

# WebArchives.ca

## Enabling Public Access to Web Archives

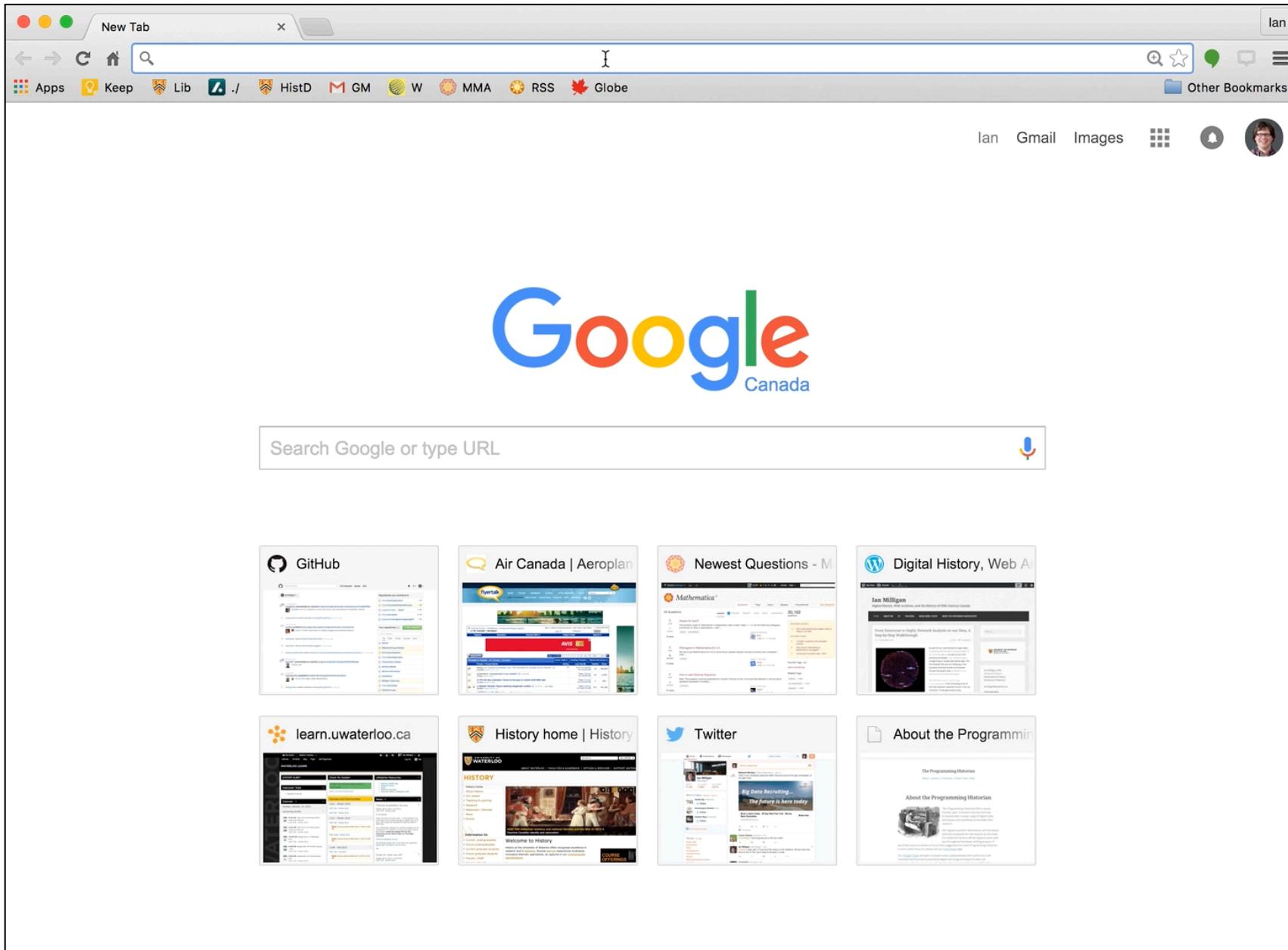
---

**Ian Milligan**  
Assistant Professor, History

