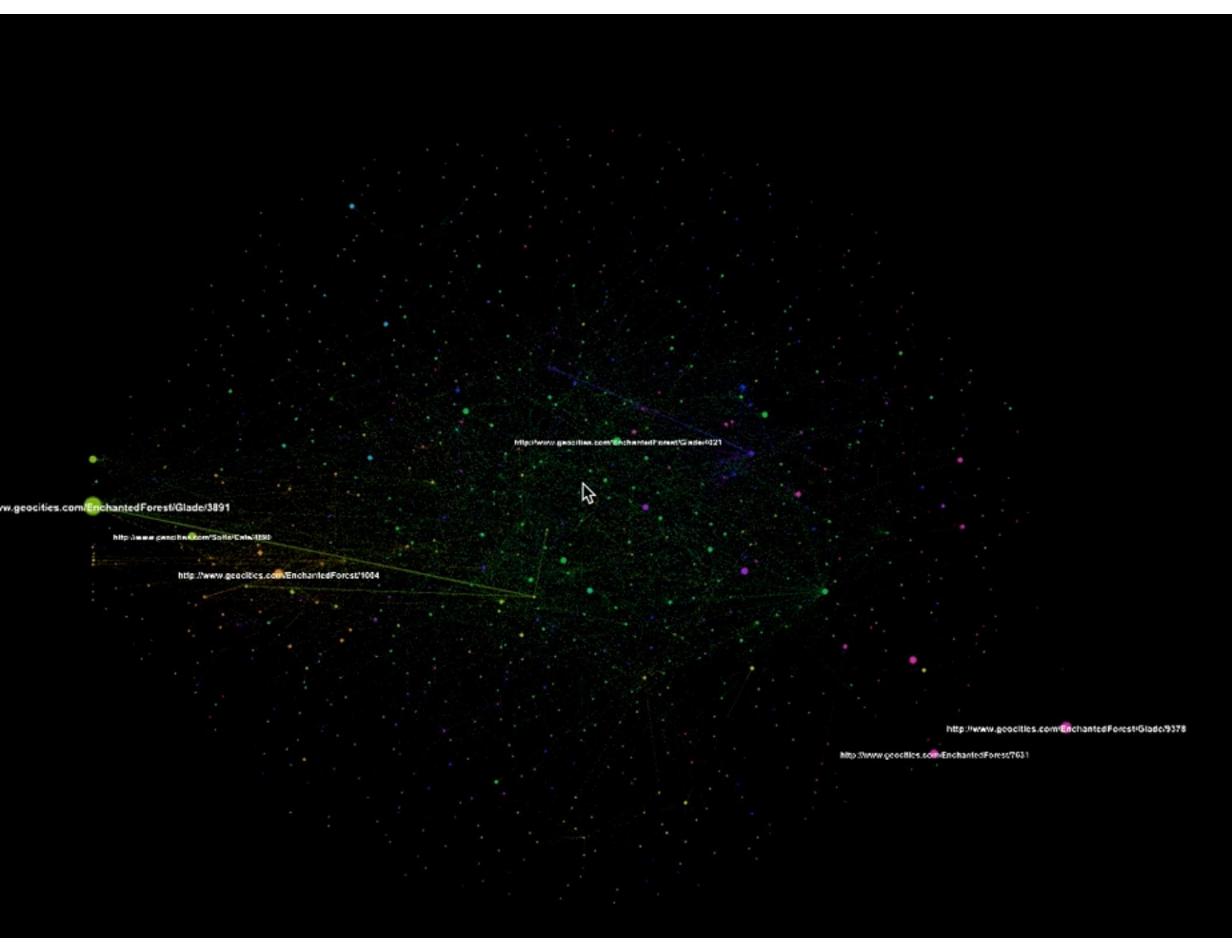
The National Security Agency cares more about metadata than content.

