

# **Studying the Web in the Shadow of Uncle Sam**

## **The Difficult Case of the Canadian Web Sphere**

**Workshop on National Webs**  
**Aarhus, Denmark**  
**8 December 2016**

---

**Ian Milligan**  
Assistant Professor  
@ianmilligan1



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

**Tom Smyth**  
Manager, Digital Operations  
@smythbound



Bibliothèque et Archives  
Canada

Library and Archives  
Canada

Arctic Ocean

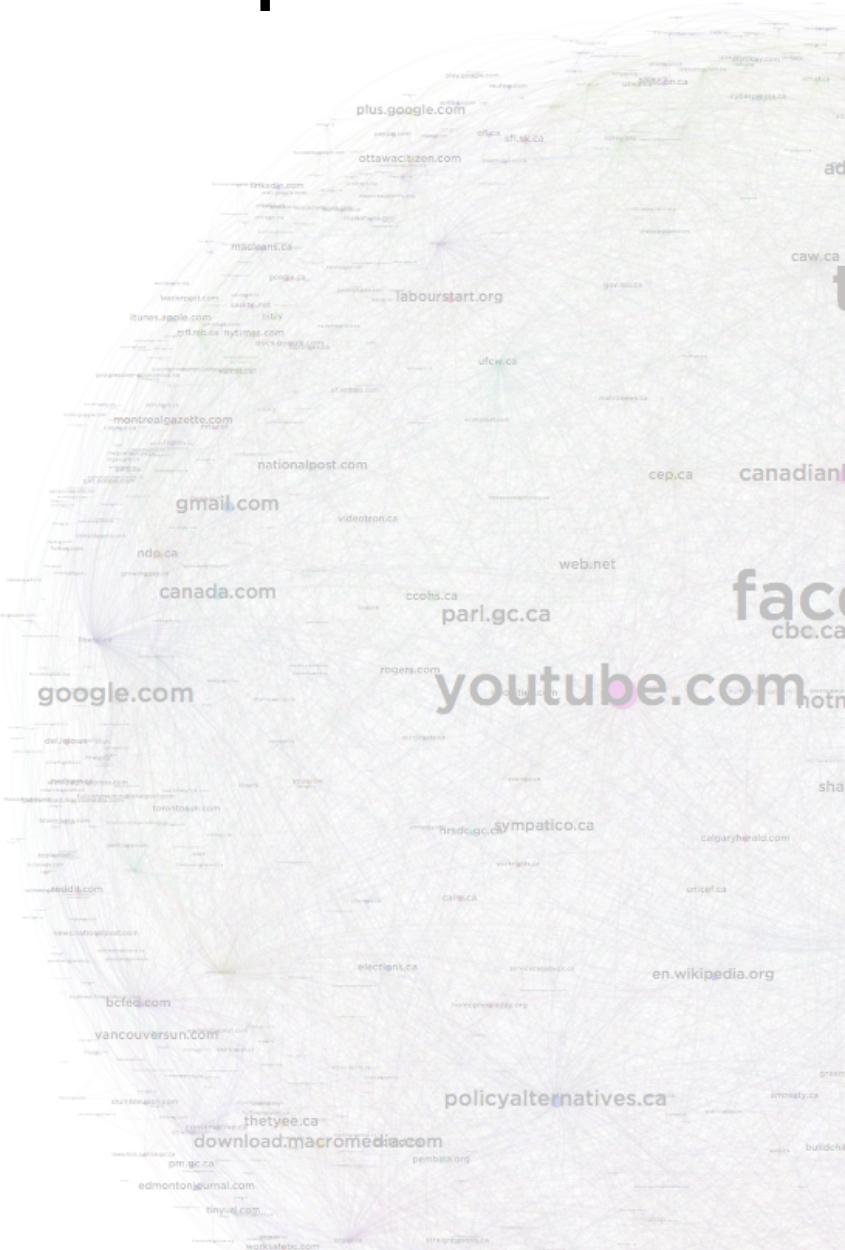
ICELAND

# Plan for the Talk

- What is the Canadian Web and why isn't the .ca enough?
- What is the current mandate to collect and preserve the Canadian web, both at a national and also institutional levels?
- What should we do moving forward?

# Our starting assumptions

- The Web preserves the early digital history of any nation – therefore absolutely essential to conduct harvests to capture Web as a primary source of the 20th and 21st century.
- Same principles and reasoning as national legal deposit programs!



