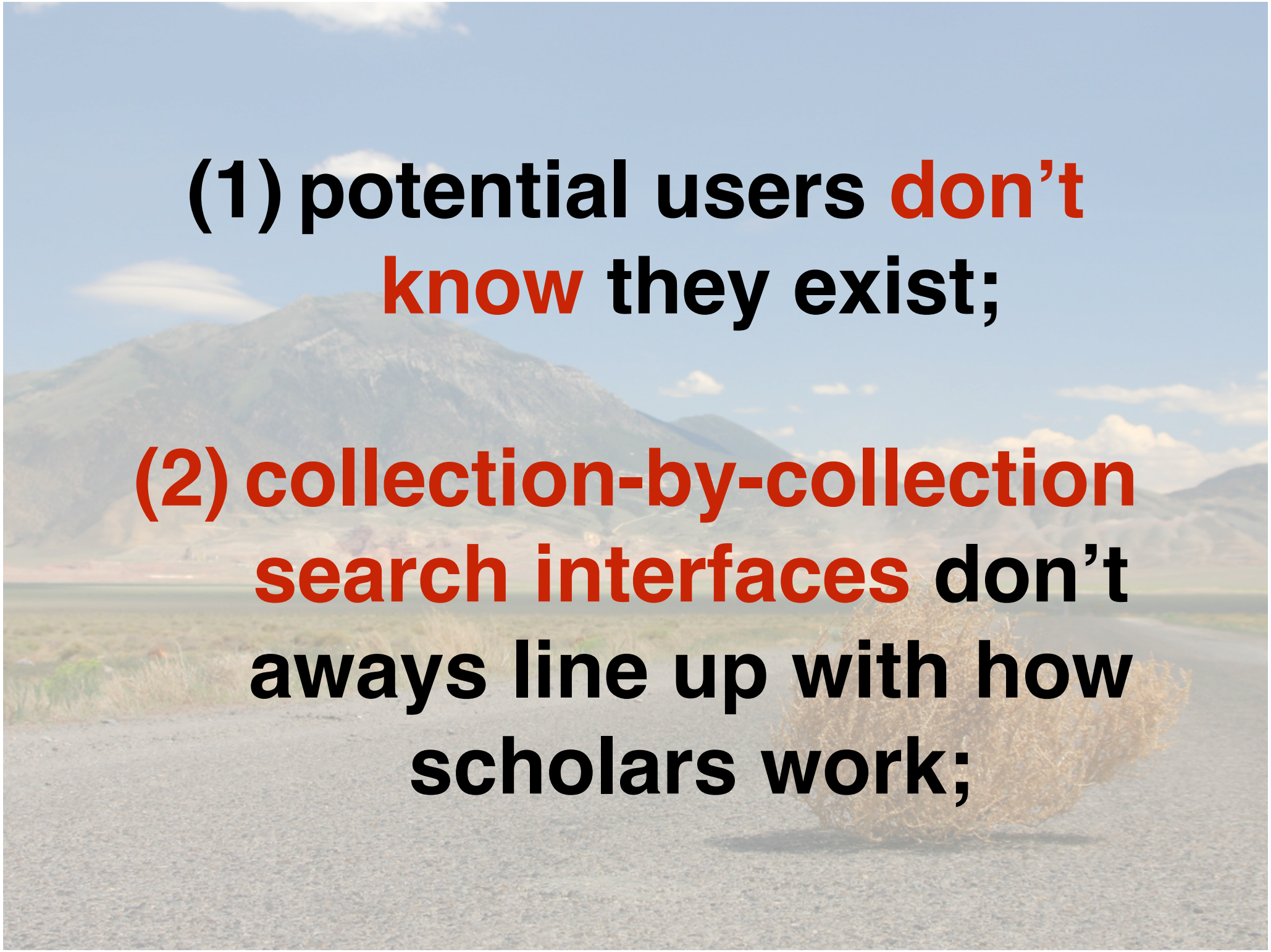
Breaking down the Silos:
The Web Archives for Longitudinal Knowledge Project

