

# **Breaking down the Silos: The Web Archives for Longitudinal Knowledge Project**

---

**Ian Milligan**  
Assistant Professor  
@ianmilligan1



**UNIVERSITY OF WATERLOO**  
**FACULTY OF ARTS**  
Department of History

You are viewing an archived web page, collected at the request of [University of Toronto](#) using [Archive-It](#). This page was captured on 20:44:25 Mar 24, 2006, 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 The information on this web page may be out of date.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.

**Welcome**

**Français**

**Become a member of Equal Voice: [Click here to signup!](#)**

**Getting To The Gate**

This online course aims to increase the number of elected women by providing practical tools for women of all ages, backgrounds and walks of life interested in running for public office

**Spotlight**

**Equal Voice Chair Rosemary Spears Honoured as YWCA Woman of Distinction for 2006.**

Mar 8, 2006.

**Women in Politics by Geoffrey Stevens.**

Feb 28, 2006.

**Harper's Cabinet (Letter to the editor of The Hill Times)**

Feb 28, 2006.

**Equal Voice: comments on the new Cabinet**

Web site hosting provided courtesy of Xynapse Inc.

[Privacy Policy](#)

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 The information on this web page may be out of date.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.

**THE TEAM THE PARTY ISSUES MEDIA CENTRE YOUR RIDING DONATE**

**Stay Informed**

**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**

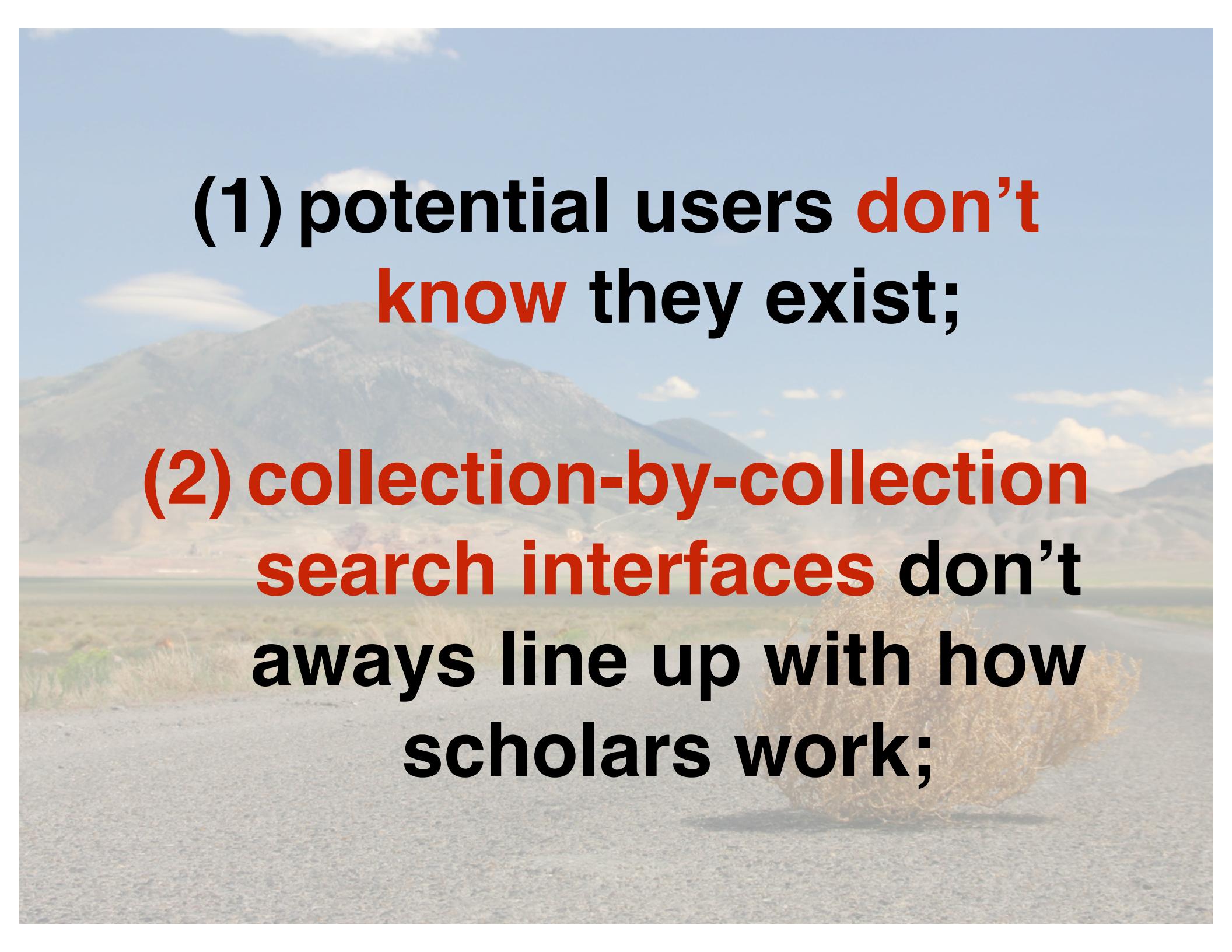
**Aboriginal Peoples' Commission**

**Senior Liberals Commission**

# We have **fantastic** web archival collections in Canada.

**But not many people  
use them...**



- 
- The background of the slide features a scenic landscape with a large, rugged mountain range under a clear blue sky with a few wispy clouds. In the foreground, there's a dirt road or path leading towards the mountains, with some dry, yellowish-brown brush on the right side.
- (1) potential users **don't** know they exist;**
  - (2) collection-by-collection search interfaces **don't** always line up with how scholars work;**

https://archive-it.org/collection/x

