Researcher Access for Web Archives

Our Experiences

MARAC 2017
Buffalo, NY
27 October 2017

Ian Milligan
Associate Professor
@ianmilligan1

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

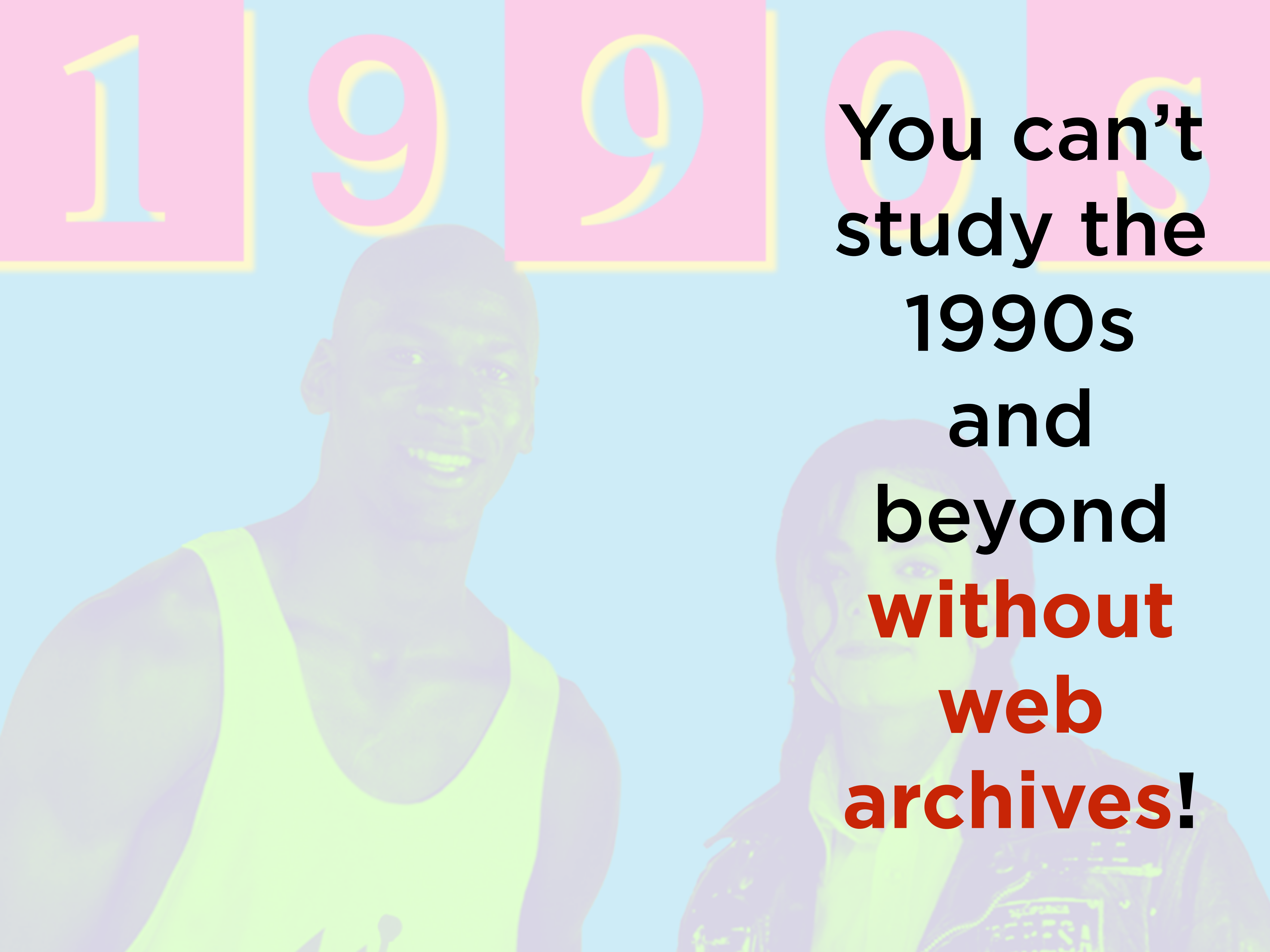
Why does a historian care so much about web archives?

Scarcity

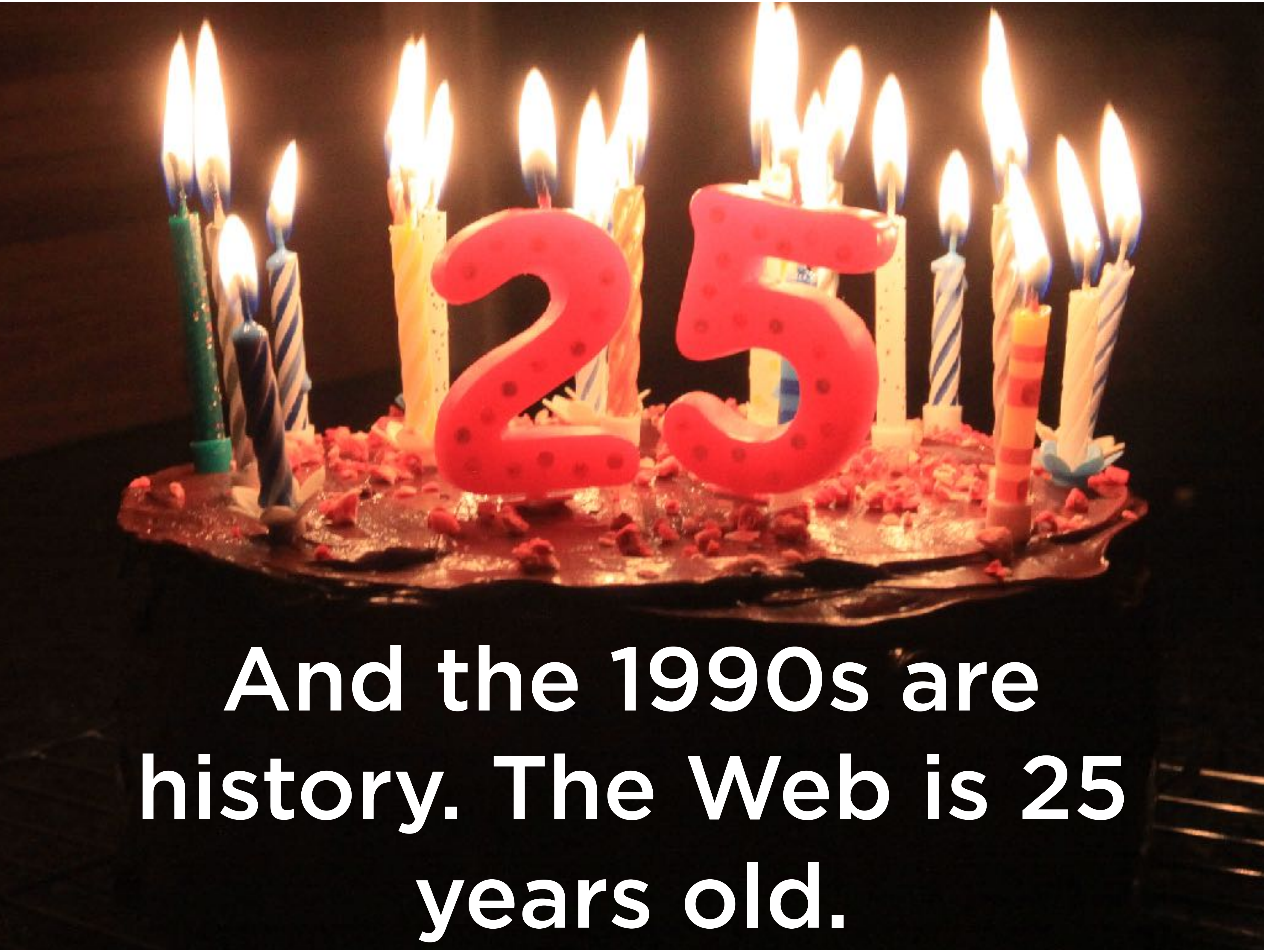
CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series V
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Subseries F Democratic Party
BOX 25
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
374
CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series V
PARTY LEADERSHIP / ADMINISTRATIVE FILE
Subseries F Democratic Party
BOX 30
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
375
370
371
372
373