UNIVERSITY OF  
**WATERLOO**



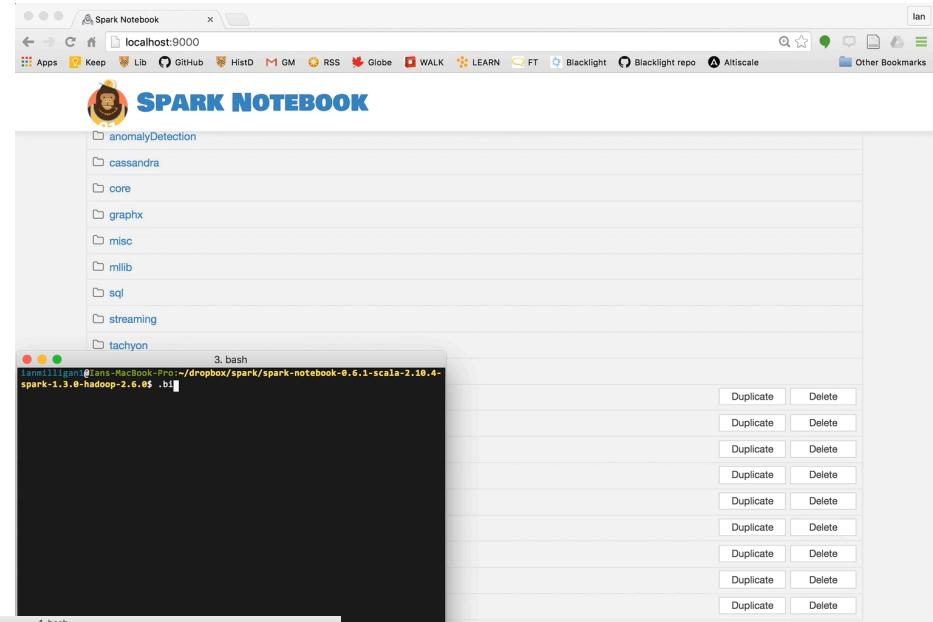
**Web Archives are great**  
– **but how do you use  
them?**