Secure https://archive-it.org/collections/227?q="Stephen+Harper"&page=1&show=Sites

lan

Enter a search term on the right to search the text within the archived pages. Or for more search options, use the Advanced Search options below.

Advanced Search

Contains all of:

Exact phrase:

Not containing:

From the Host:

ex. www.archive-it.org

Results per host:

1 (default) ▾

File format:

All formats ▾

Capture date range:

From: ▾ ▾

To: ▾ ▾

[Advanced Search](#)

[Help with Search](#)

specific URL or to search the text of archived webpages.

"Stephen Harper"

Search Clear

The following results were found for the term(s): "Stephen Harper"

- No metadata results for "Stephen Harper", but there are up to 1229211 matches within the page text.

Search Page Text

Page 1 of 61,461 (1,229,211 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/cqyzv> 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](#)

---

**Stephen Harper (pmharper) on Twitter**

URL: <http://twitter.com/PMHarper>

This text was captured on **Aug 03, 2010** [Show All Captures](#)

Stephen Harper (pmharper) on Twitter Skip past navigation On a mobile phone? Check out <m.twitter.com>! Skip to navigation Skip to sign in form Have an account? Sign in Username or email Password Remember me Forgot password? Forgot username? Already using Twitter on your phone? Get short, timely messages from Stephen Harper. Twitter is a rich source of instantly updated information. It's easy to stay updated on an incredibly wide variety of topics. Join today and follow @pmharper. Get updates via SMS by texting follow pmharper to 40404 in the United States Codes for other countries Two-way (sending and receiving) short codes: Country Code For customers of Australia 0198089488 Telstra Canada... Account Name Stephen Harper Location Ottawa, Ontario Web <http://www.conser...> Bio Prime Minister of...

Content: text/html Size: 46 KB

[More Results from twitter.com](#)

---

**The Walrus » The Man Behind Stephen Harper » Tom Flanagan » politics**

URL: <http://www.walrusmagazine.com/articles/the-man-behind-stephen-harper-tom-flanagan/politics>

This text was captured on **Aug 03, 2010** [Show All Captures](#)