~~Scarcity~~

Abundance

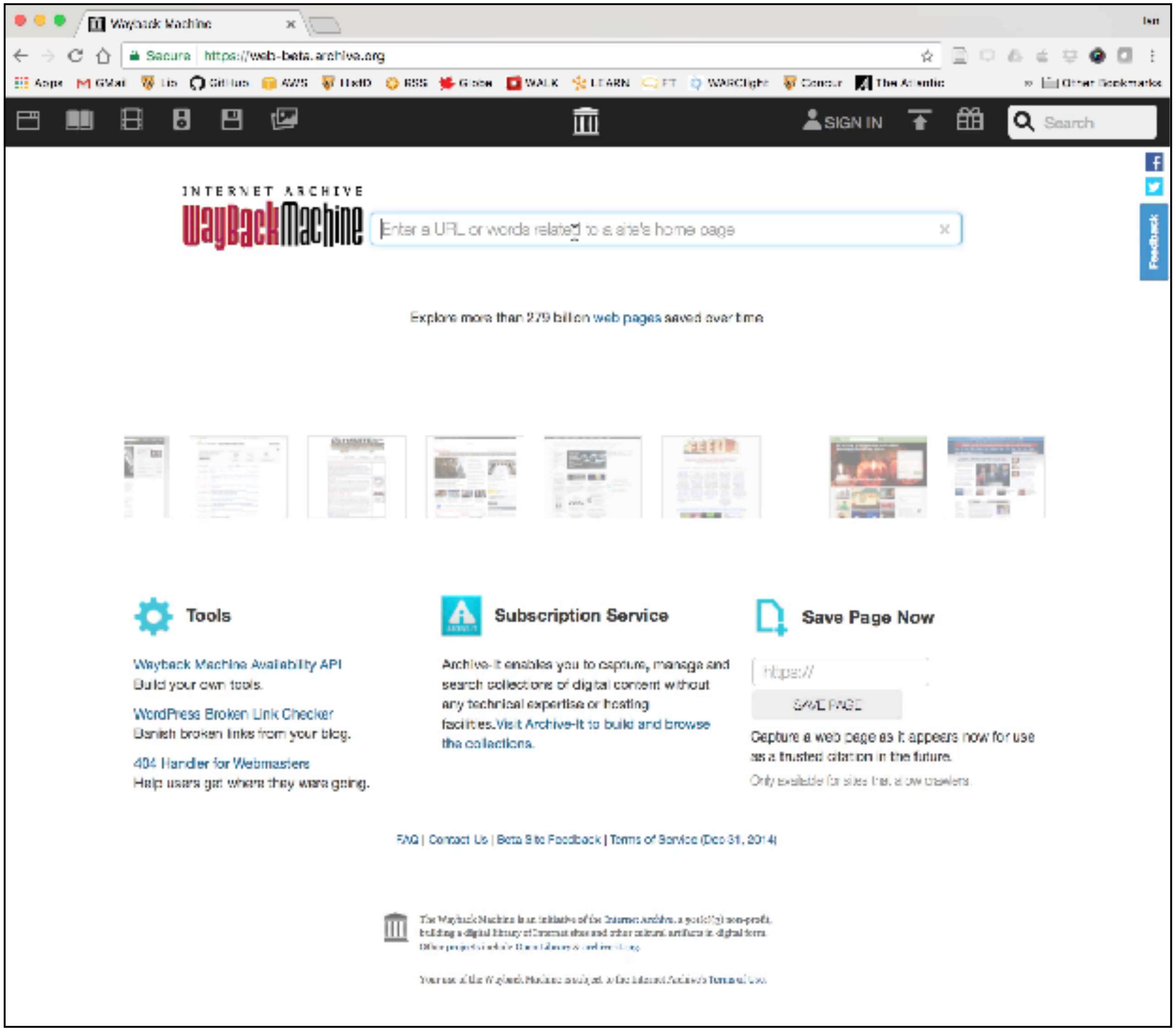


You can’t
study the
1990s
and
beyond
without
web
archives!



And the 1990s are history. The Web is 25 years old.

Current Access Method



Wayback Machine
https://web-beta.archive.org
SIGN IN
Search
INTERNET ARCHIVE
WaybackMachine
Enter a URL or words related to a site's home page
Feedback

Explore more than 279 billion web pages saved over time

Tools

Wayback Machine Availability API
Build your own tools.

WordPress Broken Link Checker
Banish broken links from your blog.

404 Handler for Webmasters
Help users get where they were going.

Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

Save Page Now

https://

SAVE PAGE

Capture a web page as it appears now for use as a trusted citation in the future.

Only available for sites that allow crawlers.

FAQ | Contact Us | Beta Site Feedback | Terms of Service (Dec 31, 2014)

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit, building a digital library of Internet sites and other cultural artifacts in digital form.
Other projects include Open Library & archive-it.org

Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

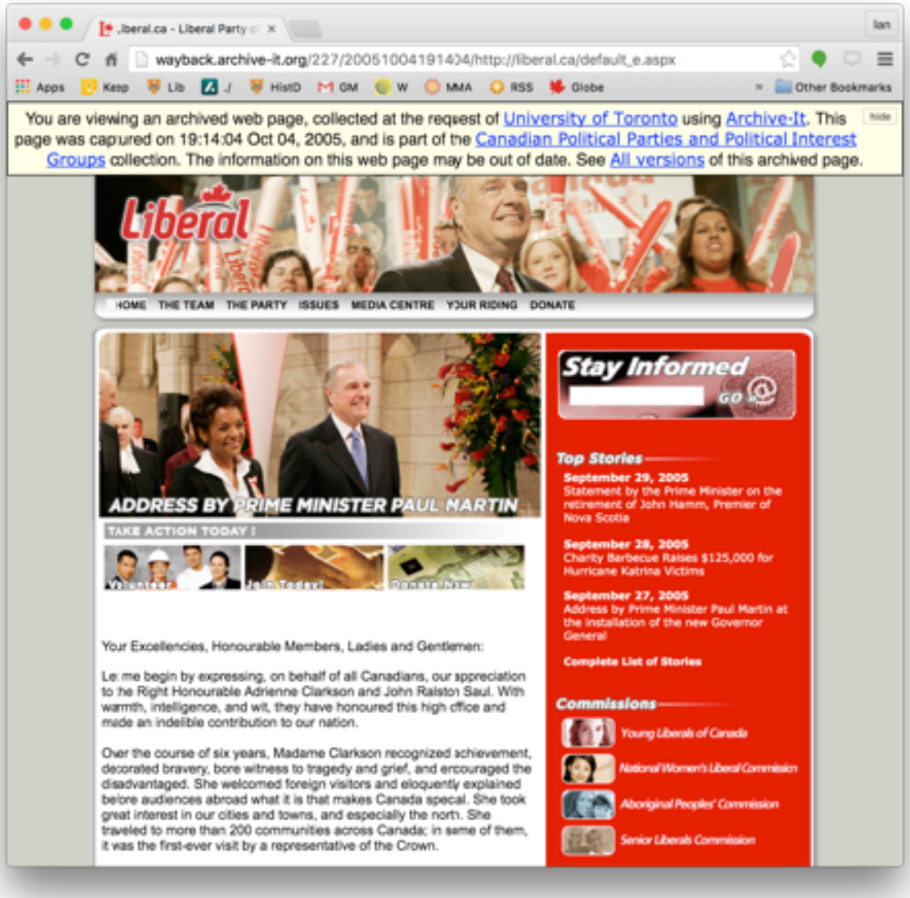
This won’t be enough!