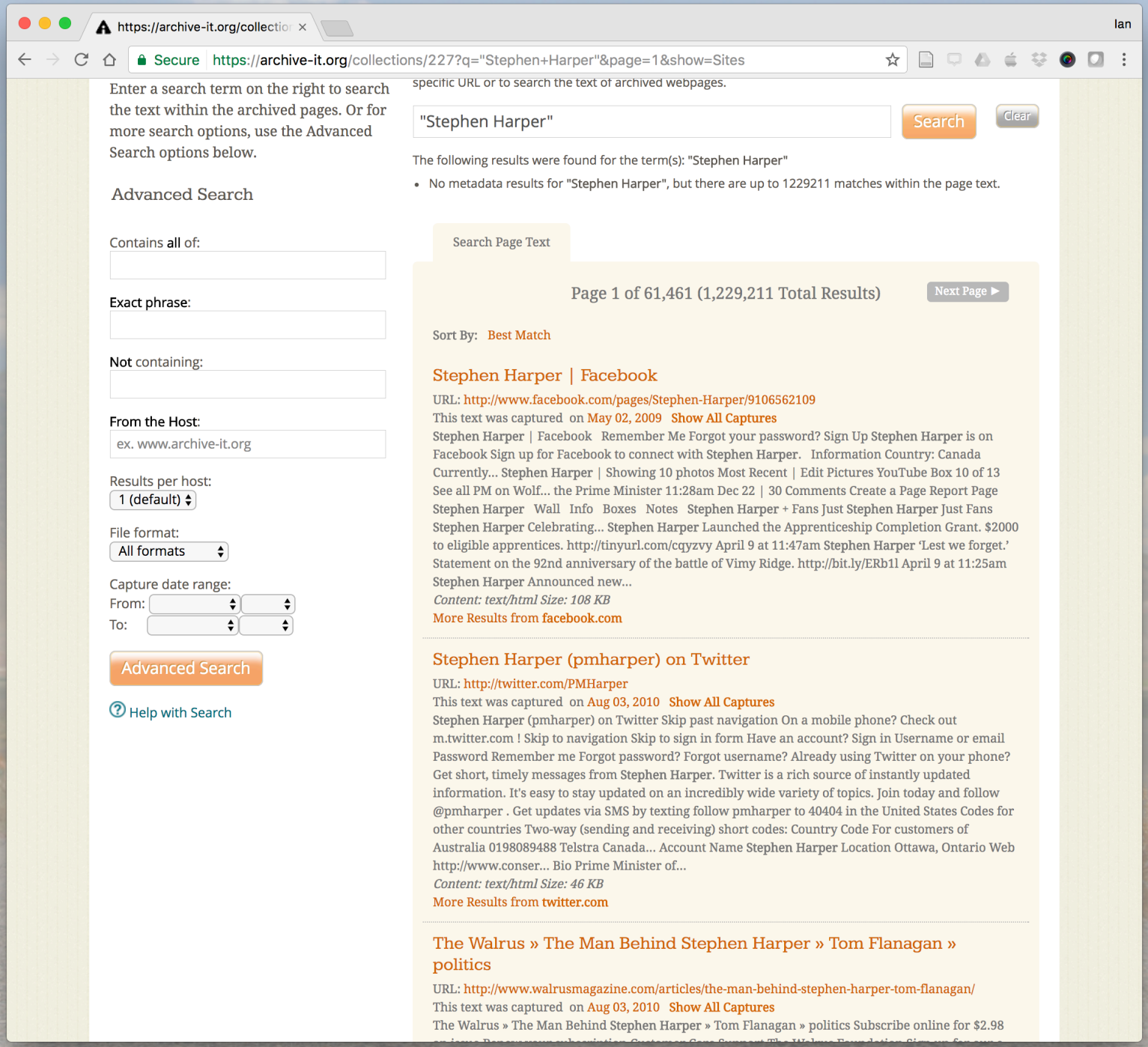
Ian Milligan
Assistant Professor
@ianmilligan1

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

We have fantastic web archival collections in Canada.