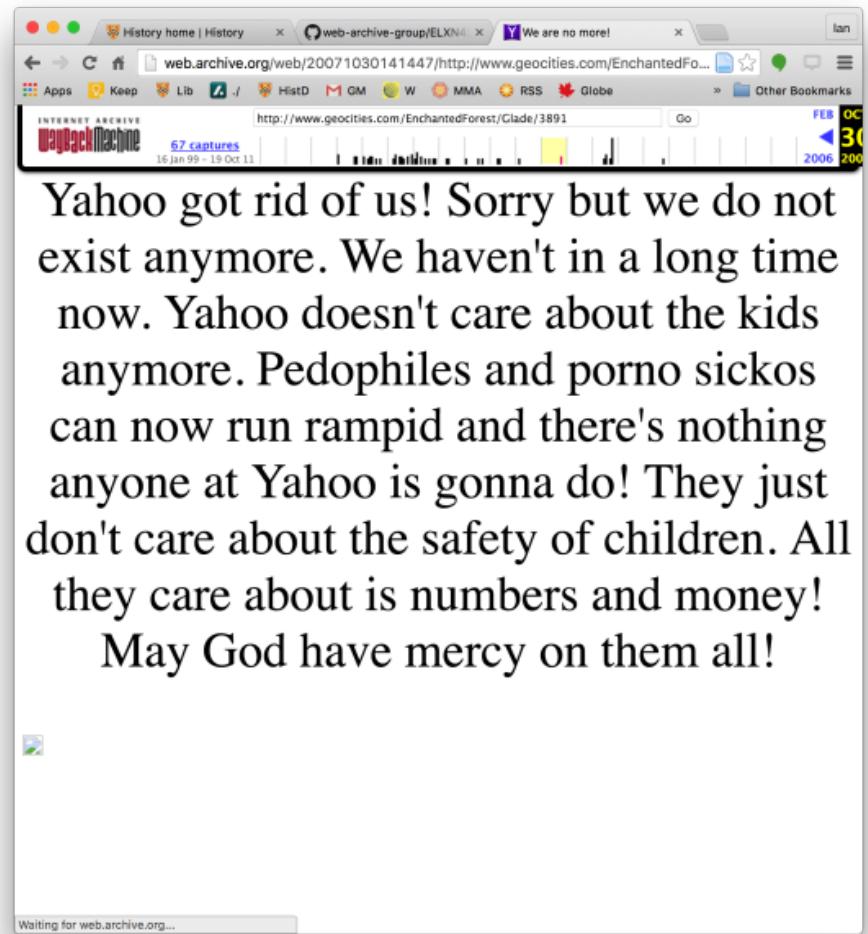
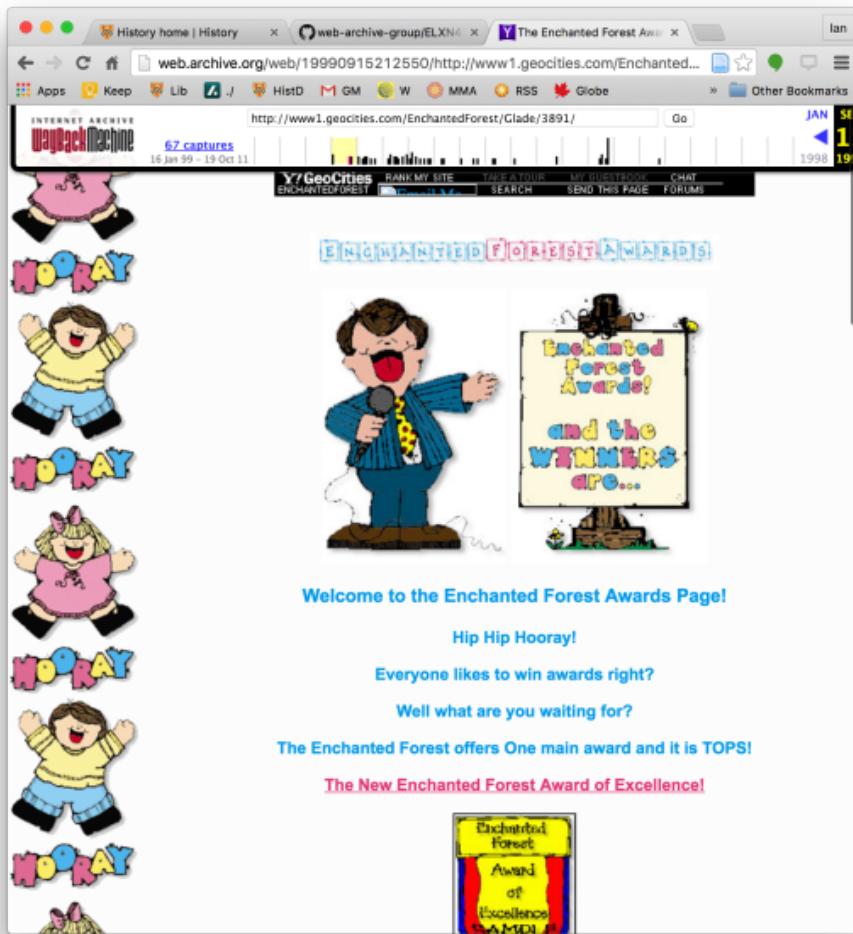
**(this also speaks to the ethical
concerns at play with this material,
another reason to have us here!)**