So what can we do?

Ask Nich Worby for data

Canadian Political Parties & Political Interest Group Collection

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



You are viewing an archived web page, collected at the request of University of Toronto using Archive-It. This page was captured on 19:14:04 Oct 04, 2005, and is part of the Canadian Political Parties and Political Interest Groups collection. The information on this web page may be out of date. See All versions of this archived page.

Your Excellencies, Honourable Members, Ladies and Gentlemen:

Let me begin by expressing, on behalf of all Canadians, our appreciation to the Right Honourable Adrienne Clarkson and John Ralston Saul. With warmth, intelligence, and wit, they have honoured this high office and made an indelible contribution to our nation.

Over the course of six years, Madame Clarkson recognized achievement, decorated bravery, bore witness to tragedy and grief, and encouraged the disadvantaged. She welcomed foreign visitors and eloquently explained before audiences abroad what it is that makes Canada special. She took great interest in our cities and towns, and especially the north. She traveled to more than 200 communities across Canada; in some of them, it was the first-ever visit by a representative of the Crown.

Top Stories

September 29, 2005
Statement by the Prime Minister on the retirement of John Hamm, Premier of Nova Scotia

September 28, 2005
Charity Barbecue Raises $125,000 for Hurricane Katrina Victims

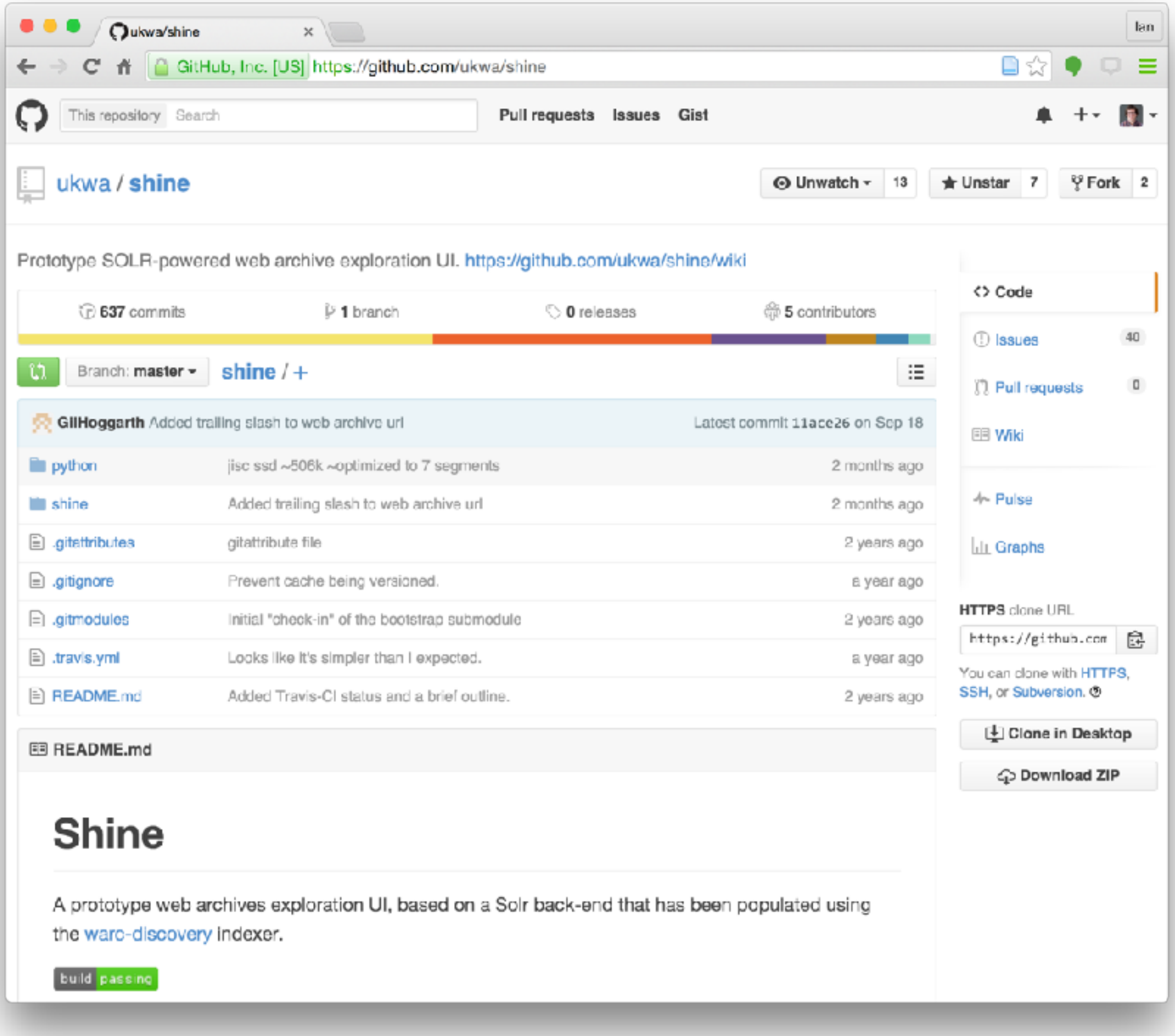
September 27, 2005
Address by Prime Minister Paul Martin at the installation of the new Governor General

Complete List of Stories

Commissions

Young Liberals of Canada
National Women's Liberal Commission
Aboriginal Peoples' Commission
Senior Liberals Commission

How could we provide better access?