# What is the Canadian Web?





**is not enough**

# History of the Canadian ccTLD

- ccTLD system (first in July 1985 with .uk)
- First .ca domain in 1988 with upei.ca
- Managed by University of British Columbia Computer Scientist John Demco



# History of the Canadian ccTLD

- **Structured**
  - Top-level domain: .ca
  - Second-level domain: province
  - Third-level domain: city
- **Standard domain: entity.city.province.ca**
- **Needed to prove national presence (i.e. business registered in two or more provinces) to have a simple .ca domain.**



# History of the Canadian ccTLD

- **June 1997:** Net97 Meeting to move domain from public to private hands;
- **2000:** Transfer domain to the Canadian Internet Registration Authority (CIRA)
  - 60,000 domains (2000)
  - 250,000 domains (2001)
  - 500,000 (2004)
  - 1,000,000 (2008)
  - 2,000,000 (2011)
  - 2,500,000 (2016)
- **Today the 14th largest TLD on the live Web.**

# History of the Canadian ccTLD

- Who uses it?
- Branch Plant Economy and Public Sector Institutions
  - ford.ca
  - starbucks.ca
  - uwaterloo.ca
  - canada.gc.ca
  - ianmilligan.ca



Canada

UNIVERSITY OF  
**WATERLOO**



Government  
of Canada



# History of the Canadian ccTLD

- **Who doesn't use it?**
  - **Media** ([globeandmail.com](http://globeandmail.com), [TorontoStar.com](http://TorontoStar.com), [FinancialPost.com](http://FinancialPost.com), [CanadianBusiness.com](http://CanadianBusiness.com))
  - **Large companies** – of largest 10, only three use .ca (leading banks, energy, resources)
  - **Many academics, individuals, and beyond**





.ca ccTLD not  
enough!

# State of the Canadian Web

- Canadian culture is produced and consumed online, and has been since the late 1990s
- Historians need to care about web archives!

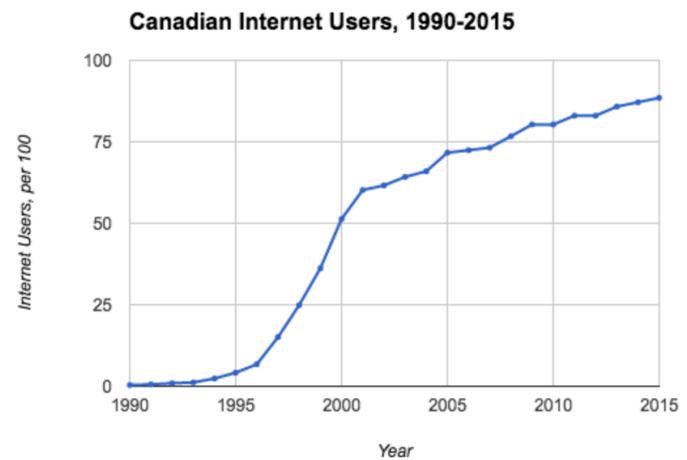


Figure 1 ref

[http://data.worldbank.org/indicator/IT.NET.USER.P2?cid=GPD\\_44&locations=CA](http://data.worldbank.org/indicator/IT.NET.USER.P2?cid=GPD_44&locations=CA)



# The 1990s are history!

The image is a collage of three distinct photographs. The leftmost photo shows two ice hockey players in action on the ice; one player wears a red and white Canadian jersey, and the other wears a white USA jersey. The middle photo is a close-up of a gas pump sign for 'PETRO-CANADA' with the price '62.9' per litre. The rightmost photo shows a construction worker in a hard hat and safety vest working on a steel structure, with a Canadian flag flying in the background.



**What is the mandate to  
collect and preserve the  
Canadian Web?**

# Canada and Digital Heritage

- ***Library and Archives Canada Act*** (2004)
  - “Representative sample of the documentary material of interest to Canada that is accessible to the public without restrictions through the Internet or any similar medium.”
  - The main instrument, section 8(2), enables **selectivity** rather than **comprehension**.
  - Gathers material as a **library** (mostly under legal deposit), no formal instruments call the web archival

**Operates mostly under  
Section 10 (Legal  
Deposit)**

# Canada and Digital Heritage

- “**publication** means any library matter that is made available in multiple copies or at multiple locations, whether without charge or otherwise, to the public generally or to qualifying members of the public by subscription or otherwise. Publications may be made available through any medium and may be in any form, including printed material, on-line items or recordings.”