2005 Canadian Federal Election









**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



The screenshot shows a web browser window with the title bar "Archives Unleashed - Web" and the URL "https://artsweb.uwaterloo.ca/archivesunleashed/". The page content includes a red header "Archives Unleashed", a black navigation bar with links for "HOME", "CALL FOR PARTICIPATION", "SCHEDULE", "PARTICIPANTS", and "ORGANIZERS". Below the navigation is a section titled "Archives Unleashed: Web Archive Hackathon, Ma" (partially cut off). It mentions the location "Toronto, Ontario, Canada: March 3-5, 2016" and describes the workshop's purpose: "This workshop, with the generous support of the Social Sciences and Humanities Research Science Foundation, the University of Waterloo, the University of Toronto, Rutgers University Outaouais, the Internet Archive, Library and Archives Canada, and Compute Canada, presents collaboratively unleash our web collections, exploring cutting-edge research tools while focusing on future directions in web archive analysis." Another section below discusses the hackathon's goals: "This hackathon will bring together a small group of 20-30 researchers to collaboratively develop new tools and approaches to web archives. While there has been considerable discussion about web forums or mechanisms for coordinated, mutually informing development efforts have been few, this is an opportunity to collaboratively unleash our web collections, exploring cutting-edge research tools while focusing on broad-based consensus on future directions in web archive analysis." At the bottom, there is a "Sponsoring Universities" section featuring logos for the University of Waterloo and the University of Toronto.

Case Two: Datathons

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



The screenshot shows a web browser window displaying the 'Archives Unleashed' website. The title bar reads 'Archives Unleashed - Web'. The URL in the address bar is 'https://artsweb.uwaterloo.ca/archivesunleashed/'. The page features a red header with the text 'Archives Unleashed'. Below the header is a navigation menu with links for 'HOME', 'CALL FOR PARTICIPATION', 'SCHEDULE', 'PARTICIPANTS', and 'ORGANIZERS'. The main content area has a large heading 'Archives Unleashed: Web Archive Hackathon, Ma'. Below it, text specifies 'Toronto, Ontario, Canada: March 3 -5, 2016'. A detailed description follows, mentioning the support of various organizations and the goal of the hackathon. Another paragraph describes the purpose of the hackathon, emphasizing collaboration and research tools. At the bottom, there is a section titled 'Sponsoring Universities' with logos for the University of Waterloo and the University of Toronto.

Archives Unleashed

HOME CALL FOR PARTICIPATION SCHEDULE PARTICIPANTS ORGANIZERS

Archives Unleashed: Web Archive Hackathon, Ma

Toronto, Ontario, Canada: March 3 -5, 2016

This workshop, with the generous support of the Social Sciences and Humanities Research Science Foundation, the University of Waterloo, the University of Toronto, Rutgers University Outaouais, the Internet Archive, Library and Archives Canada, and Compute Canada, present collaboratively unleash our web collections, exploring cutting-edge research tools while fos on future directions in web archive analysis.

This hackathon will bring together a small group of 20-30 researchers to collaboratively dev and approaches to web archives. While there has been considerable discussion about web forums or mechanisms for coordinated, mutually informing development efforts have been an opportunity to collaboratively unleash our web collections, exploring cutting-edge resea broad-based consensus on future directions in web archive analysis.

Sponsoring Universities

UNIVERSITY OF WATERLOO

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)



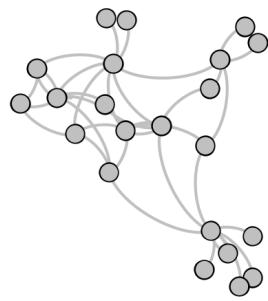
Day One: Socializing



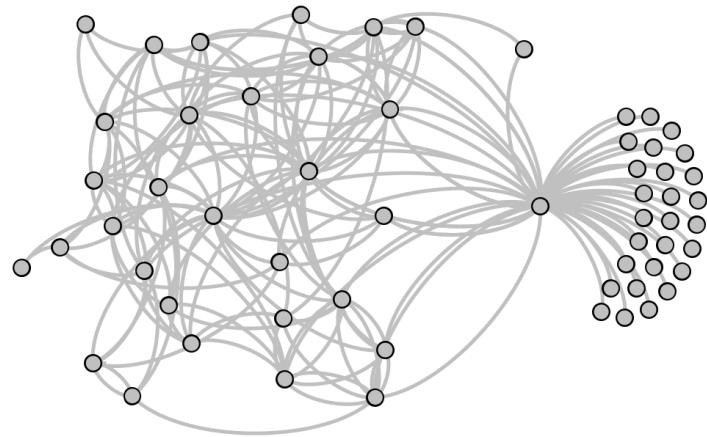
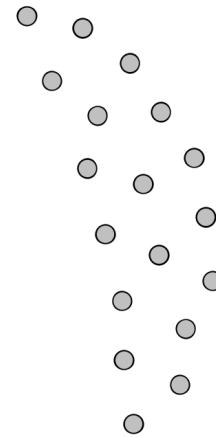
Day Two - Three: Work



Forming Connections



Exchanging ideas *before...*



and after.