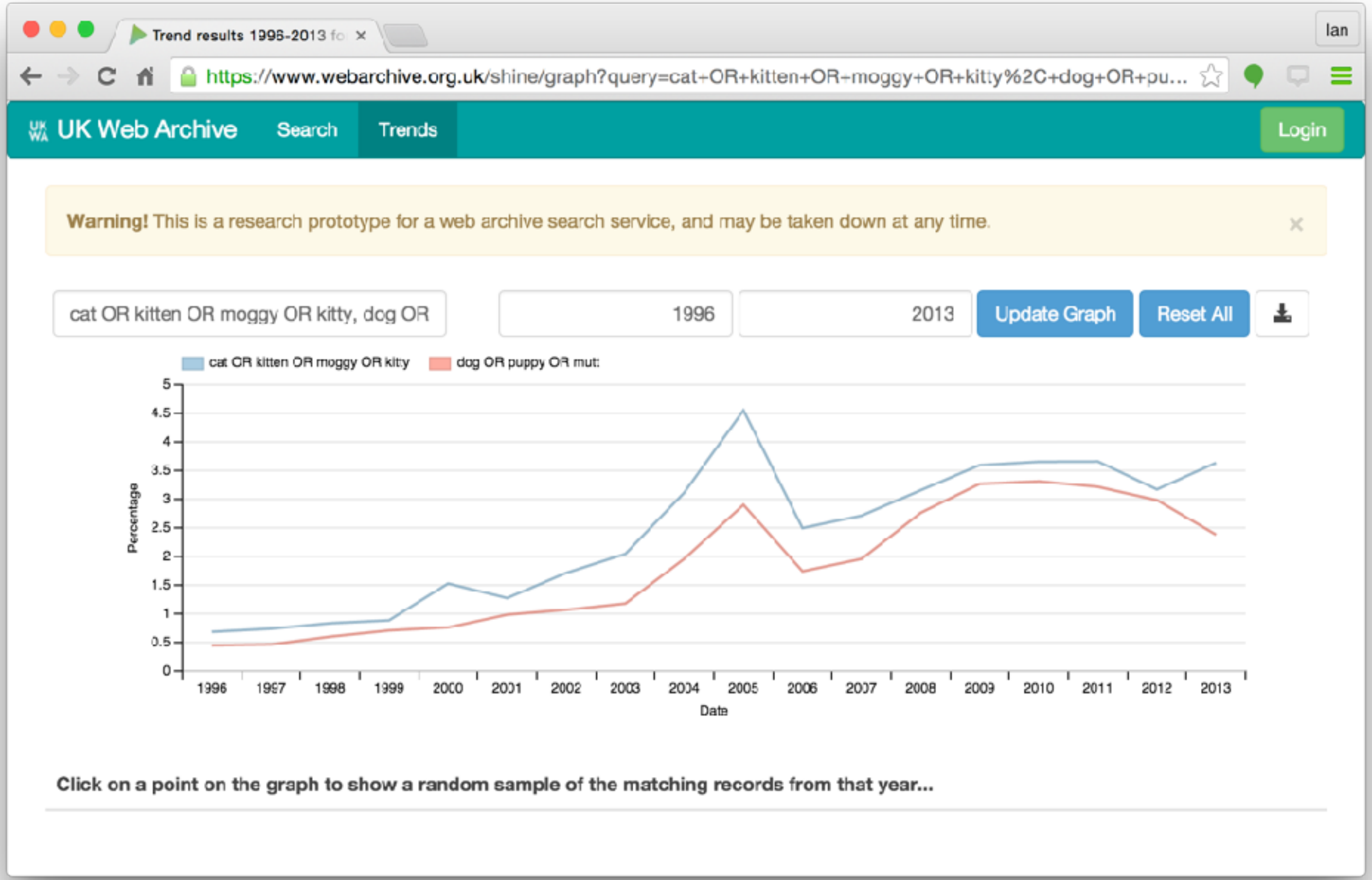
Prototype SOLR-powered web archive exploration UI. https://github.com/ukwa/shine/wiki

You can clone with HTTPS, SSH, or Subversion.

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

Shine

A prototype web archives exploration UI, based on a Solr back-end that has been populated using the warc-discovery indexer.



Trend results 1996-2013 fo
Ian
https://www.webarchive.org.uk/shine/graph?query=cat+OR+kitten+OR+moggy+OR+kitty%2C+dog+OR+pu...
UK WA UK Web Archive
Search
Trends
Login

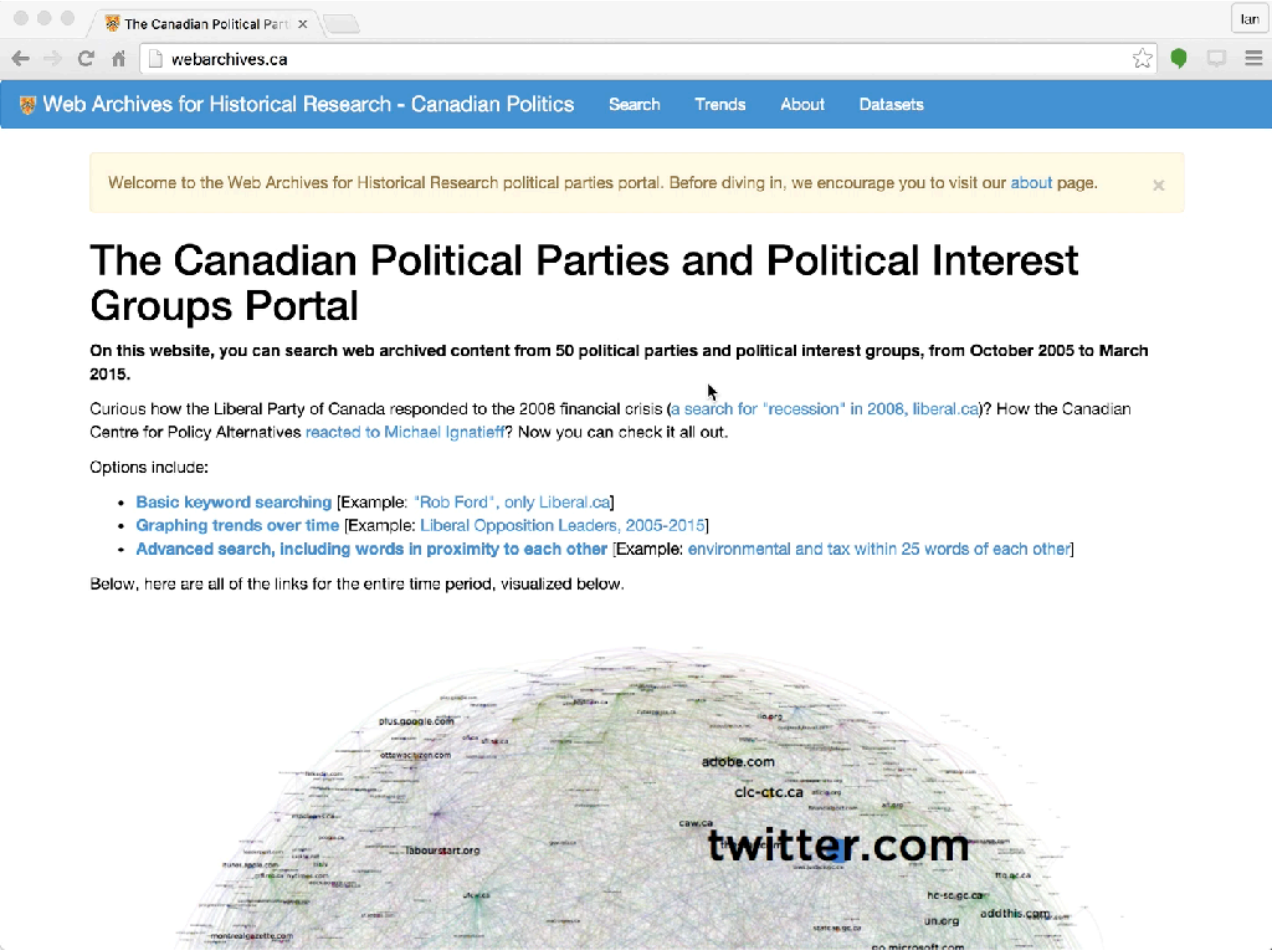
Warning! This is a research prototype for a web archive search service, and may be taken down at any time.

cat OR kitten OR moggy OR kitty, dog OR

1996

2013

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



Web Archives for Historical Research - Canadian Politics Search Trends About Datasets

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.