But not many people
use them…


(1) potential users don't know they exist;
(2) collection-by-collection search interfaces don't aways line up with how scholars work;


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

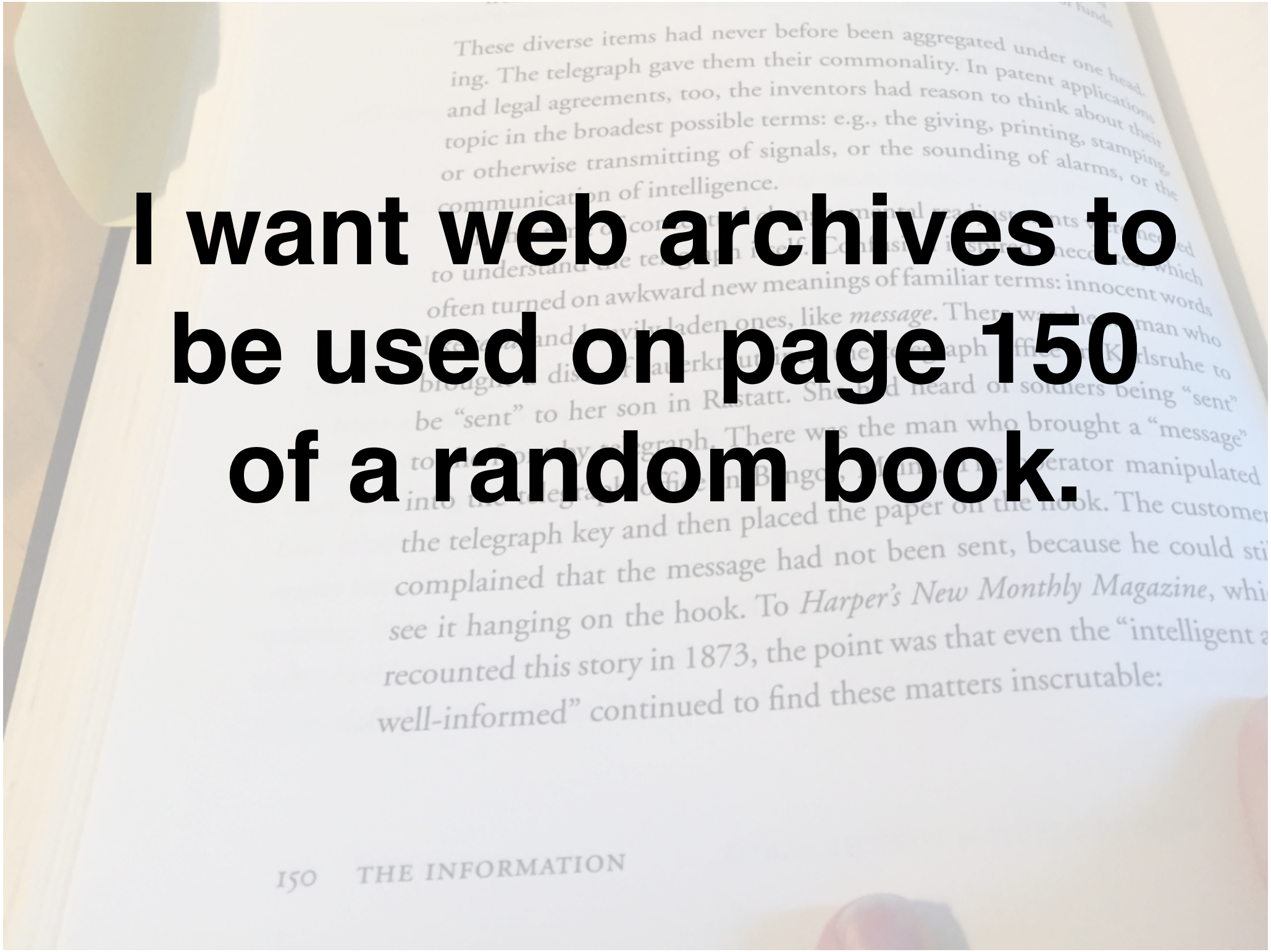

Canadian
archival data is
silo'd by
institution and
collection.