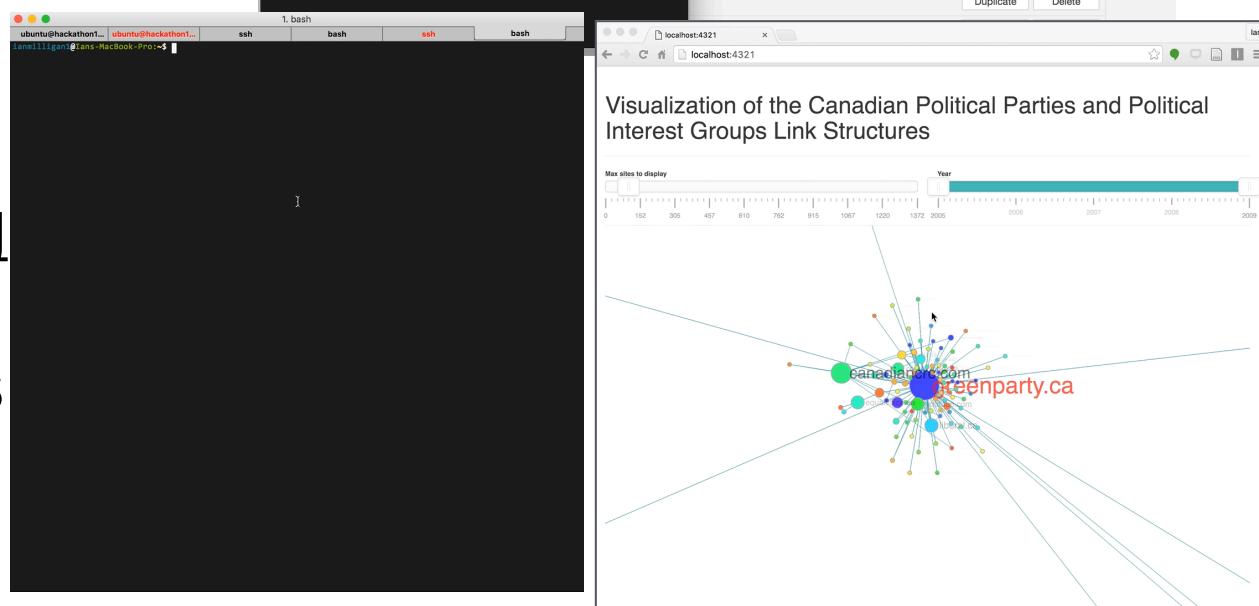
**Limited to close, slow reading.**

# Two Models

- **The coding, computational approach discussed in the last session (warcbase)**



- warcbase.org
- docs.warcbase.org
- If you have the files storage, etc.



**Lighter-weight interfaces  
needed for public,  
occasional research use.**

BRITISH  
LIBRARY

YORK  
UNIVERSITÉ  
UNIVERSITY

UNIVERSITY OF  
WATERLOO



Andrew Jackson, Jimmy Lin, Ian Milligan, and Nick Ruest, “**Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities**,” *Joint Conference on Digital Libraries 2016*, Newark NJ.

**In short: we will need  
search engines.**



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

# Standard SERP Inadequate

The screenshot shows a web browser window with the URL <https://archive-it.org/collections/227?q=%22Stephen+Harper%22&page=1&show=Sites>. The page title is "Canadian Political Parties and Political Interest Groups" collected by "University of Toronto". The page is archived since Oct, 2005.

On the left, there is a sidebar with search filters:

- Contains all of:
- Exact phrase:
- Not containing:
- From the Host:  ex. [www.archive-it.org](http://www.archive-it.org)
- Results per host:  1 (default)
- File format:  All formats
- Capture date range:  
From:   
To:

The main content area shows a search bar with "Stephen Harper" and a "Search" button. Below it, a message states: "The following results were found for the term(s): "Stephen Harper"" followed by a bullet point: "No metadata results for "Stephen Harper", but there are up to 1211638 matches within the page text."

At the bottom, there is a summary of the first result:

Page 1 of 60,582 (1,211,638 Total Results) [Next Page ▶](#)

Sort By: [Best Match](#)

**Stephen Harper | Facebook**  
URL: <http://www.facebook.com/pages/Stephen-Harper/9106562109>  
This text was captured on **May 02, 2009** [Show All Captures](#)

Stephen Harper | Facebook Remember Me Forgot your password? Sign Up Stephen Harper is on Facebook Sign up for Facebook to connect with Stephen Harper. Information Country: Canada Currently... Stephen Harper | Showing 10 photos Most Recent | Edit Pictures YouTube Box 10 of 13 See all PM on Wolf... the Prime Minister 11:28am Dec 22 | 30 Comments Create a Page Report Page Stephen Harper Wall Info Boxes Notes Stephen Harper + Fans Just Stephen Harper Just Fans Stephen Harper Celebrating... Stephen Harper Launched the Apprenticeship Completion Grant. \$2000 to eligible apprentices. <http://tinyurl.com/cqyzyv> April 9 at 11:47am Stephen Harper 'Lest we forget.' Statement on the 92nd anniversary of the battle of Vimy Ridge. <http://bit.ly/ERb1l> April 9 at 11:25am Stephen Harper Announced new...  
Content: text/html Size: 108 KB  
[More Results from facebook.com](#)

**Overview first, zoom and  
filter, then details-on-demand.**

**Schneiderman's mantra (1996)**

# **Close - Medium - Distant**

- **Distant Reading**: Billions of documents
- **Close Reading**: One document
- **“Middle game”**: Moving between these levels

# Building Portals

- Democratizing access so that historians can use them.
- Building **transparent indexes**.
- But they have to be useful and tested...