# Canada and Digital Heritage

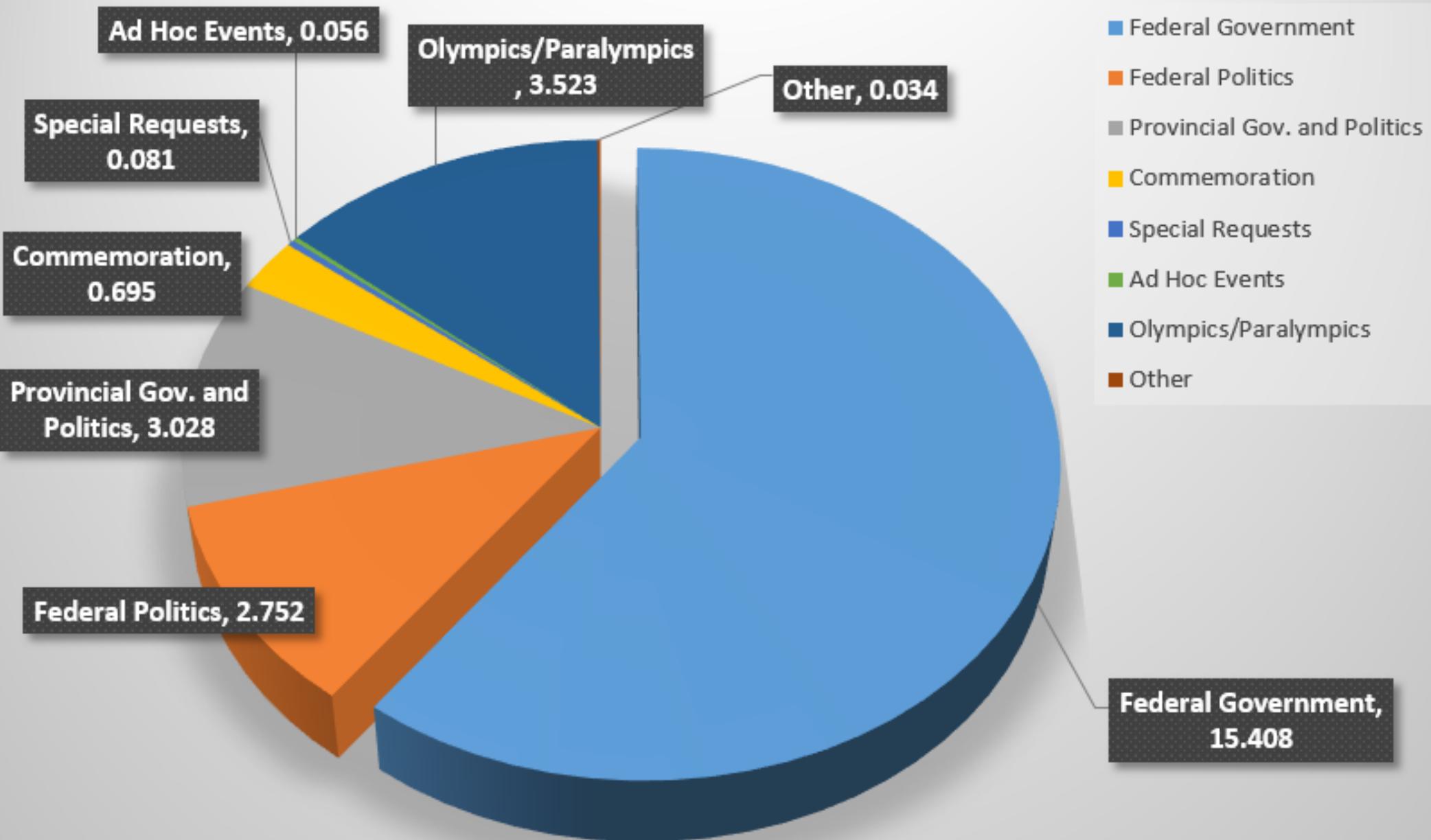
- **Publisher** means a person who makes a publication available in Canada that the person is authorized to reproduce or over which the person controls the content. It does not include a person who only distributes a publication.