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**

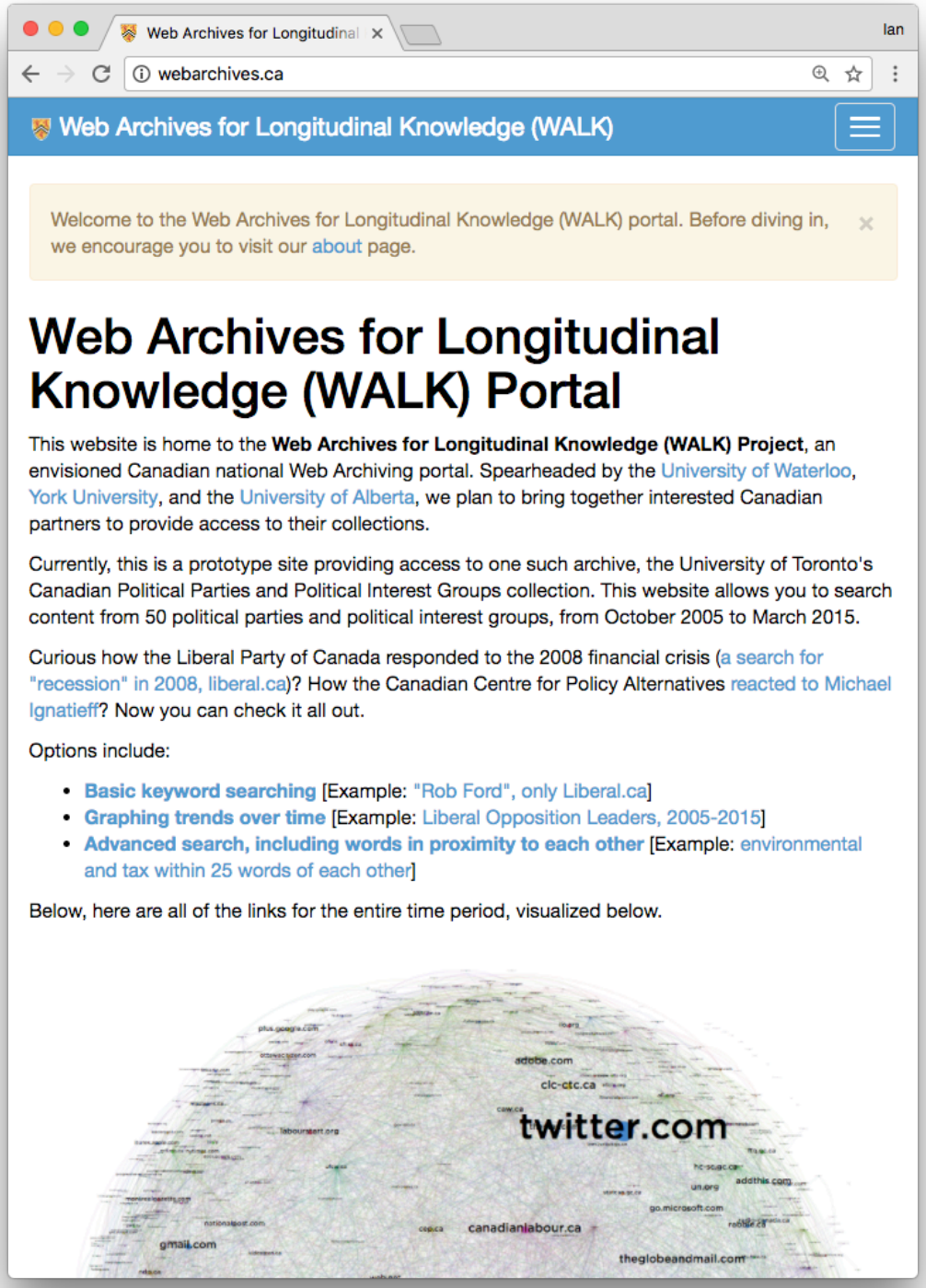

Right now, to use Canadian web archives – you have to really want to use them.


i.e. you need
to be an
expert.