The screenshot shows a web browser window with the title "Archive-It - Canadian Political Parties and Political Interest Groups". The URL in the address bar is <https://archive-it.org/collections/227>. The page features a navigation bar with links for HOME, EXPLORE, LEARN MORE, and CONTACT US. A large "ARCHIVE-IT" logo is visible. The main content area displays a collection titled "Canadian Political Parties Groups" collected by "University of Toronto". It includes information about being archived since Oct, 2005, and a description mentioning national Canadian political parties and a number of specific subjects like Politics & Elections. 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 for Title (A-Z), Title (Z-A), URL (A-Z), and URL (Z-A).

# Pivotal Changes in Canadian Politics, 2005-2015

- Militarization of Canadian society?
- Change from ‘natural governing party’ of Liberals to Conservatives
- Major policy changes on foreign policy, environment, etc.
- Sweeping change to how we understand our history
- How to measure?



# Canadian Political Parties & Political Interest Group Collection

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



# 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 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 header includes the Archive-It logo, navigation links for HOME, EXPLORE, LEARN MORE, and CONTACT US, and a tagline: "The leading web archiving service for collecting and accessing cultural heritage on the web. Built at the Internet Archive". Below the header, the breadcrumb navigation shows "Explore > University of Toronto > Canadian Political Parties and Political Interest Groups". The main content area features the University of Toronto Libraries logo and the title "Canadian Political Parties and Political Interest Groups". It indicates the collection was "Collected by: University of Toronto", "Archived since: Oct, 2005", and describes it as "Canadian Political Parties and Political Interest Groups will archive the websites of all the national Canadian political parties, and a number of special interest groups across the political spectrum". The collector is listed as "Collector: University of Toronto". A search bar at the bottom left contains the query "Stephen Harper". The results summary states "Page 1 of 60,657 (1,213,132 Total Results)" and "Sort By: Best Match". The results list includes a link to "Stephen Harper | Facebook" with a URL of <http://www.facebook.com/pages/Stephen-Harper/9106562109>. The page also includes a "Search Page Text" input field and a "Next Page ▶" button.

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](#)

Great research question that our  
contemporary historians were  
studying (Canada changing)

+

Great collection (all the political  
parties + many interest groups)

+

Open Source Software





With Nick Ruest (York University), Nich Worby (University of Toronto), Jimmy Lin (University of Waterloo)

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

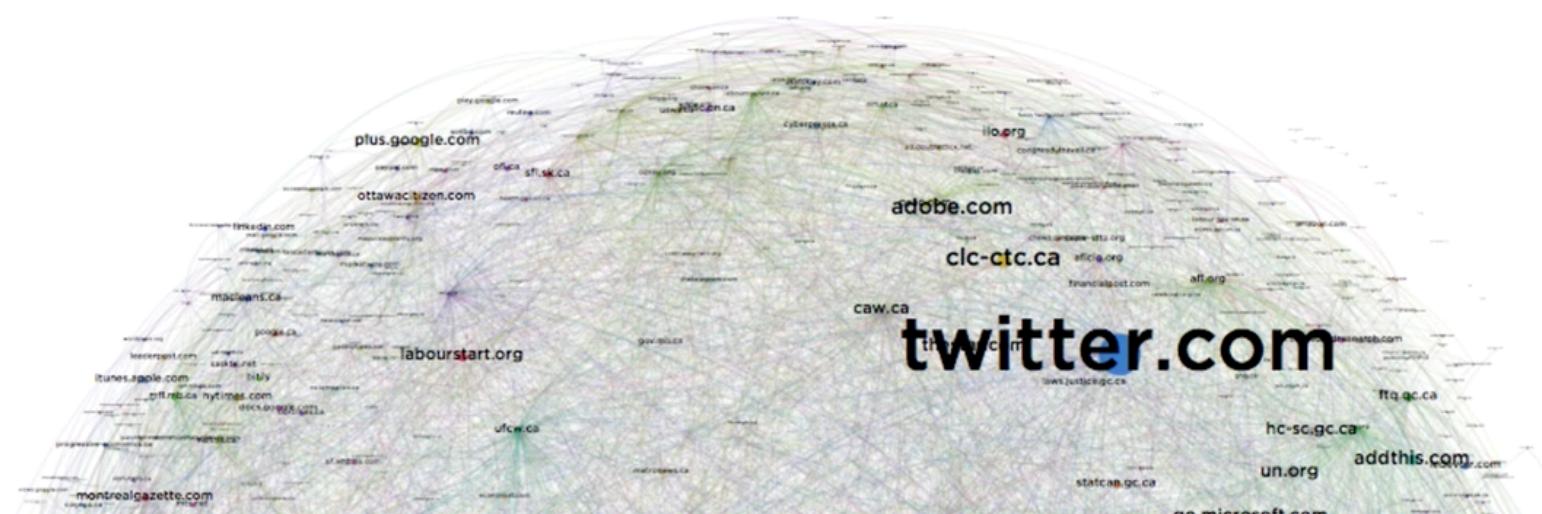
**On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.**