Matt Weber + Teis Moller
Kristensen

The Projects

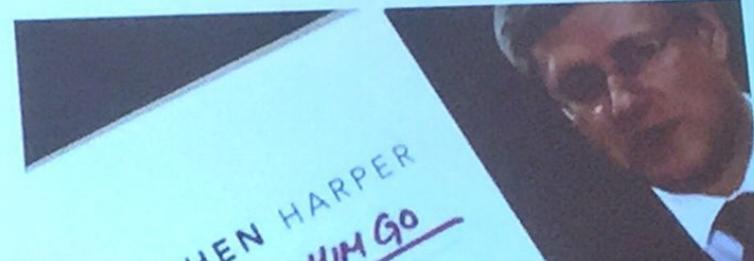
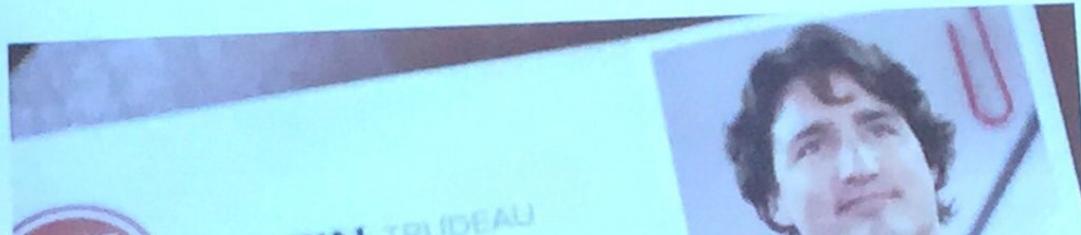
A screenshot of a Mac OS X desktop showing a web browser window. The window title is "hackathon/Projects at master". The address bar shows "GitHub, Inc. [US] https://github.com/web-archive-group/hackathon/tree/master/Projects". The bookmarks bar includes links to Apps, Keep, Lib, GitHub, HistD, GM, RSS, Globe, WALK, LEARN, and FT, along with "Other Bookmarks". The main content area displays a README.md file titled "Project Links". It lists six projects:

- Helge Holzmann, Jaspreet Singh, Vinay Goel - Searching, Mining Everything**
This project enhanced [Archive Spark](#).
Richard Rath, Todd Suomela, Kathrine Cook, and Evan Light - First Nations Representations and Government Data
This project looked at how First Nations were represented as well as government data questions more generally.
- Shane Martin, Eric Oosenbrug, and Jeremy Wiebe - GraphX enhancements for Warcbase.**
They have a GitHub repository [here](#) that provides insight on the D3.js side of the project
- Sawood Alam and Mat Kelly - Interplanetary Wayback (ipwb)**
They have a GitHub repository [here](#) that provides the initial code for integrating WARCs and the IPFS.
- Nathalie Casemajor, Neha Gupta, Petra Galuscakova, Ruqin Ren, Ryan Deschamps, Rosa Iris Rodriguez Rovira, and Sylvain Rocheleau - Canadian Politics**
They have a slidedeck, available [here in this repository](#).
- Kim Pham, Kyle Parry, Emily Maemura, and Niel Chah - "I-know-words-and-images"**
They have a GitHub repository [here](#).
- Alexander Nwala, Allison Hegel, Federico Nanni, Jonathan Armoza, Kelsey Utne, Nick Ruest, Yu Xu - "Tracking Discourse on Social Media"**
They have a slidedeck, available [here as a Google Slides presentation](#). Alternatively, a [PDF version is available here as well](#).