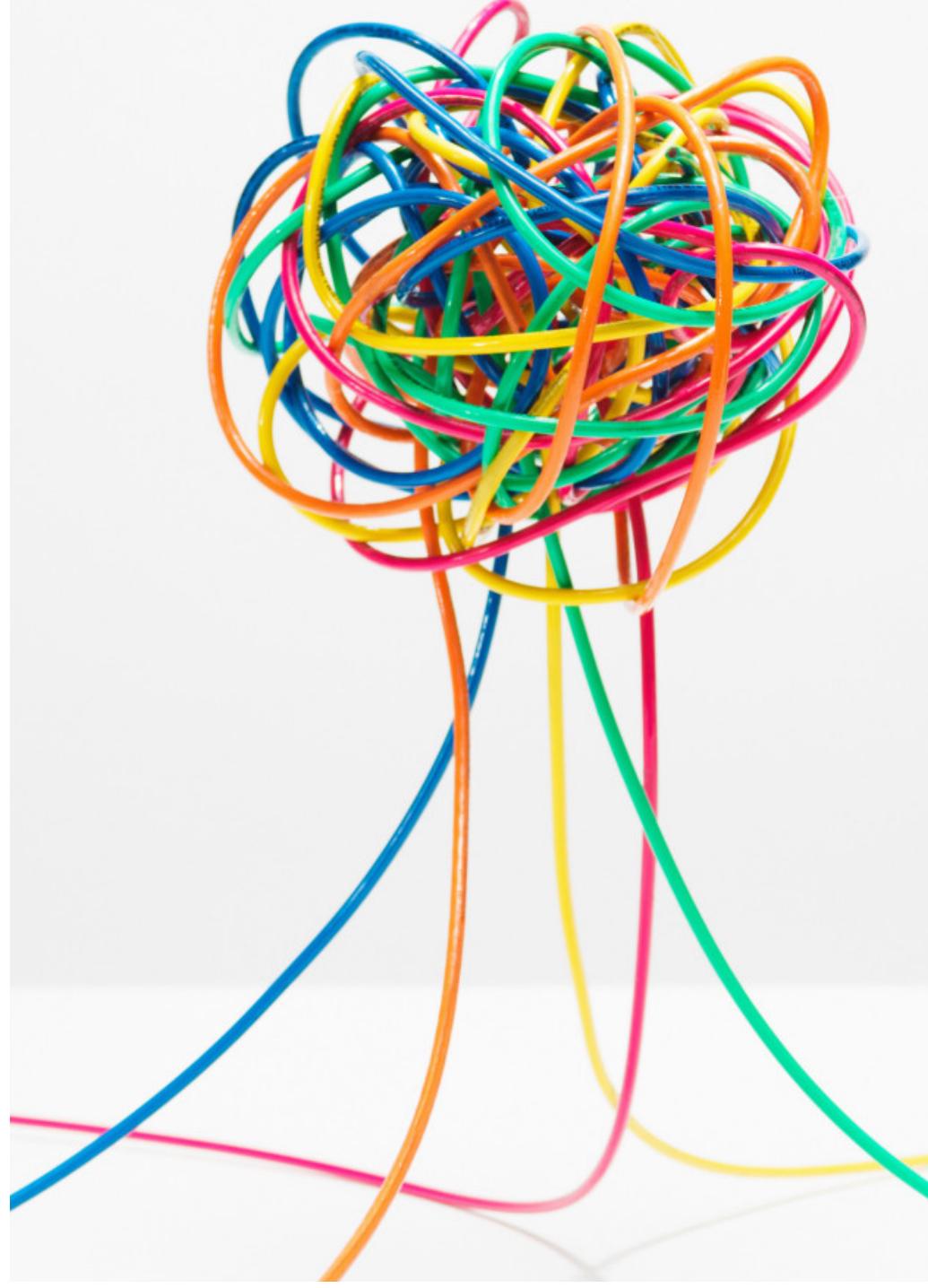
The Walrus » The Man Behind Stephen Harper » Tom Flanagan » politics Subscribe online for \$2.98