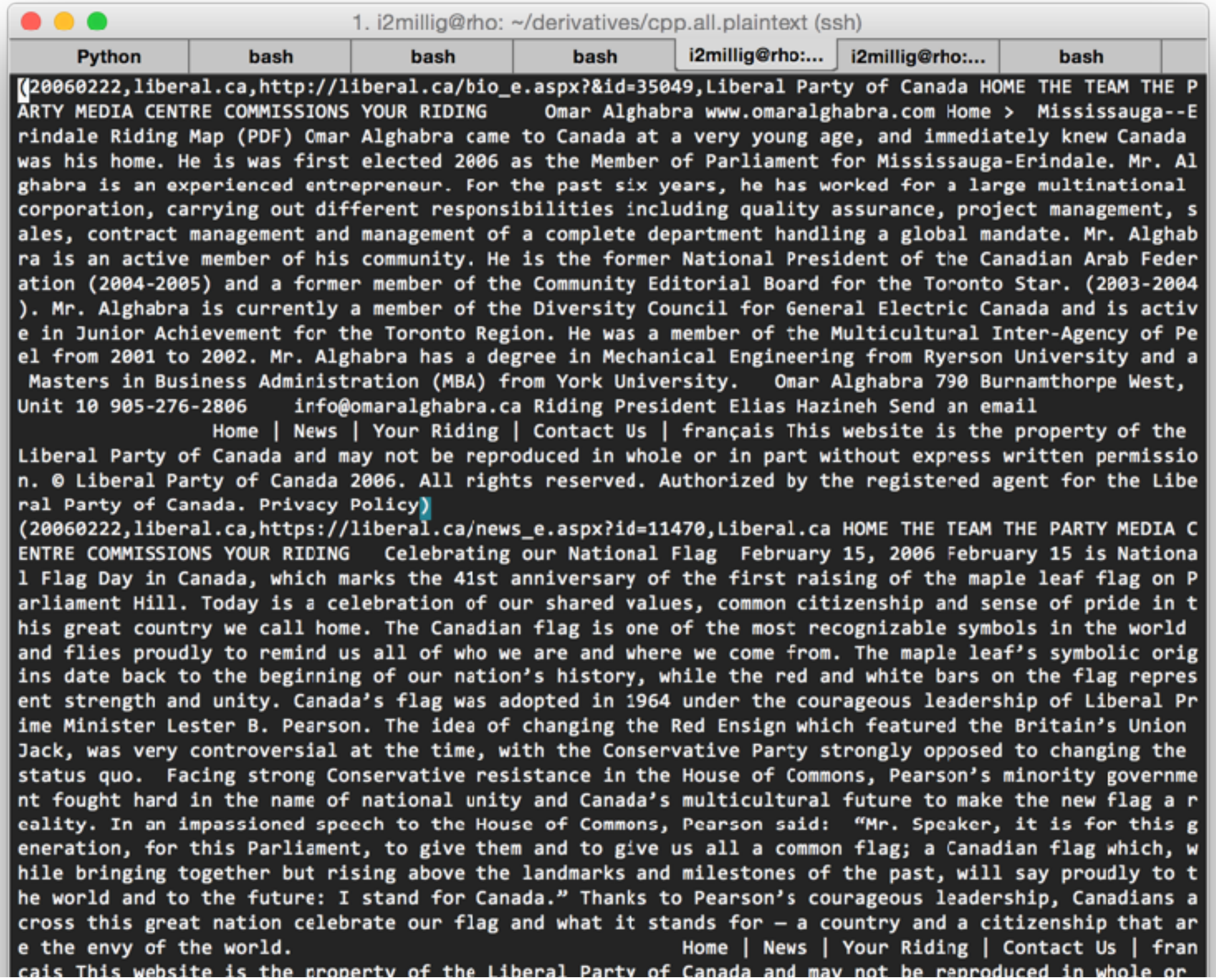
Options include:

- **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.

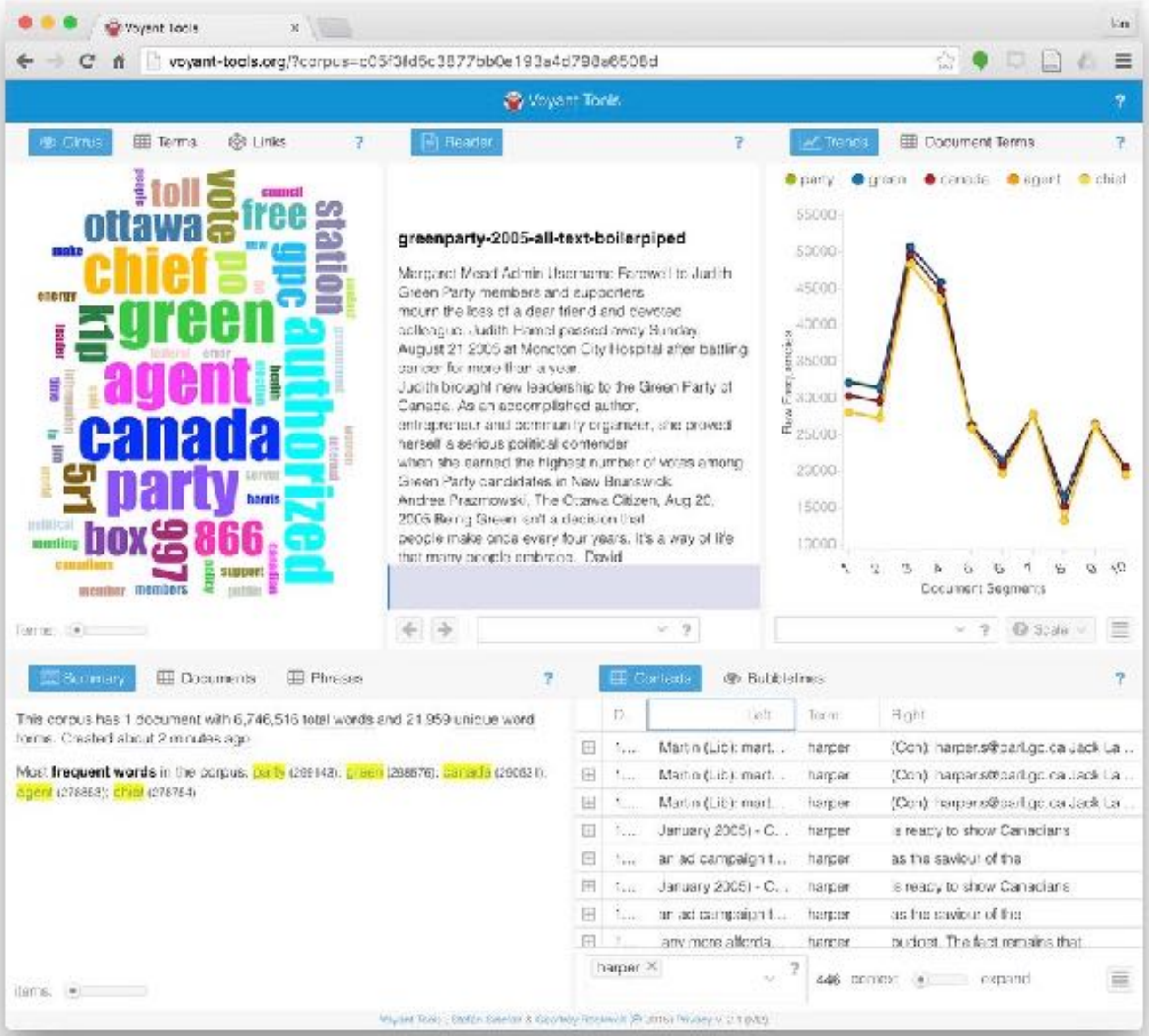
We also wanted to explore derivative datasets for researcher access with our new project.