Day Three: Awards

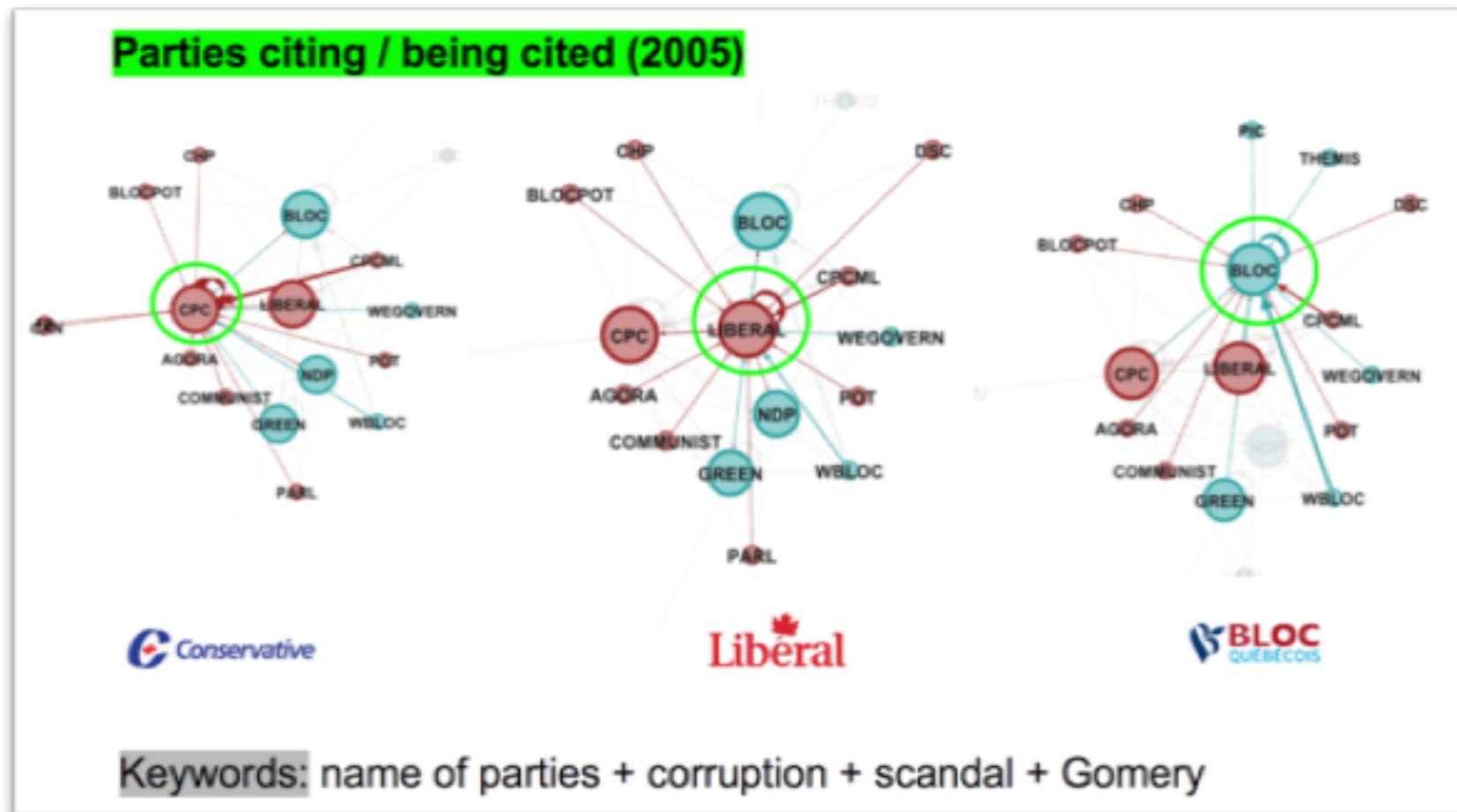
MAGES: NEGATIVE ADS

How often do political parties
use images
of their opponents?



Sample Projects:

How do Canadian political parties represent their opponents?



Sample Projects:

How do Canadian political parties represent their opponents?

FACE DETECTION + NAME DETECTION
=> / DBpedia
=> json

Vision Case Studies Try Our Demo Documentation Support Get a free API Key Visual JSON

Tags
No_tags NaN

Text Within Image
justin trudeau i

Gender
Male 0.999

Celebrity
Justin Trudeau
https://www.dbpedia.org/resource/Justin_Trudeau 0.999
Person Politician MemberOfParliament

Age
35-44 0.431

neutral: 

anger: 

disgust: 

fear: 

happiness: 

sadness: 

This figure displays two screenshots of the Vision API interface. The top screenshot shows the results for a photo of Justin Trudeau, with detected text 'JUSTIN TRUDEAU HIGH RISK' and various demographic and identity details. The bottom screenshot shows the results for a video frame of a person speaking, with a bar chart indicating the six primary emotions: neutral, anger, disgust, fear, happiness, and sadness.

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).

The screenshot shows a GitHub repository page for 'web-archive-group/WALK'. The repository has 20 commits, 1 branch, and 0 releases. The commits are listed as follows:

Commit	Description
Call-Notes	initial commit
Config	updated profile file
MOU	sample MOU, for Issue #4
Scripts	crashed; restarting on additional collections
Warbase	checking in new scripts for April 26
README.md	initial commit

WALK

Project Description

Web Archives for Longitudinal Knowledge (WALK) is a Compute Canada Research Platforms project. Over together 20 core years and 35 TB of storage over the next three years (valued at \$11,635), we will bring Canadian Archive-It partners together with our analytic tools. Right now, there are great silos; WALK will hopefully begin to end that. The team consists of Ian Milligan and Nick Ruest, a great team of librarians and researchers at the University of Alberta (including Geoff Harder, Todd Suomela, Sonya Betz, Peter Binkley, Geoffrey Rockwell, Jefferson Bailey, and John Simpson).