However, LAC has largely approached web archiving along the lines of **sampling**.

# Collections to date

- **Federal government** (2005, 2006, 2007, 2013-15, 2016-)
- **Provincial/Territorial governments** (2006, 2008, 2009)
- **Thematic collections**
  - Olympics, Elections (2006, 2008, 2011, 2015)
  - Commemorations (state funerals, royalty visits, etc.)
- **Event-based harvesting** (2013, 2015, 2016 onwards)
- **Preservation harvesting** (upon request, or when resource taken offline)

# LAC Web Archival Data by Collection Category (TB)



<http://webarchive.bac-lac.gc.ca/?lang=en>

or

<http://archivesduweb.bac-lac.gc.ca/?lang=fr>

**So that's the scope of  
things.. what should we  
do?**

**It is the responsibility of the national memory institution to steward documentary heritage, and to ensure that it can evolve and manage the next generation mediums for recording and transmitting that heritage, as they are innovated.**

**Who can fill in the  
gaps?**



# Thanks America!

- **Internet Archive**
  - 2011 Wide Web Scrape (wide-00002)
    - 2.2 billion URLs, 29 million hosts
    - 160,884 .ca hosts, or 0.5%
    - 15th most-scraped top-level domain
  - Wide but shallow

# Moving beyond the .ca?



# Archive-It

- **Twelve Canadian Universities**

- Toronto, Alberta, Waterloo, Winnipeg, Wilfrid Laurier, Victoria, Saskatchewan, Manitoba, British Columbia, Carleton Dalhousie, Simon Fraser

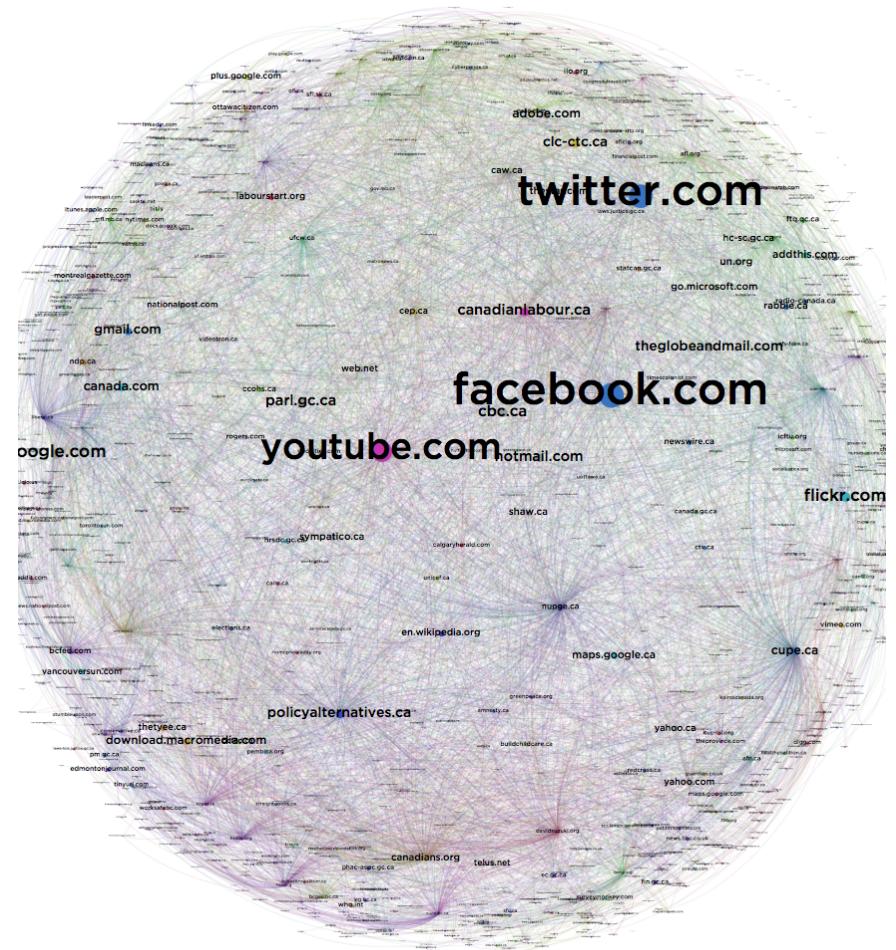
- **Other national memory institutions**