Extract all Text



```
(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 is 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
el 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 permissio
n. © 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



greenparty-2005-all-text-boilerpiped

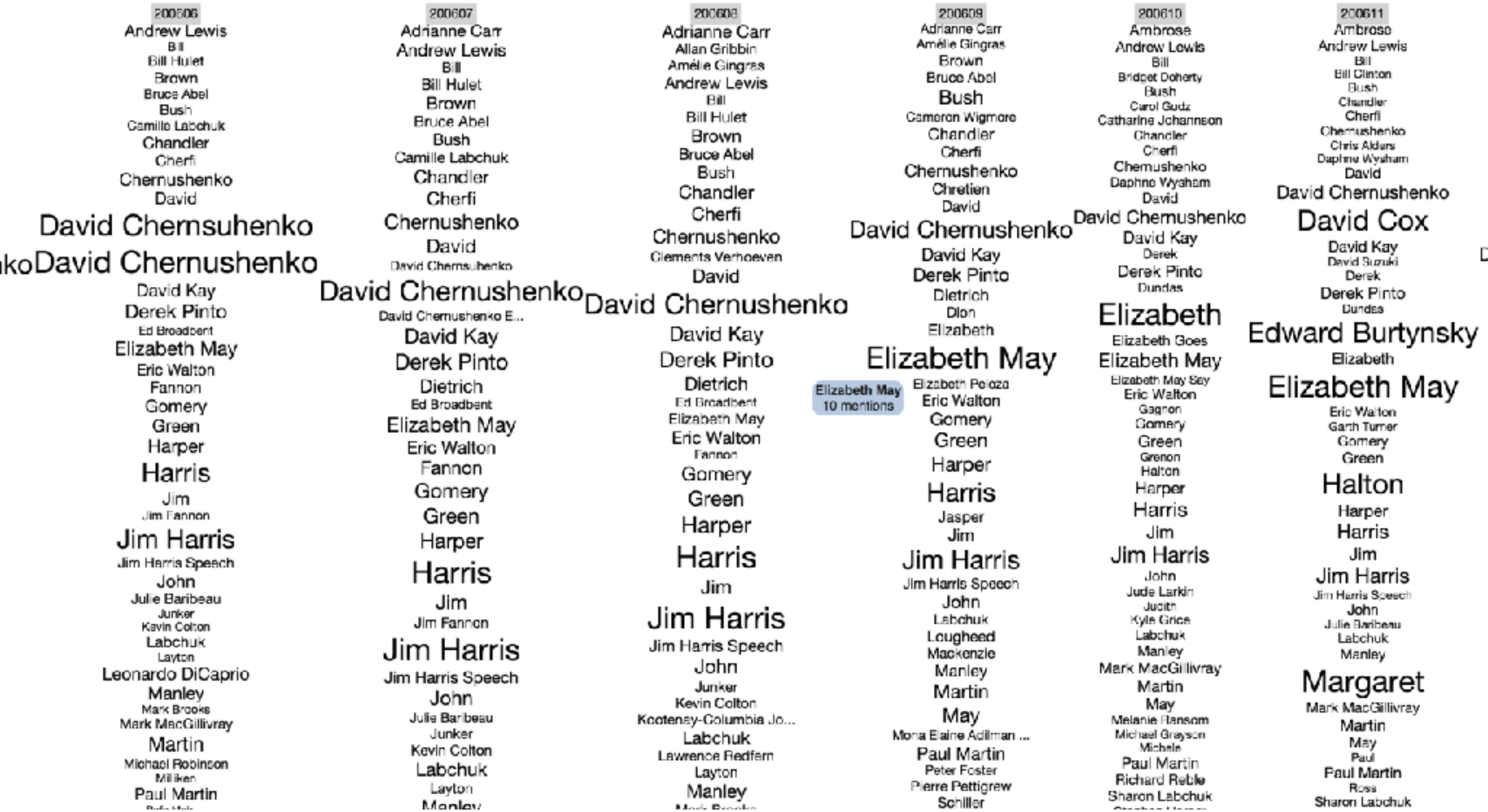
Margaret Mead Admin Username Farewell to Judith
Green Party members and supporters
mourn the loss of a dear friend and devoted
colleague. Judith Hamel passed away Sunday
August 21 2005 at Moncton City Hospital after battling
cancer for more than a year.
Judith brought new leadership to the Green Party of
Canada. As an accomplished author,
entrepreneur and community organizer, she proved
herself a serious political contender
when she earned the highest number of votes among
Green Party candidates in New Brunswick.
Andrea Prazmowski, The Ottawa Citizen, Aug 20,
2005 Being Green isn't a decision that
people make once every four years. It's a way of life
that many people embrace. David

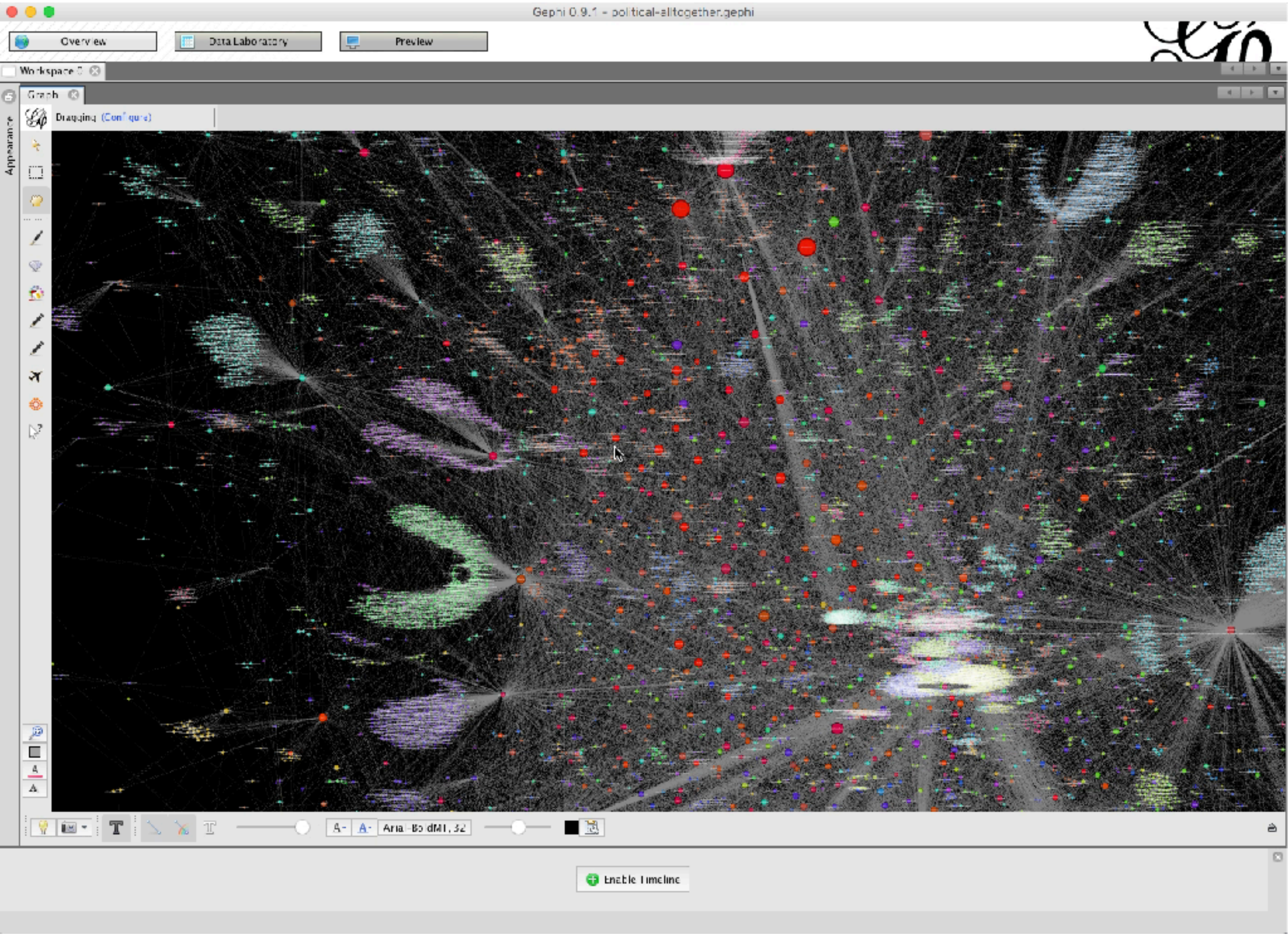
This corpus has 1 document with 6,746,516 total words and 21,959 unique word forms. Created about 2 minutes ago.

Most **frequent words** in the corpus: party (289142); green (288676); canada (280631); agent (278868); chief (278754)

D	Left	Term	Right
