COVID, the Crown in Canada, and Ukraine: Recent Collections and the Forthcoming Launch of the new Canadian Web Archive

Tom J. Smyth

Program Manager,

Web and Social Media Preservation Program

The National Web

Archive is now 109+ terabytes and over 3(.018) billion assets (& 86.8 million Tweets!)

What is Web Archiving?

- Is an

internationally-practiced, digital preservation and curation based discipline that guarantees future access to unique resources from the Web

- "Web Archiving" involves the use of specialized hardware and software to target, make precise copies, download to local servers, and emulate the original published and interactive context of web resources via an access portal – while securing their data permanently on the institution's side via digital preservation

Canadian Legislative

Context:
LAC Act section 8 (2)

Powers of Librarian and Archivist

- *8 (1) The Librarian and Archivist may do anything that is conducive to the attainment of the objects of the Library and Archives of Canada, including*

Sampling from Internet

- *(2) In exercising the powers referred to in paragraph (1)(a) and for the purpose of preservation, the Librarian and Archivist may take, at the times and in the manner that he or she considers appropriate, a representative sample of the documentary material of interest to Canada that is accessible to the public without restriction through the Internet or any similar medium.*

Acquisition Methodology

1. Comprehensive collection of the GC web presence (2005 -)

2. Thematic web and social media research collections (2009 -)

3. Events-based collections and news media (2013 -) 4.

Rescue or preservation harvesting (2005 -)

5. Acquisition of nominated resources (2005 -)

More than Data Management & Collection…

- More than web resources, we curate research data collections with the requirements of digital humanities/scholars in mind.
- Digital documentary heritage evidencing *“Canadian living history”.*
- “20 years from now, when a digital historian sits down to write the history of COVID-19 and its impact on Canada, what primary sources and data would they wish they had?” Our curation.
- *COVID-19 has demonstrated that web archiving is one of the few immediate actions information professionals may take now -- to preserve a fulsome historical timeline and its primary resources.*

A Word on

Twitter…

• We’re monitoring the ‘Qwitter’ / situation •

Had you heard of Mastodon before Musk? •

(Doesn’t yet have the scale)

• Some tools do exist for collecting Mastodon data •

Since we collect thematically, LAC won’t be doing emergency Twitter archive exports

• Celebrities / people seem to be choosing Instagram…

COVID-19

Documenting COVID-19 in Canada…

Activity:	Total:
Total news/media websites crawled daily for COVID-19, YTD	34
Total non-media resources selected for COVID-19, YTD	2,025
Total digital assets collected for COVID-19, YTD	594 mil
Total data collected for COVID-19, YTD	19.88 TB
Tweets captured for COVID-19, YTD	4.90 mil

Death of HM Queen Elizabeth II; ascension of HM King

Charles III

Crown in Canada Collection

Activity:	Total:

Total non-media resources selected, YTD	78
Total digital assets collected for COVID-19, YTD	~330,329
Total data collected for COVID-19, YTD	81.8 GB

IIPC WEBINAR: WEB ARCHIVING THE WAR IN UKRAINE

The main focus of this webinar is present efforts around archiving Telegram which became one of the most important communication platforms during the war in Ukraine (OPORA survey, June 2022). The leads of the new IIPC collaborative collection "War in Ukraine" will also give an update on the results of the first crawl. Recordings now available for IIPC members: https://netpreserve.org/members-only-archive/ukraine-webinar/

I. The Telegram Archive of the War: The Center for Urban History in Lviv

DATE

31 Aug 2022

Expired!

TIME

3:00 PM - 4:15 PM

LOCAL TIME

Timezone: America/Toronto
Date: 31 Aug 2022
Time: 11:00 AM - 12:15 PM

CATEGORY

LAC (and IIPC) Ukraine Collection

Activity:	Total:
Total non-media resources selected, YTD	174
Total digital assets collected for COVID-19, YTD	~1.86 mil
Total data collected for COVID-19, YTD	272 GB

Presenting…

The Greatly Improved!

Government of Canada Web Archive

Canadian Web Archive

The Web Archive Access Portal of Library and Archives Canada

Search

All Collections | Search

Advanced Search

◉ Keyword Search ○ URL Search

Browse collections

Coronavirus/COVID-19

This collection provides access to websites related to the COVID-19 pandemic in Cana captured between February 2020 and the present. These resources document the response to the crisis from public health agencies, governments, charities, and other groups, and the impact of the pandemic on life in Canada.

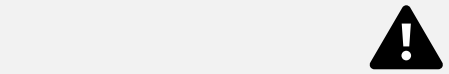
Collections Available for Launch

- Truth and Reconciliation Web Archive (curated in collaboration with the National Centre for Truth and Reconciliation and the U of Winnipeg and

Manitoba)

- COVID-19 and its Impacts in Canada
- All federal data collected since 2005
- Approximately ~65 of 109 tera will be available

Federal Collections Discovery



Advanced Full Text Search

Thematic Collection Discovery

Collection Development Approach & Curation / Discovery

- Thematic web archival collections curation is driven by specialized collection development policies (e.g., COVID)
- The CD policy defines the goals of the collection and defines the subthemes to be documented (e.g., impact of COVID on business and economy at all levels)
- Each subtheme receives a priority level and a quality control service level, and are then assigned to curators to select pertinent resources for inclusion in the collection

Integration with Fed Search

- ~3 billion documents to index is no small task!
-

Integration of web archives full text search with LAC

collections fed search (as a facet) will be investigated

- In the interim, we are generating library bibrecords for web archives at collection level
- The collection-level bibrecords will aid users in fed / library search to discover and access LAC's web collections

A Word on Data and Search vs Curation

- (Imagine curating / arranging every document or producing a bibrecord for each resource?)

- To avoid untenable work, we must index at domain level; behind this arrangement, thousands of resources and snapshots likely exist (e.g., Prime Minister)
- Thus, not all seeds in the archive will be discoverable / seen by browsing at this time; users should rely on drilling down into collections and targeted full text search

“Curated Discovery and Search”• GCWA

- Thematic Collections
  - Olympic and Paralympics
    - Olympics and Paralympics, Vancouver

2010 Economic Impact on Vancouver/Canada <Infrastructure Development>

Full Text search

Federal and Provincial Elections, 2005-present

- National Historical Commemorations, 2006-present
- Media and Newspaper Collections, 2016-
- Indigenous Collections
- Canada 150
- Royal Commissions and Commissions of Inquiry
- Events-based Collections (Lac-Megantic, Forest Fires, Humboldt…)
- Historical HTML Official Pubs and Bibliographies (1994-2005) •

Many more

Future State

• Curate and roll out all historical collections (2023-) •
Proceed to possible *.ca domain crawls in the near future •
Watch the situation with:

- COVID-19
- Public Order Emergency Commission (POEC)
- Twitter
- Ukraine
- Alberta Sovereignty Act

Tom J. Smyth, MA, MISt, RMP

Manager, Web and Social Media Preservation Program Tom.Smyth@bac-lac.gc.ca
@smythbound

Questions/comments welcome

archivesweb-webarchives@bac-lac.gc.ca