- Regional Municipality of Waterloo, National Gallery of Canada, Library and Archives Canada

The screenshot shows the Archive-It search results page for "University of Toronto". At the top, there's a navigation bar with links for HOME, EXPLORE, LEARN MORE, and CONTACT US, along with a logo for THE INTERNET ARCHIVE. A banner on the right side reads "The leading web archiving service for collecting and accessing cultural heritage on the web. Built at the Internet Archive". Below the banner, the "University of Toronto" entry is displayed, featuring its logo and basic information: "Archive-It Partner Since: Sep, 2005", "Organization Type: Colleges & Universities", and "Organization URL: http://www.utoronto.ca/". A detailed description follows, mentioning it's a network of 30 collections with over 15 million holdings, ranking third among research libraries in North America, and serving researchers across the university and around the world. The main content area is titled "Narrow Your Results" and lists categories: Subject, Creator, Publisher, Collector, Curator, and Notes. Under "Subject", there are links for "Politics & Elections (5)", "Society & Culture (4)", "Government (3)", "Government - Cities (2)", and "Arts & Humanities (1)". Other sections show results for "University of Toronto Libraries (2)", "Munk School of Global Affairs (1)", "University of Toronto Archives and Records Management Services (1)", "Hermonie Xie (1)", and "Notes (1)". A search bar at the bottom allows users to enter search terms or search the page text.

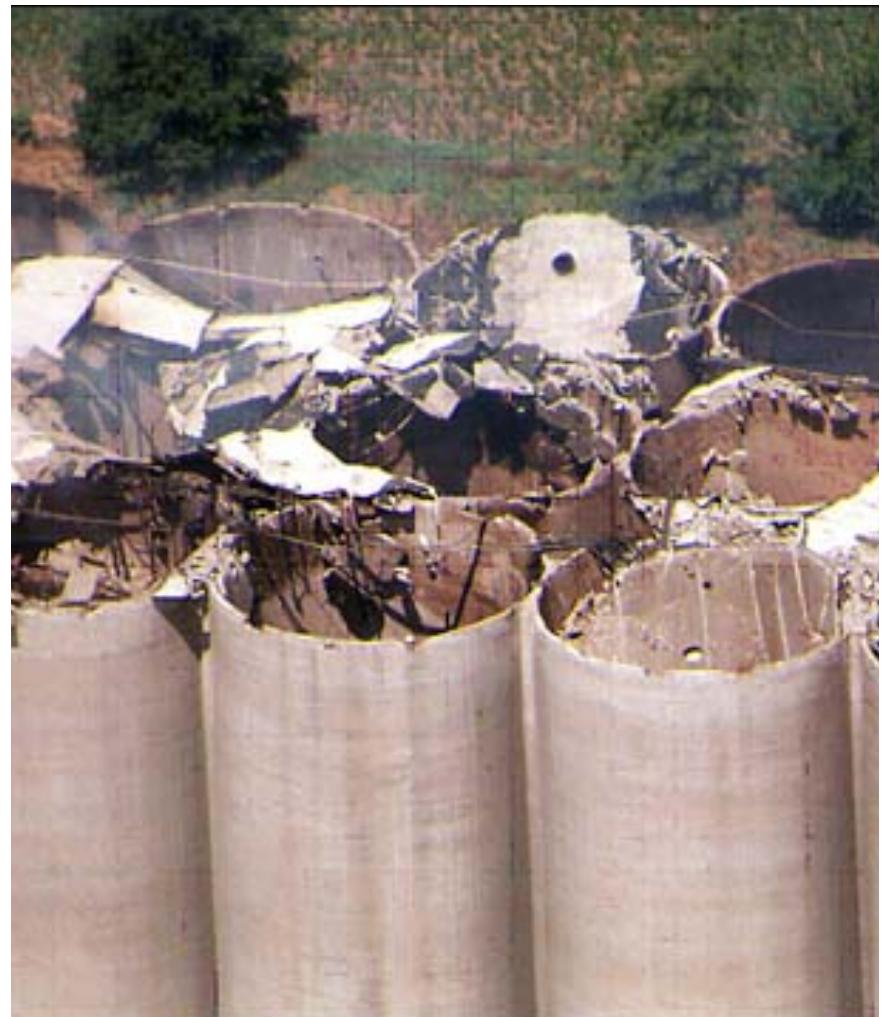
# Archive-It

- Thematic approach to Canadian culture – very important!
  - Example: Canadian Political Parties and Interest Group Collection



# Archive-It

- Silo'ed data
- Problem of overlap
- **Event-based model misses content that doesn't belong to an event**
- How do document the social history & daily life of Canadian communities?