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.

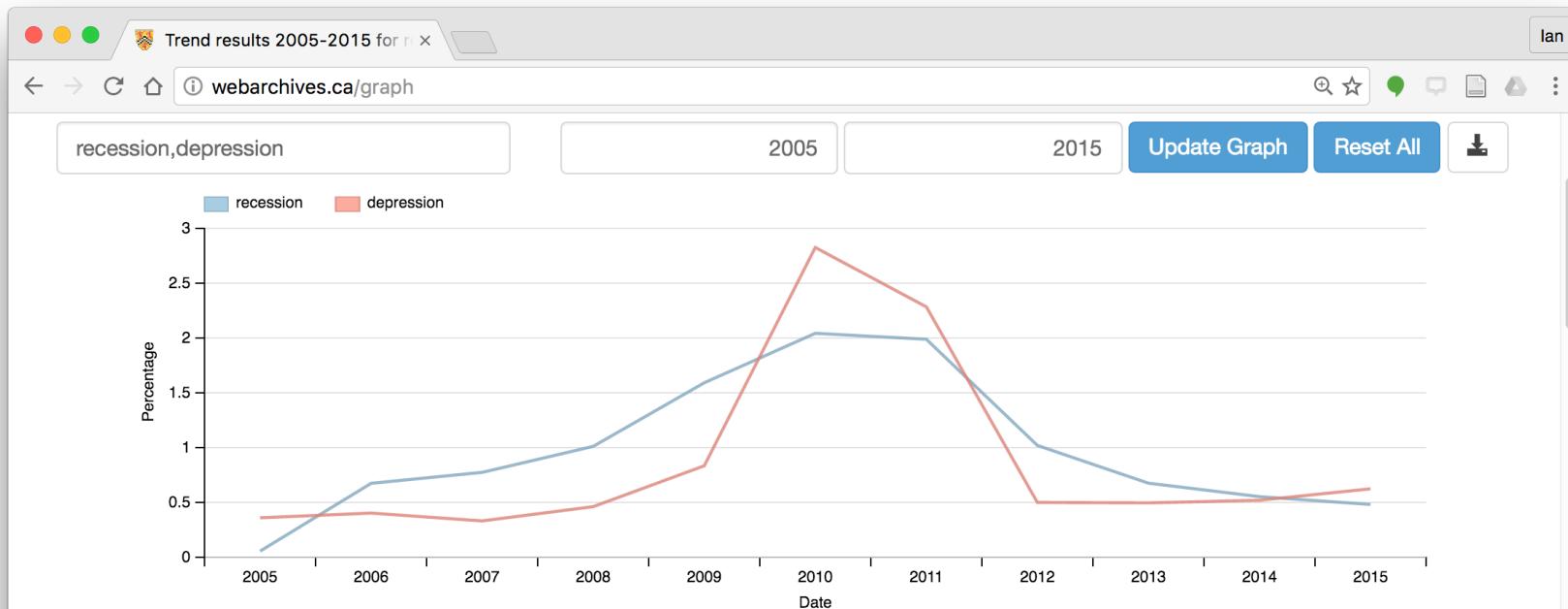
Options include:

- **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



# Trend Diagram



Distant =

Middle Game

Close =

## Some Sample Queries to Get Started With:

- "[Public Transit](#)": Liberals/NDP/Conservatives. We use quotation marks to search on the exact phrase.
- "[Tar Sands](#)": Amongst the three major parties, who uses and subsequently stops using this phrase for Canada's Oil Sands?
- "[Climate Change](#)": Similar to the above two queries, we can see this issue rise and possibly wane in public interest.

Feel free to take these examples and adapt them to your own purposes. If you find something interesting, [please let us know!](#)

Found 100 samples matching '[recession](#)' from 2010.

Matching Text	Link
<a href="#">recession</a> , and if persisted in, a depression. This method hits eve... <a href="#">canadianactionparty.ca</a>	
. eeds forty percent of income and is rising. Today, with a <a href="#">recession</a>   starting	
prevent a <a href="#">recession</a> at home rather than support the value of the dollar abr...	
...mblay=1064%5D" target="_blank">A Slowdown or a//: <a href="#">Recession</a> in the U.S. in 2008?</a>), we are	
...lemented," said James.“For families struggling with the <a href="#">recession</a> a	<a href="#">ndp.ca</a>
...the financial burden we must live in today in this time of <a href="#">recession</a> . I haven't	<a href="#">ndp.ca</a>

# Five Things We Learned

- Political parties delete content
- User-generated comments were more common in political parties
- Absences can be more informative than presences
- We can see the rise/fall of prominent people
- Enabling user access is truly transformative

# Next Steps

The screenshot shows a web browser window with the title bar "Web Archives for Longitudinal Knowledge (WALK) | webarchives.ca". The address bar contains "webarchives.ca". The page itself has a blue header with the WALK logo and navigation links for Search, Trends, About, and Datasets. A yellow callout box in the center of the page says "Welcome to the Web Archives for Longitudinal Knowledge (WALK) portal. Before diving in, we encourage you to visit our [about](#) page." Below this, the main content area features a large heading "Web Archives for Longitudinal Knowledge (WALK) Portal". A paragraph explains the project is a Canadian national Web Archiving portal involving the University of Waterloo, York University, and the University of Alberta. It notes it's a prototype site for political parties and interest groups from 2005 to 2015. Another paragraph discusses the Liberal Party's response to the 2008 crisis and Michael Ignatieff's reaction. A section titled "Options include:" lists three items: "Basic keyword searching [Example: 'Rob Ford', only Liberal.ca]", "Graphing trends over time [Example: Liberal Opposition Leaders, 2005-2015]", and "Advanced search, including words in proximity to each other [Example: environmental and tax within 25 words of each other]". At the bottom, a note says "Below, here are all of the links for the entire time period, visualized below."

## Web Archives for Longitudinal Knowledge (WALK) Portal

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

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

Curious how the Liberal Party of Canada responded to the 2008 financial crisis ([a search for "recession" in 2008, liberal.ca](#))? How the Canadian Centre for Policy Alternatives [reacted to Michael Ignatieff](#)? Now you can check it all out.

Options include:

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

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



# Thanks!

compute | calcul  
canada | canada



Social Sciences and Humanities  
Research Council of Canada

Conseil de recherches en  
sciences humaines du Canada

Canada

**Ian Milligan**  
Assistant Professor, History

UNIVERSITY OF  
**WATERLOO**