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

Enter the Web Archives for Longitudinal Knowledge Project



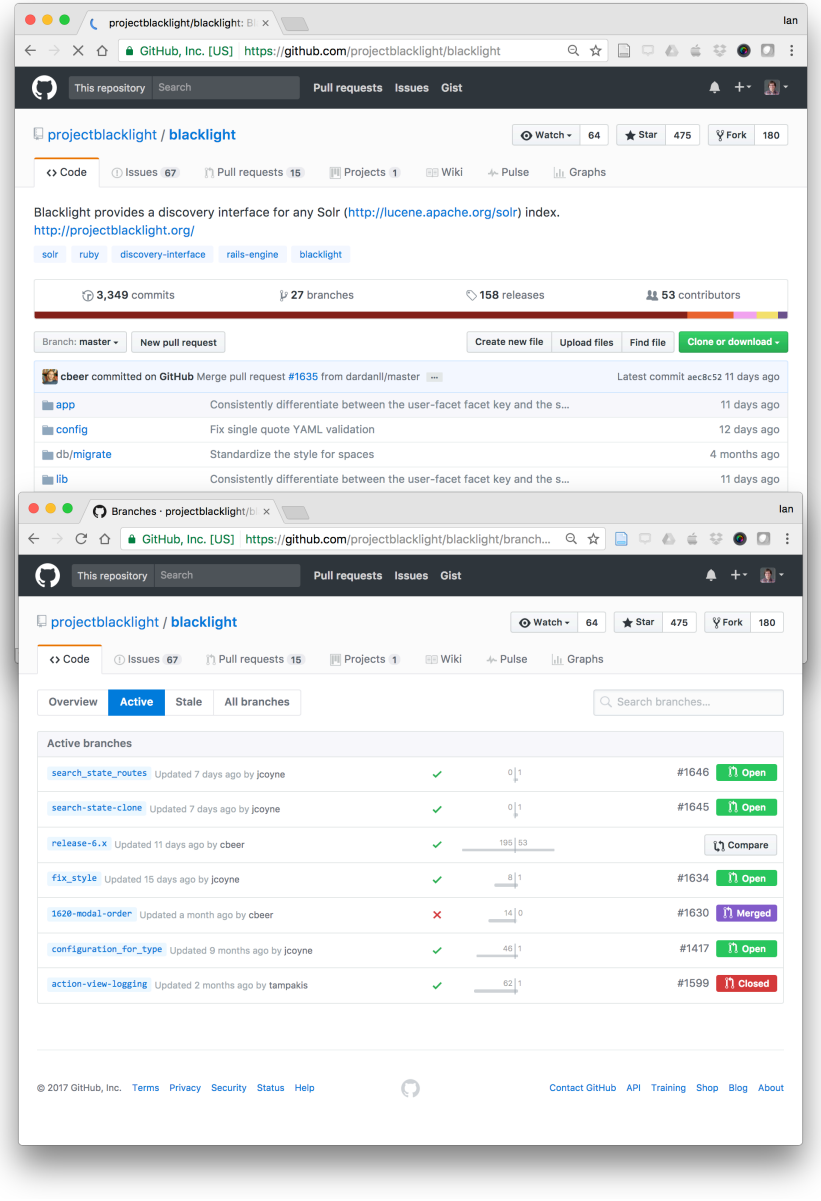

We want to break down silos, bring Canadian web archives into a centralized portal with access to derivative datasets.

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).