Archive-It - Canadian Political | x Ian

https://archive-it.org/collections/227

Apps Keep Lib GitHub AWS HistD GM RSS Globe WALK LEARN FT Other Bookmarks Login

HOME EXPLORE LEARN MORE CONTACT US

The leading web archiving service for collecting and accessing cultural heritage on the web  
Built at the [Internet Archive](#)

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

UNIVERSITY OF TORONTO LIBRARIES

## Canadian Political Parties and Political Interest Groups

Collected by: University of Toronto

Archived since: Oct, 2005

Description: Canadian Political Parties and Political Interest Groups will archive the websites of all of the national Canadian political parties, and a number of special interest groups across the political spectrum.

Subject: [Politics & Elections](#)

Collector: [University of Toronto](#)

### Narrow Your Results

Subject

Sort By: Count | (A-Z)

New Democratic Party of Canada (2)

Assembly of First Nations (1)

Bloc Québécois (1)

Canada First (1)

Canada West Foundation (1)

More ▾

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here

Search

Clear

Sites

Search Page Text

Page 1 of 1 (78 Total Results)

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



**Need to  
collaborate,  
cooperate, and  
ensure things don't  
fall in the gaps...**

# Library and Archives Canada

- **Truth and Reconciliation Commission**
  - Led by Universities of Winnipeg, Manitoba, and the National Centre for Truth and Reconciliation
  - LAC forms core of collaborative collection, allowing others to concentrate on regional resources and perspectives
  - 200 seeds
- Universities of Winnipeg and Manitoba developing metadata schema specific to the TRC
- Central hub at the NCTR



# Towards a National Web Archiving Strategy

- **Web Archives for Longitudinal Knowledge (WALK) Project** - Waterloo and York (Nick Ruest), supported by Compute Canada
  - Current partners: Alberta, Toronto, SFU, Winnipeg, Victoria, Dalhouse
  - ~ 20 TB of Web Archives
- Common discoverability interface
  - Project Blacklight portal
  - Common solr index



# Towards a National Web Archiving Strategy

- **Providing all derivatives to researchers** (via project webpage, GitHub, and our provincial library consortium institutional repository)
- **Providing crawl analytics**
- Becoming a one stop shop for all special collections across the country?



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.

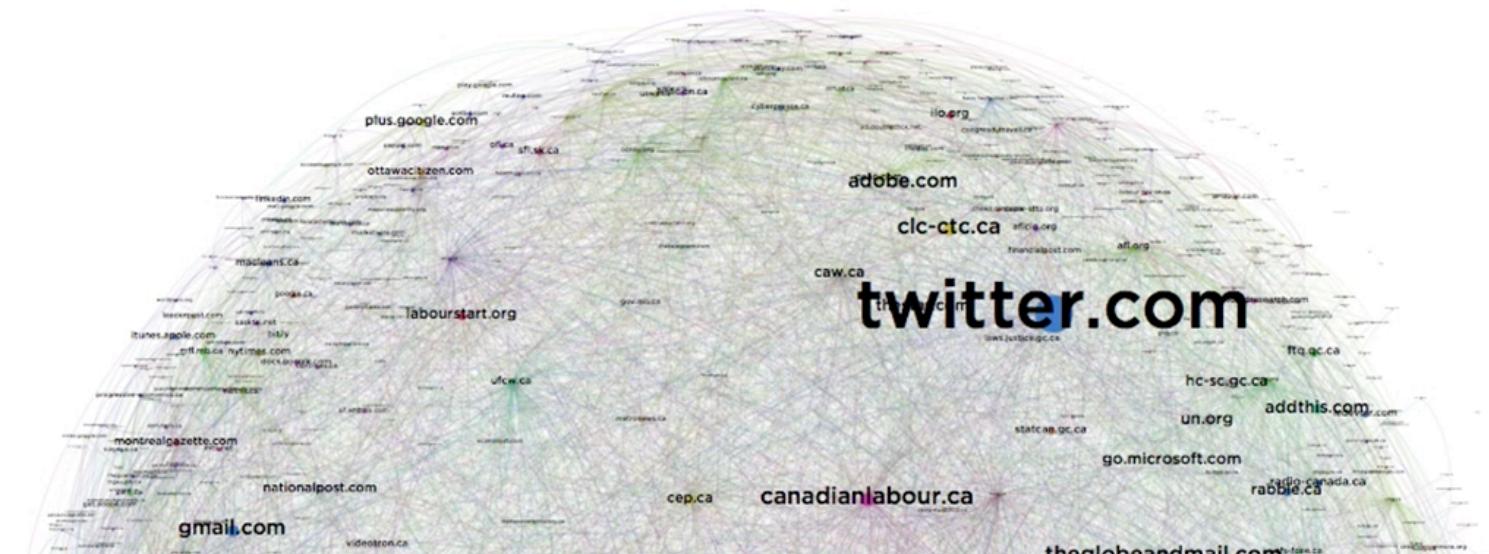
Currently, this is a prototype site providing access to one such archive, the University of Toronto's Canadian Political Parties and Political Interest Groups collection. This website allows you to search content from 50 political parties and political interest groups, from October 2005 to March 2015.