1...	Martin (Lib): mart...	harper	(Con): harper.s@parl.gc.ca Jack La...
1...	Martin (Lib): mart...	harper	(Con): harper.s@parl.gc.ca Jack La...
1...	Martin (Lib): mart...	harper	(Con): harper.s@parl.gc.ca Jack La...
1...	January 2005) - C...	harper	is ready to show Canadians
1...	an ad campaign t...	harper	as the saviour of the
1...	January 2005) - C...	harper	is ready to show Canadians
1...	an ad campaign t...	harper	as the saviour of the
1...	any more afforda...	harper	budget. The fact remains that

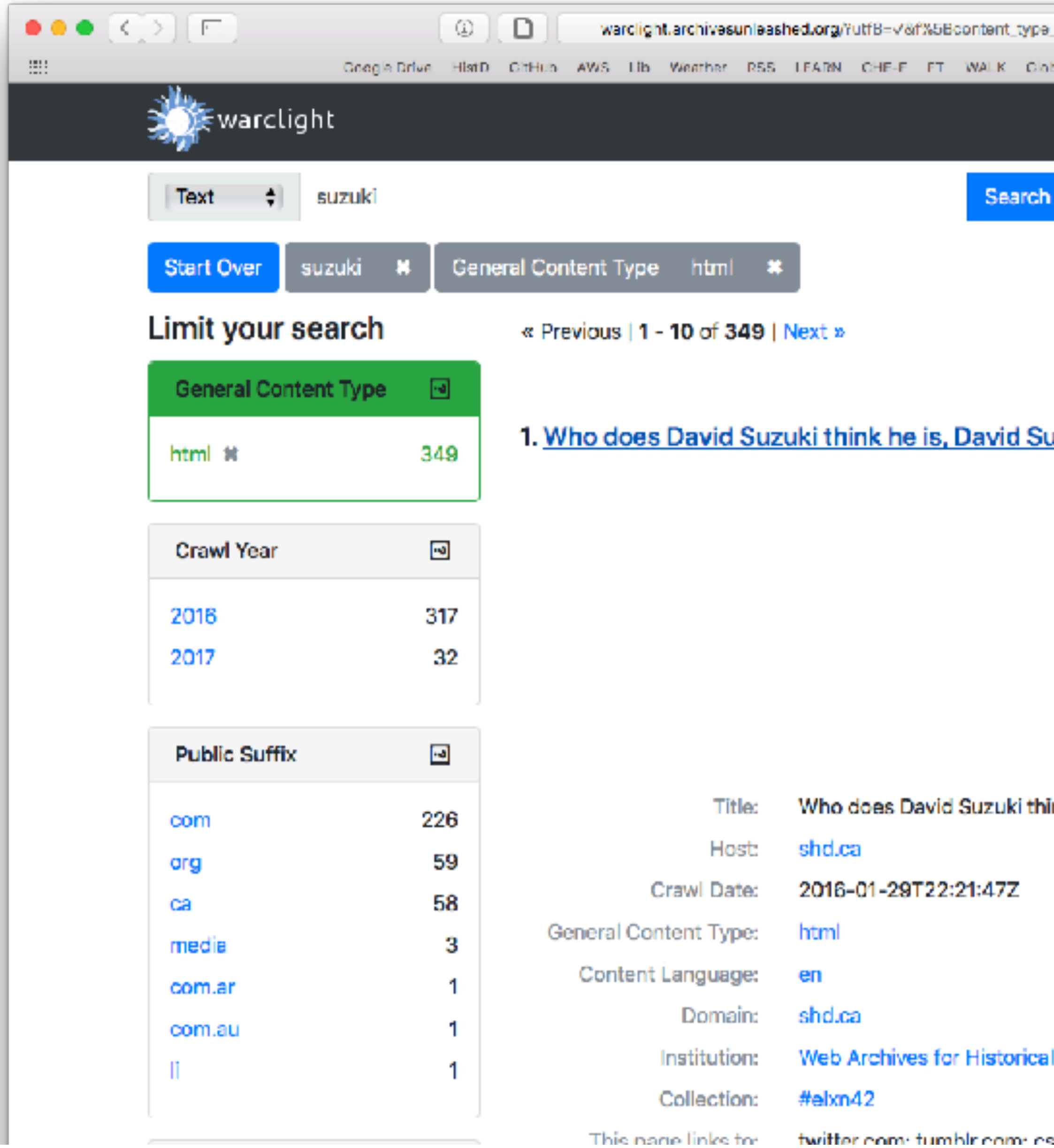
Extract Entities



Gephi 0.9.1 - political-alltogether.gephi
Overview
Data Laboratory
Preview
Workspace 0
Graph
Appearance
Dragging (Configure)
Arial-BoldMT, 32
Enable Timeline

Next Step

- The **Archives Unleashed Cloud**
  - Under development - moving Shine to a Blacklight-based system for community involvement;
  - A GUI-based workbench for uploading WARCs or using Internet Archive APIs to extract derivatives from your collections to do text analysis, networks, other DH work.
- **Set of regional events and workshops - with travel funding - stay tuned!**



suzuki

suzuki ✖ | General Content Type html ✖

« Previous | 1 - 10 of 349 | Next »

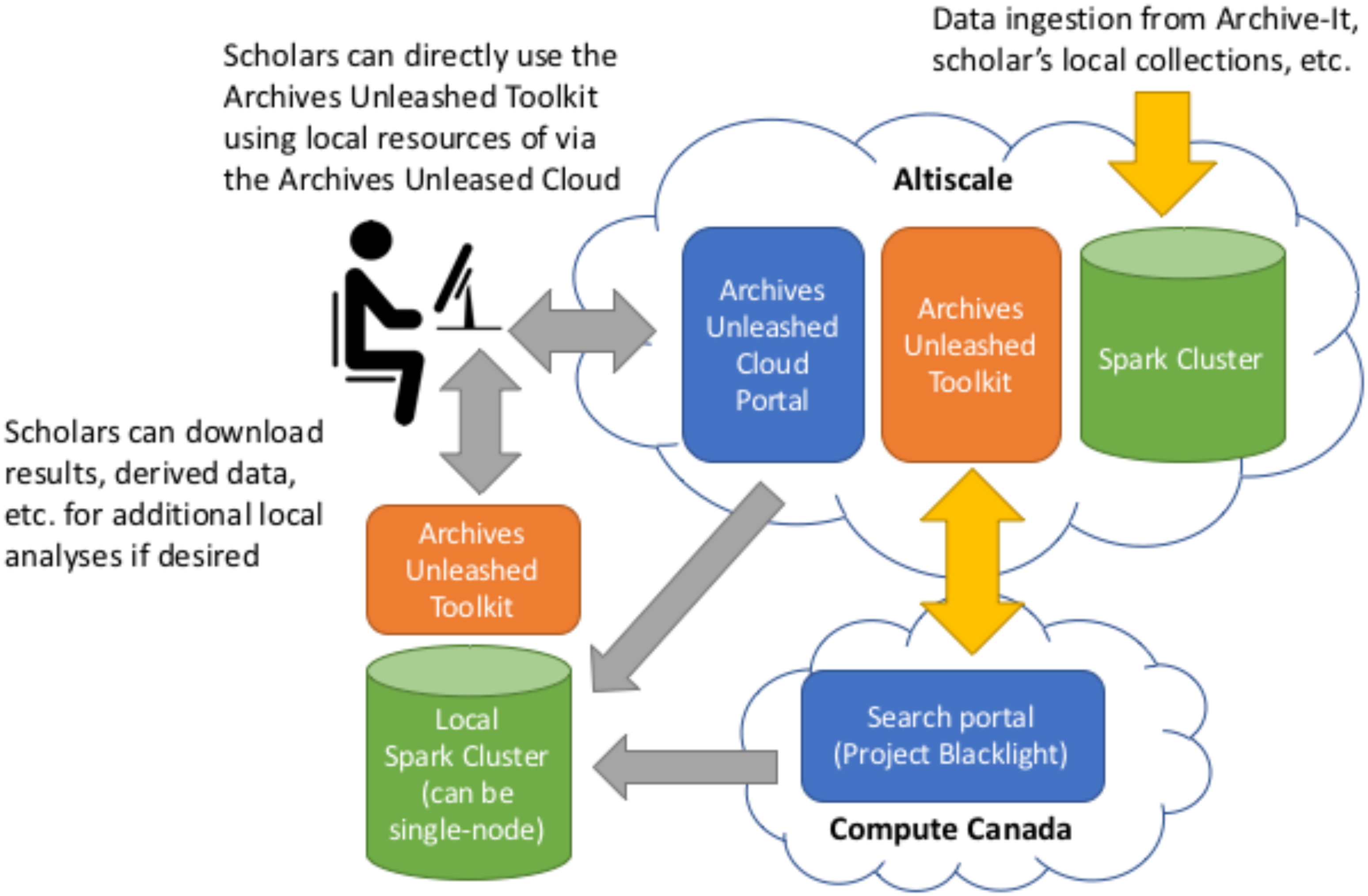
General Content Type	
html ✖	349

Crawl Year	
2016	317
2017	32

Public Suffix	
com	226
org	59
ca	58
media	3
com.ar	1
com.au	1
li	1