**Canadian  
archival data is  
silo'd by  
institution and  
collection.**

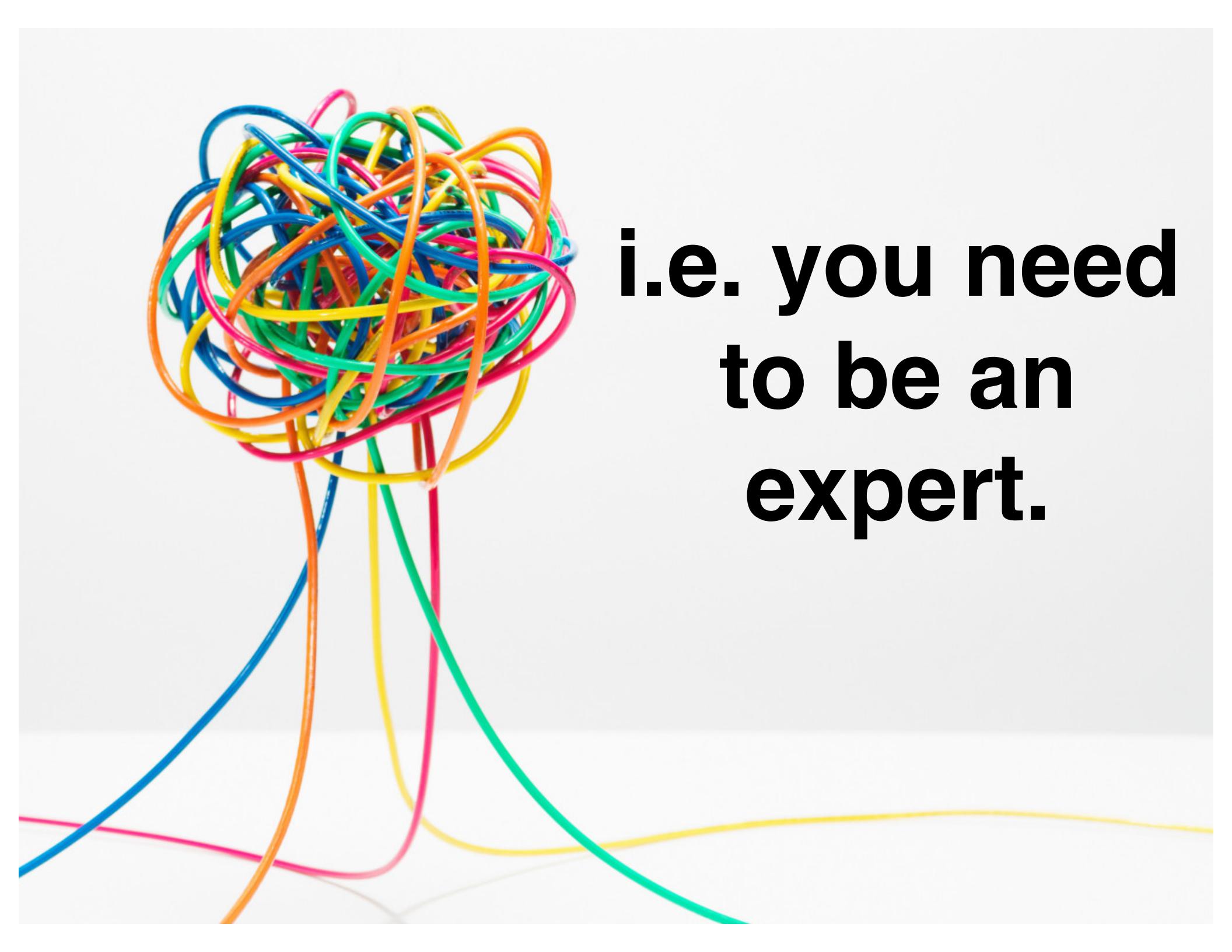


# The Canadian Web Archival Landscape

- Currently around 25 Canadian collecting institutions, amassing around ~ 130 collections
- Most use Archive-It as a back-end provider of web archival services



**Right now, to  
use Canadian  
web archives –  
you have to  
really want to  
use them.**



i.e. you need  
to be an  
expert.

# I want web archives to be used on page 150 of a random book.

These diverse items had never before been aggregated under one heading. The telegraph gave them their commonality. In patent applications and legal agreements, too, the inventors had reason to think about their topic in the broadest possible terms: e.g., the giving, printing, stamping, or otherwise transmitting of signals, or the sounding of alarms, or the communication of intelligence. To understand the telegraph, it is necessary to understand the technology itself. Confusion is inevitable, which often turned on awkward new meanings of familiar terms: innocent words like *message* and *operator*, vily laden ones, like *message*. There was the man who brought a *message* to the telegraph office in Baden-Baden. The operator manipulated the telegraph key and then placed the paper on the hook. The customer complained that the message had not been sent, because he could still see it hanging on the hook. To *Harper's New Monthly Magazine*, which recounted this story in 1873, the point was that even the "intelligent and well-informed" continued to find these matters inscrutable:

# Enter the Web Archives for Longitudinal Knowledge Project

lan

Web Archives for Longitudinal Knowledge (WALK)

Welcome to the Web Archives for Longitudinal Knowledge (WALK) portal. Before diving in, we encourage you to visit our [about](#) page.

## Web Archives for Longitudinal Knowledge (WALK) Portal

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

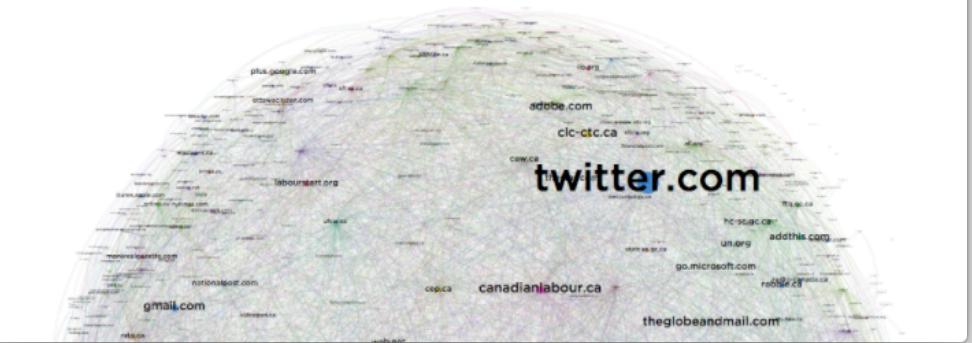
Currently, this is a prototype site providing access to one such archive, the University of Toronto's Canadian Political Parties and Political Interest Groups collection. This website allows you to search content from 50 political parties and political interest groups, from October 2005 to March 2015.