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.



lan

GitHub, Inc. [US] https://github.com/web-archive-group/WALK-CrawlVis

Apps GMail Lib GitHub AWS HistD RSS Globe WALK LEARN FT Blacklight Blacklight repo Concur Other Bookmarks

This repository Search Pull requests Issues Gist

web-archive-group / WALK-CrawlVis Unwatch 1 Star 0 Fork 0

Code Issues 2 Pull requests 0 Projects 0 Wiki Pulse Graphs Settings

No description or website provided. — Edit

29 commits	2 branches	0 releases	1 contributor
Branch: master	New pull request	Create new file Upload files Find file	Clone or download
ianmilligan1 final md fix on link			Latest commit f70472f 3 days ago
Code	fixed typo	3 days ago	
Raw	intermediate processing	3 days ago	
crawl-sites	checking in new analyses	3 days ago	
README.md	final md fix on link	3 days ago	
WORKFLOW.md	checking in workflow	2 months ago	
d3.min.js	adding here	4 months ago	
d3.v3.min.js	switching to v3 min	4 months ago	
README.md			

## WALK-CrawlVis

This repo contains the crawl visualizations for WALK project collections.

### Info

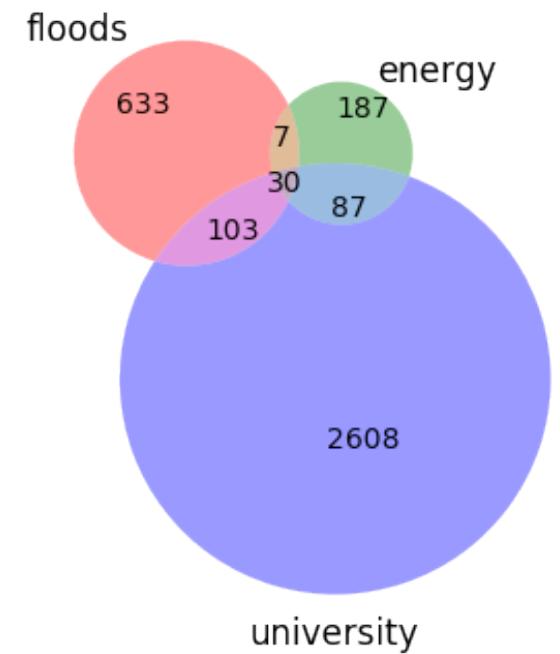
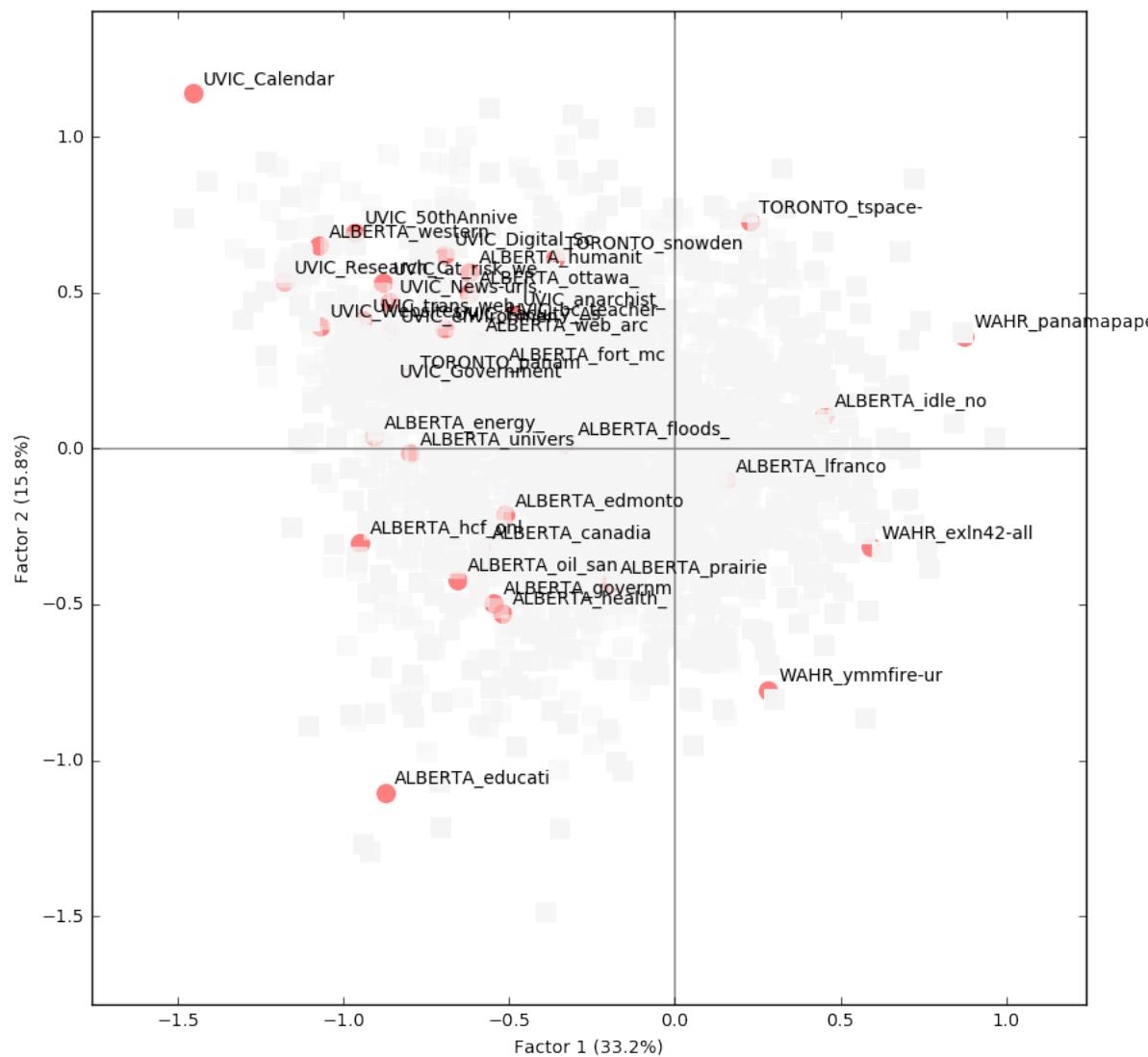
We are still working on basic descriptions for collections. In short, what you might expect to find isn't always what you will find. The "humanities computing" collection, for example, was probably a class project not a web archive of humanities computing projects.

### Current List of Visualizations



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

# Exploring collection coverage and curatorial models



# **Moving towards domain crawling with CIRA**

# Conclusions

- Difficult to define the Canadian (not just .ca) top-level domain;
- Difficult to resource – how to make web archiving a priority over other, more traditional activities?
- But scattered activity, together, points towards a vibrant future.

In Canada – our goal is to start pressuring our heritage institutions to...



**MAKE IT SO!**



Social Sciences and Humanities  
Research Council of Canada

Conseil de recherches en  
sciences humaines du Canada

Canada



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



Bibliothèque et Archives  
Canada



UNIVERSITY OF  
**WATERLOO**

Library and Archives  
Canada

# Thanks!

---

**Ian Milligan**  
Assistant Professor  
@ianmilligan1



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

**Tom Smyth**  
Manager, Digital Operations



Bibliothèque et Archives  
Canada

Library and Archives  
Canada