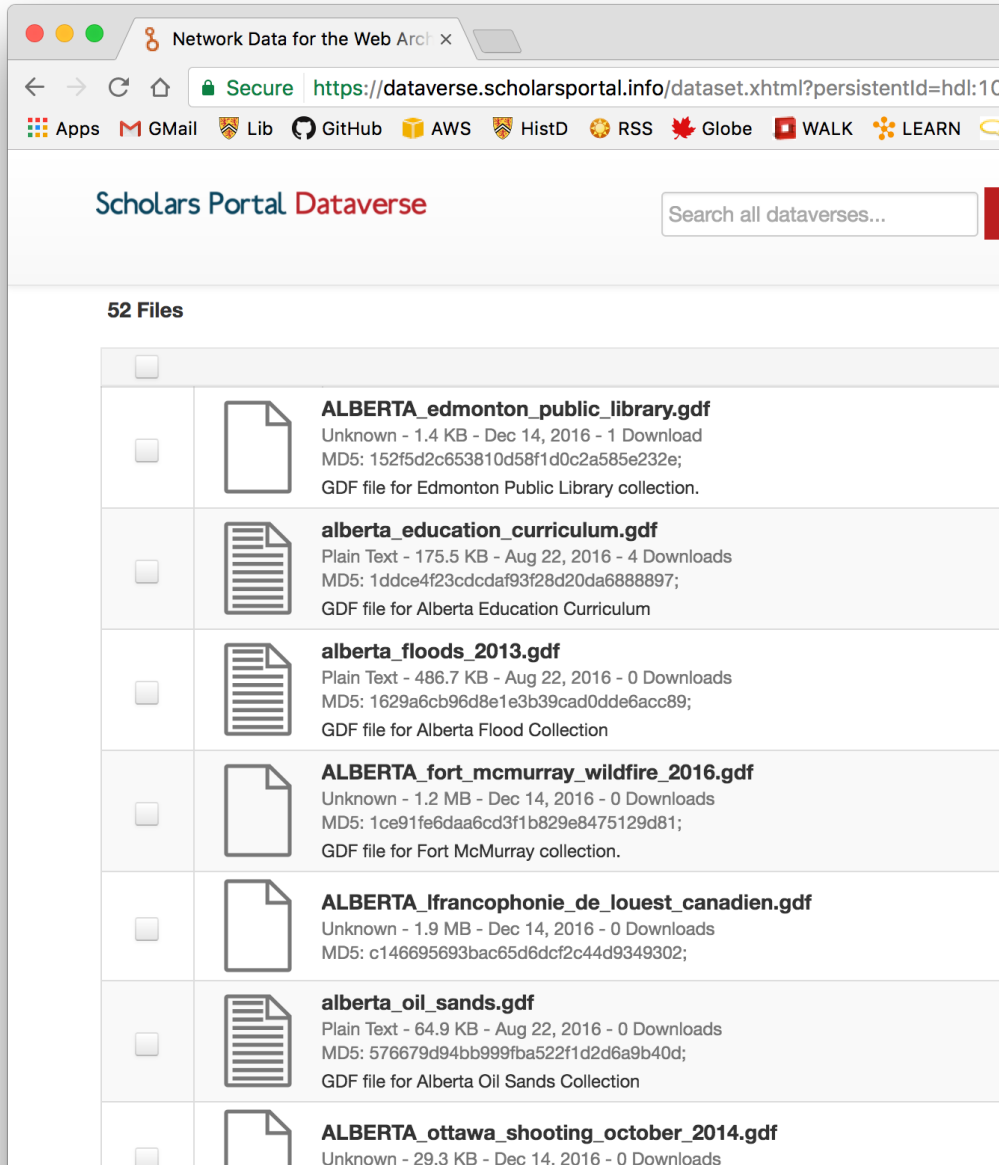
WALKin’

- 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!)