Archive-It Collections

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

The screenshot shows a web browser window with the URL <https://archive-it.org/collections/227>. The page title is "Archive-It - Canadian Political Interest Groups". The header includes links for HOME, EXPLORE, LEARN MORE, and CONTACT US. The main content area displays a collection titled "Canadian Political Parties Groups" collected by "University of Toronto". It notes that the collection was archived since Oct, 2005, and describes it as containing national Canadian political parties, and a number of specific political interest groups. Below this, there's a section titled "Narrow Your Results" with a search bar and buttons for "Sites" and "Search Page Text". A footer at the bottom right shows "Page 1 of 1 (54)" and sorting options.

Explore >> University of Toronto >> Canadian Political Parties and Political Interest Groups

[Canadian Political Parties Groups](#)
Collected by: [University of Toronto](#)
Archived since: Oct, 2005
Description: Canadian Political Parties and Political Interest Groups, national Canadian political parties, and a number of specific political interest groups.
Subject: [Politics & Elections](#)
Collector: [University of Toronto](#)

Narrow Your Results

Enter search terms here

Sites Search Page Text

Page 1 of 1 (54)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Current Interface

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

The screenshot shows a web browser window 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 a header with the Archive-It logo, navigation links for HOME, EXPLORE, LEARN MORE, and CONTACT US, and a tagline about collecting and accessing cultural heritage on the web. Below the header, it says "Explore >> University of Toronto >> Canadian Political Parties and Political Interest Groups". The main content area displays the "Canadian Political Parties and Political Interest Groups" collection, which was collected by the University of Toronto and archived since Oct, 2005. It includes a brief description, subject information (Politics & Elections), and a collector note. A search bar at the bottom allows users to search for specific terms within the collection results.

Building Portals

- Democratizing access so that people can use them.

The screenshot shows a web browser window with the URL <https://archive-it.org/collections/227>. The page title is "Archive-It - Canadian Political Parties and Political Interest Groups". The header includes links for HOME, EXPLORE, LEARN MORE, and CONTACT US. The main content area features a large "ARCHIVE-IT" logo. To the right, there is a summary box for the collection:

- Canadian Political Parties Groups**
- Collected by: [University of Toronto](#)
- Archived since: Oct, 2005
- Description: Canadian Political Parties and Political Interest Groups, national Canadian political parties, and a number of specific political interest groups.
- Subject: [Politics & Elections](#)
- Collector: [University of Toronto](#)

Below this, there is a section titled "Narrow Your Results" with a list of subjects:

- New Democratic Party of Canada (2)
- Assembly of First Nations (1)
- Bloc Québécois (1)
- Canada First (1)
- Canada West Foundation (1)

At the bottom, there are buttons for "More ▾", "Sites", "Search Page Text", and "Page 1 of 1 (54)".

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 lan

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

File	Description	Time
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](#)





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

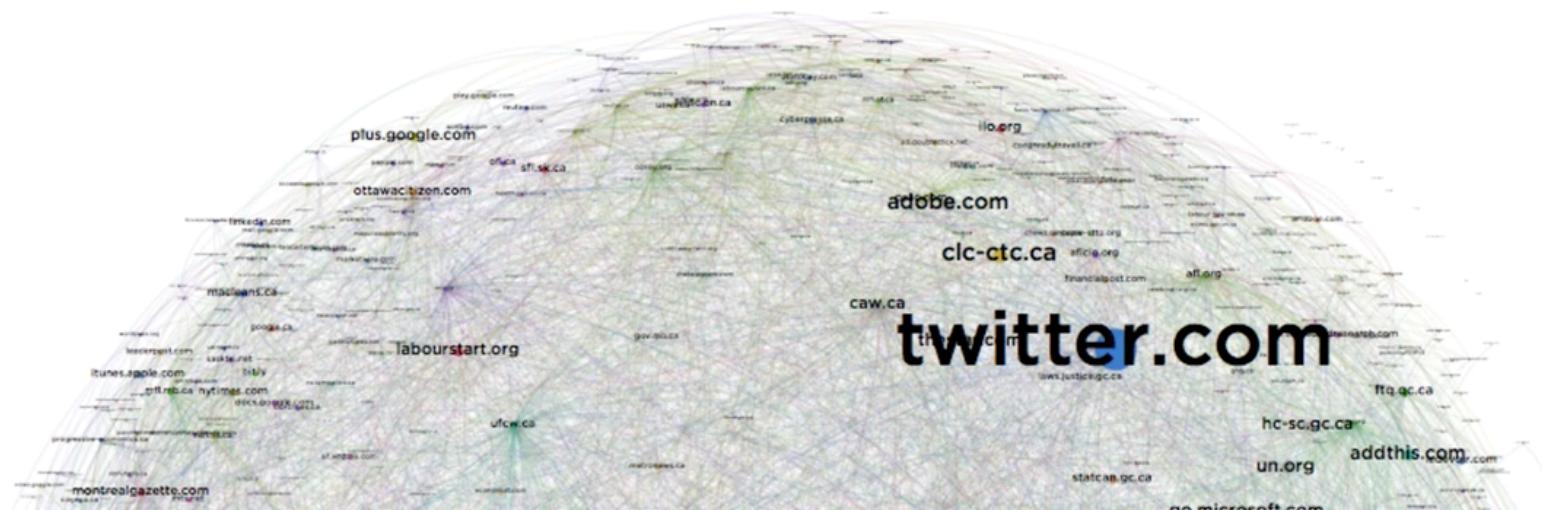
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