1. Who does David Suzuki think he is, David Su

Title:	Who does David Suzuki thi
Host:	shd.ca
Crawl Date:	2016-01-29T22:21:47Z
General Content Type:	html
Content Language:	en
Domain:	shd.ca
Institution:	Web Archives for Historical
Collection:	#elxn42
This page links to:	twitter.com; tumblr.com; cs

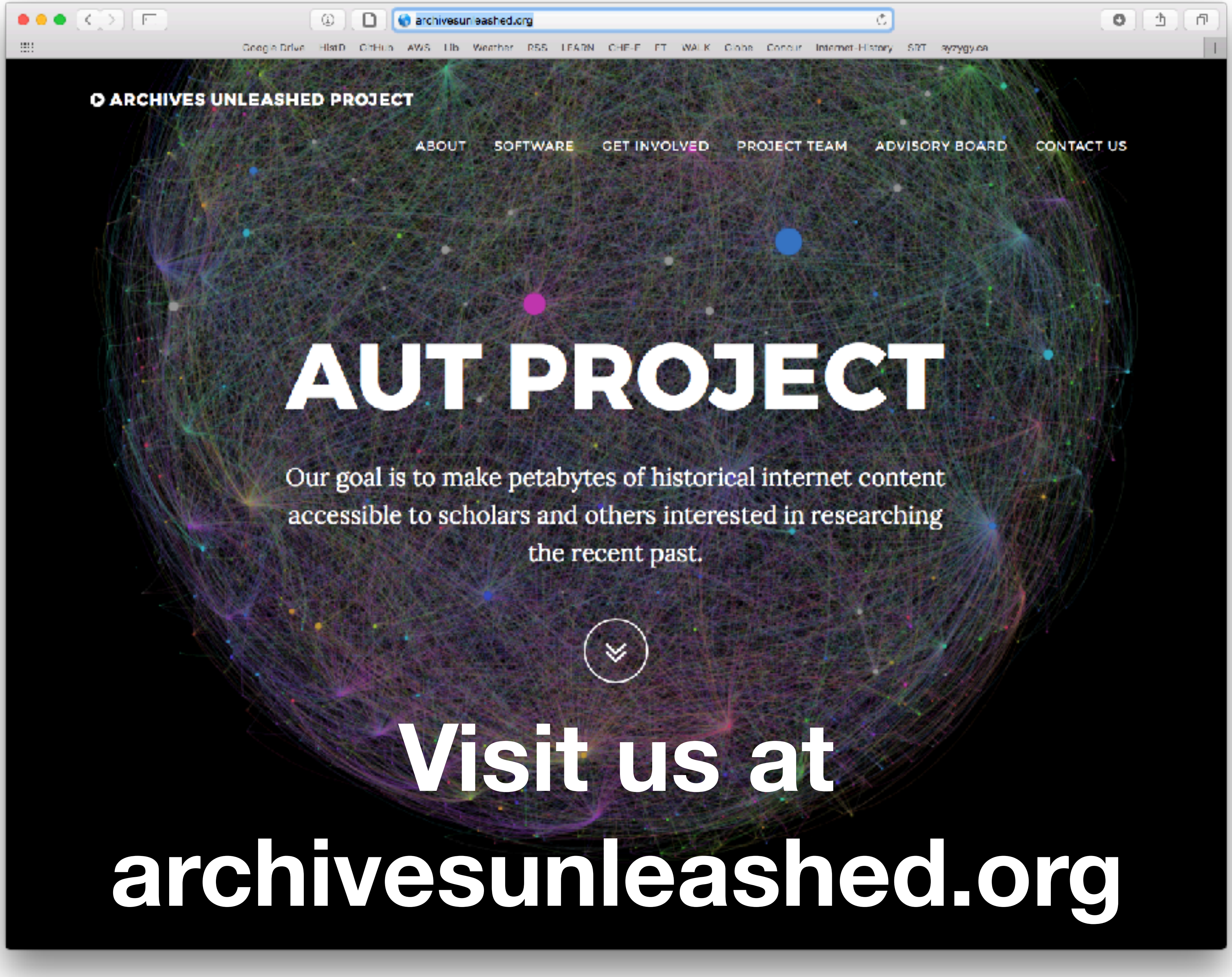


Scholars can directly use the Archives Unleashed Toolkit using local resources of via the Archives Unleased Cloud

Data ingestion from Archive-It, scholar's local collections, etc.

Scholars can download results, derived data, etc. for additional local analyses if desired

Altiscale
Archives Unleashed Cloud Portal
Archives Unleashed Toolkit
Spark Cluster
Archives Unleashed Toolkit
Local Spark Cluster (can be single-node)
Search portal (Project Blacklight)
Compute Canada



ARCHIVES UNLEASHED PROJECT

ABOUT SOFTWARE GET INVOLVED PROJECT TEAM ADVISORY BOARD CONTACT US

AUT PROJECT

Our goal is to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past.

Visit us at archivesunleashed.org

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada

Canada

Ontario

compute canada | calcul canada

UNIVERSITY OF WATERLOO

YORK UNIVERSITÉ UNIVERSITY

THE ANDREW W. MELLON FOUNDATION

Thanks very much!

Ian Milligan
Associate Professor
@ianmilligan1

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History