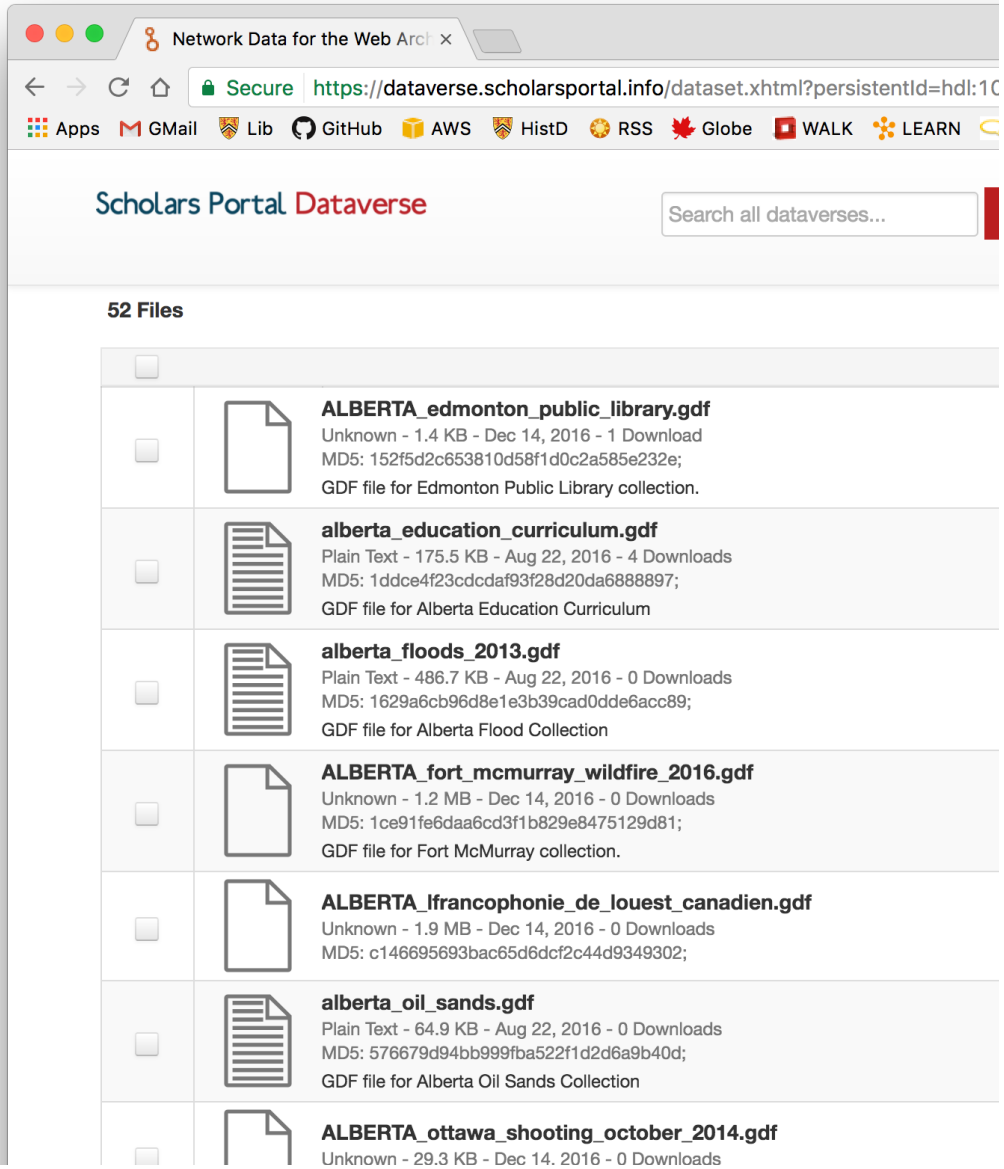
WALKin’

- 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



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.



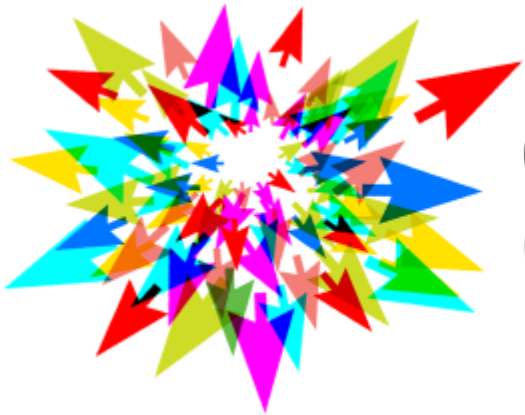
The goal: One central hub for web archiving search, research, derivatives. So researchers can cite web archives on page 150 of a book without needing to be an expert.

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Ontario



compute canada | calcul canada

UNIVERSITY OF WATERLOO

Thanks!

Ian Milligan

Assistant Professor

@ianmilligan1

UNIVERSITY OF WATERLOO

FACULTY OF ARTS

Department of History