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

Options include:

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

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



A large, light-colored concrete bridge girder is shown in the background, tilted at an angle. It appears to be part of a larger structure that has collapsed or is being demolished. In the foreground, there is a construction site with several green shipping containers. Two workers wearing hard hats and safety vests are standing near a row of concrete blocks. A white vehicle is partially visible on the right side of the frame. The sky is clear and blue.

**We want to break down  
silos, bring Canadian  
web archives into a  
centralized portal with  
access to derivative  
datasets.**

# WALKin'

- **Institutions with signed MOUs:**  
Toronto, Alberta, Victoria,  
Winnipeg, Dalhousie, Simon  
Fraser University
- 61 collections (~ half of Canadian  
web archival collections)
- 16 TB of WARC files
- Developing new Solr front end  
based on **Project Blacklight**.  
Data stored on Compute  
Canada/Dataverse (currently  
indexed 120 million records).

The screenshot shows a web browser window titled "Blacklight" at "localhost:3000". The main content area displays a search interface for "Web Archives For Longitudinal Knowledge" (WALK). On the left, there's a sidebar with a search bar and dropdown menus for "All Fields", "General Content Type" (audio, html, image, other, pdf, powerpoint, text, video, word), "Domain", "Links Domains", "Institution", "Collection Name", and "Collection Number". The right side features a large text block about the project's mission to use Big Data to reshape historical knowledge, followed by a circular network graph of domain connections. Below this are sections for "Our partners" featuring logos and names of partner institutions: UNIVERSITY OF ALBERTA LIBRARIES, DALHOUSIE UNIVERSITY, and SFU SIMON FRASER UNIVERSITY ENGAGING THE WORLD.

# WALKin'

- Project Blacklight, bringing us into a broader community
- Better APIs, bug fixes, and ability to interact with, give back, communicate with a larger GLAM community (many of you!)

The top screenshot shows the GitHub repository page for `projectblacklight/blacklight`. It displays basic repository statistics: 3,349 commits, 27 branches, and 15 pull requests. The bottom screenshot shows the 'Active' tab of the branches page, listing several branches with their latest commits and statuses.

Branch	Commit	Status
<code>search_state_routes</code>	Updated 7 days ago by <code>jcoyne</code>	<code>#1646</code> Open
<code>search-state-clone</code>	Updated 7 days ago by <code>jcoyne</code>	<code>#1645</code> Open
<code>release-6.x</code>	Updated 11 days ago by <code>cbeer</code>	<code>#1620</code> Compare
<code>fix_style</code>	Updated 15 days ago by <code>jcoyne</code>	<code>#1634</code> Open
<code>1620-modal-order</code>	Updated a month ago by <code>cbeer</code>	<code>#1630</code> Merged
<code>configuration_for_type</code>	Updated 9 months ago by <code>jcoyne</code>	<code>#1417</code> Open
<code>action-view-logging</code>	Updated 2 months ago by <code>tampakis</code>	<code>#1599</code> Closed

# WALKin'

- For every collection, we use warcbase to generate derivatives like:
  - Domain counts
  - URL counts
  - Full text
  - Domain-To-Domain links
- Upload selected datasets to Dataverse

The screenshot shows a web browser window titled "Network Data for the Web Arch". The address bar indicates a secure connection to <https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=hdl:10>. The page header includes a "Secure" lock icon, the URL, and various navigation and search links. The main content area is titled "Scholars Portal Dataverse" and features a search bar. Below the title, it says "52 Files". A list of files is displayed in a table format:

<input type="checkbox"/>		<b>ALBERTA_edmonton_public_library.gdf</b> Unknown - 1.4 KB - Dec 14, 2016 - 1 Download MD5: 152f5d2c653810d58f1d0c2a585e232e; GDF file for Edmonton Public Library collection.
<input type="checkbox"/>		<b>alberta_education_curriculum.gdf</b> Plain Text - 175.5 KB - Aug 22, 2016 - 4 Downloads MD5: 1ddce4f23cdcdaf93f28d20da6888897; GDF file for Alberta Education Curriculum
<input type="checkbox"/>		<b>alberta_floods_2013.gdf</b> Plain Text - 486.7 KB - Aug 22, 2016 - 0 Downloads MD5: 1629a6cb96d8e1e3b39cad0dde6acc89; GDF file for Alberta Flood Collection
<input type="checkbox"/>		<b>ALBERTA_fort_mcmurray_wildfire_2016.gdf</b> Unknown - 1.2 MB - Dec 14, 2016 - 0 Downloads MD5: 1ce91fe6daa6cdcf1b829e8475129d81; GDF file for Fort McMurray collection.
<input type="checkbox"/>		<b>ALBERTA_lfrancophonie_de_louest_canadien.gdf</b> Unknown - 1.9 MB - Dec 14, 2016 - 0 Downloads MD5: c146695693bac65d6dcf2c44d9349302;
<input type="checkbox"/>		<b>alberta_oil_sands.gdf</b> Plain Text - 64.9 KB - Aug 22, 2016 - 0 Downloads MD5: 576679d94b999fba522f1d2d6a9b40d; GDF file for Alberta Oil Sands Collection
<input type="checkbox"/>		<b>ALBERTA_ottawa_shooting_october_2014.gdf</b> Unknown - 29.3 KB - Dec 14, 2016 - 0 Downloads

# WALKin'

- So networks scholars can play with hyperlink graphs; people can see collection coverage; we can run correspondence analysis; digital humanists can do sophisticated text analysis.

The screenshot shows a web browser window titled "Network Data for the Web Arch". The address bar indicates a secure connection to <https://dataVERSE.scholarsportal.info/dataset.xhtml?persistentId=hdl:10>. The page header includes links for Apps, Gmail, Lib, GitHub, AWS, HistD, RSS, Globe, WALK, LEARN, and HELP. The main content area is titled "Scholars Portal DataVERSE" and features a search bar with the placeholder "Search all dataVERses...". Below the search bar, it says "52 Files". A list of files is displayed in a table format:

<input type="checkbox"/>		<b>ALBERTA_edmonton_public_library.gdf</b> Unknown - 1.4 KB - Dec 14, 2016 - 1 Download MD5: 152f5d2c653810d58f1d0c2a585e232e; GDF file for Edmonton Public Library collection.
<input type="checkbox"/>		<b>alberta_education_curriculum.gdf</b> Plain Text - 175.5 KB - Aug 22, 2016 - 4 Downloads MD5: 1ddce4f23cdcdaf93f28d20da6888897; GDF file for Alberta Education Curriculum
<input type="checkbox"/>		<b>alberta_floods_2013.gdf</b> Plain Text - 486.7 KB - Aug 22, 2016 - 0 Downloads MD5: 1629a6cb96d8e1e3b39cad0dde6acc89; GDF file for Alberta Flood Collection
<input type="checkbox"/>		<b>ALBERTA_fort_mcmurray_wildfire_2016.gdf</b> Unknown - 1.2 MB - Dec 14, 2016 - 0 Downloads MD5: 1ce91fe6daa6cd3f1b829e8475129d81; GDF file for Fort McMurray collection.
<input type="checkbox"/>		<b>ALBERTA_lfrancophonie_de_louest_canadien.gdf</b> Unknown - 1.9 MB - Dec 14, 2016 - 0 Downloads MD5: c146695693bac65d6dcf2c44d9349302;
<input type="checkbox"/>		<b>alberta_oil_sands.gdf</b> Plain Text - 64.9 KB - Aug 22, 2016 - 0 Downloads MD5: 576679d94bb999fba522f1d2d6a9b40d; GDF file for Alberta Oil Sands Collection
<input type="checkbox"/>		<b>ALBERTA_ottawa_shooting_october_2014.gdf</b> Unknown - 29.3 KB - Dec 14, 2016 - 0 Downloads

**The goal: One central hub for web archiving search, research, derivatives. So researchers can cite web archives on page 150 of a book without needing to be an expert.**



Social Sciences and Humanities  
Research Council of Canada

Conseil de recherches en  
sciences humaines du Canada

Canada



**compute** | **calcul**  
canada | canada



UNIVERSITY OF  
**WATERLOO**

# Thanks!

---

**Ian Milligan**  
Assistant Professor  
@ianmilligan1



**UNIVERSITY OF WATERLOO**  
**FACULTY OF ARTS**  
Department of History