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.

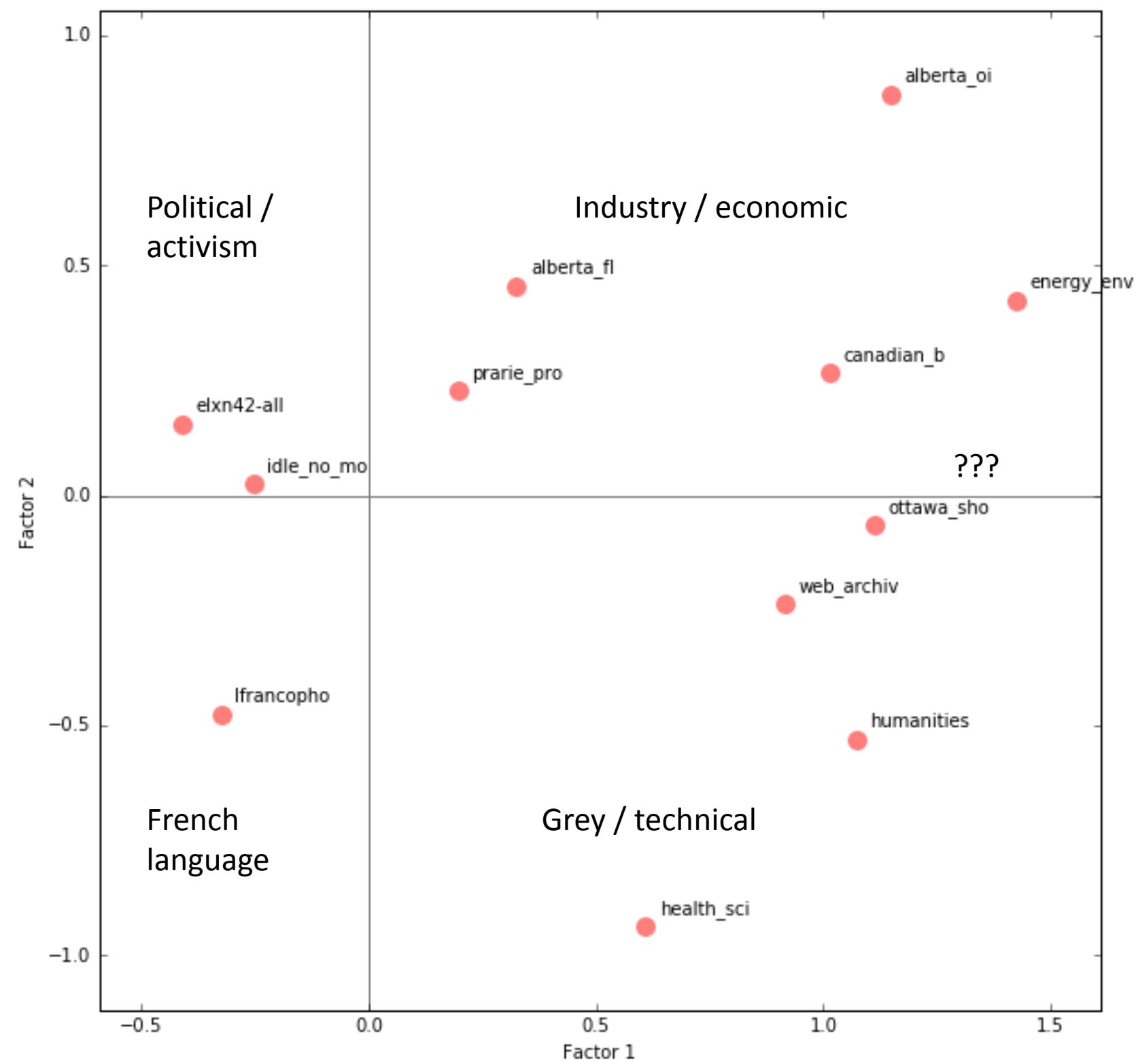
Options include:

- [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]

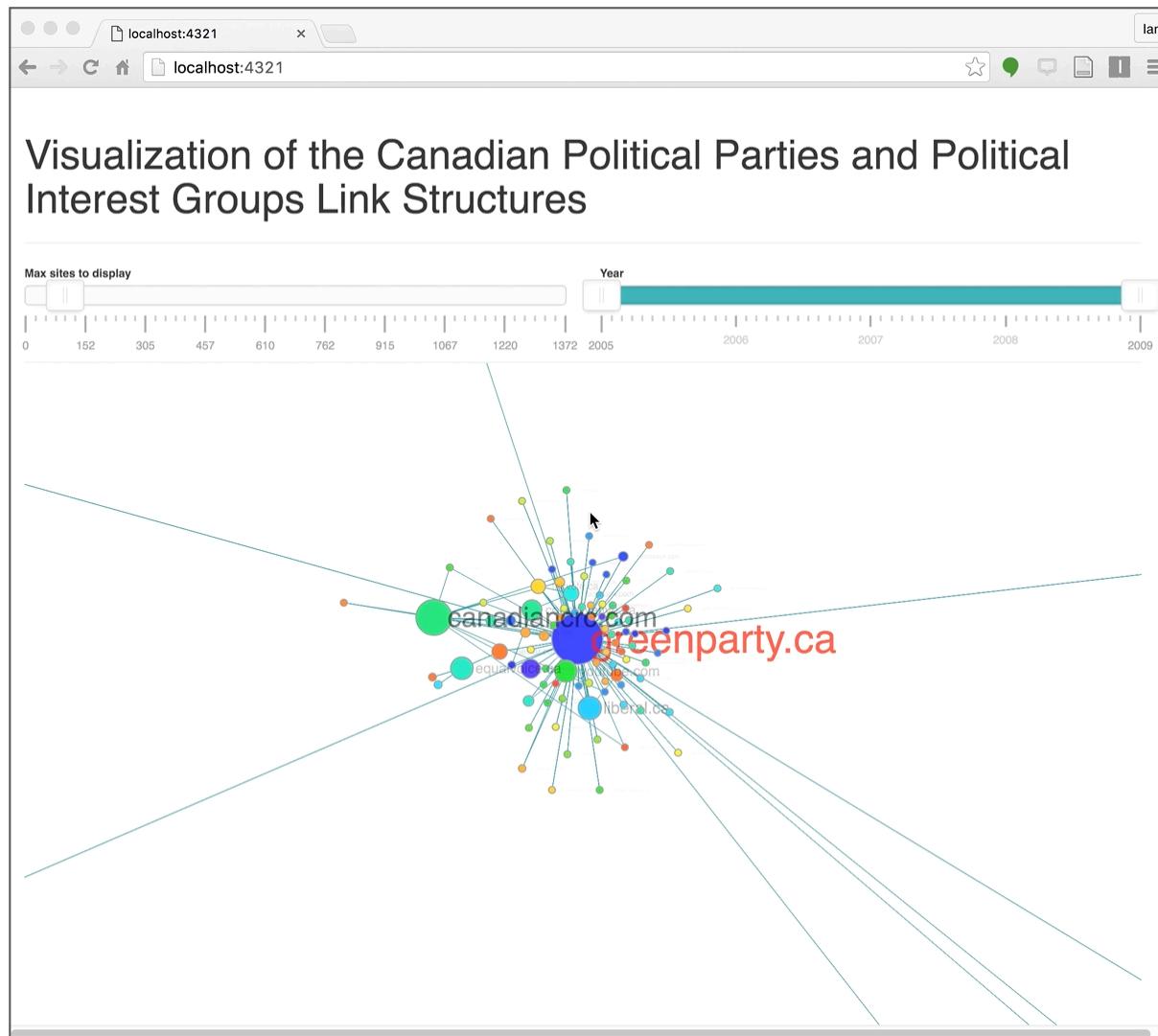
Below, here are all of the links for the entire time period, visualized below.



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



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



compute • calcul
CANADA



UNIVERSITY OF
WATERLOO

Thanks very much!

Questions?

Ian Milligan
Assistant Professor
@ianmilligan1



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History