Rapid Response Methodologies and Projects: Documenting National Historic Events at Library and Archives Canada

Tom J. Smyth

Manager, Web and Social Media Preservation Program

Digital Collections Operations Division

@smythbound

Library and Archives Canada

Library and Archives Canada

Bibliothèque et Archives Canada

Canadä

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.*

Web Archival Acquisition Methodology

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

2. **Thematic web & 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-)

Major Historical Events and Impacts on Curation Methodology

- In the early program history (~2011), ad hoc reaction to major events consisted of scrambling to get the best media resources documenting the issues

- We needed a means to get ahead of events (that by nature could not be planned or anticipated)

- In 2016 we began daily harvest of a broad range of news media; this began picking up information on COVID-19 when it emerged in the news in December 2019.

Major Historical Events and Impacts on Curation Methodology (2)

- Project management methodology: define the themes to document within the historical event, by what priority/sequence, scale or degree of comprehensiveness, quality control level
- Allow the scope to evolve over time as the historical event itself warrants / based on he new themes and resources generated
- Reduce the collection of redundant social media across platforms
- Rely on API based social media collection where possible (Twarc)
- Approach the collection of social media by theme; target and collect e.g., the most pertinent hashtags from Twitter by highest traffic

Canadian News Media Collection (2016-)

- Automated collection of news media currently comprises:
  - 335 seeds
  - 4,535 crawls
  - ~656 million objects
  - ~18 Terabytes and growing daily
- Daily news collection captures front page news and supplies an ongoing historical documentation/timeline for major events
- Frees resources to curate and concentrate on selecting non-media resources related to a given theme for better ROI
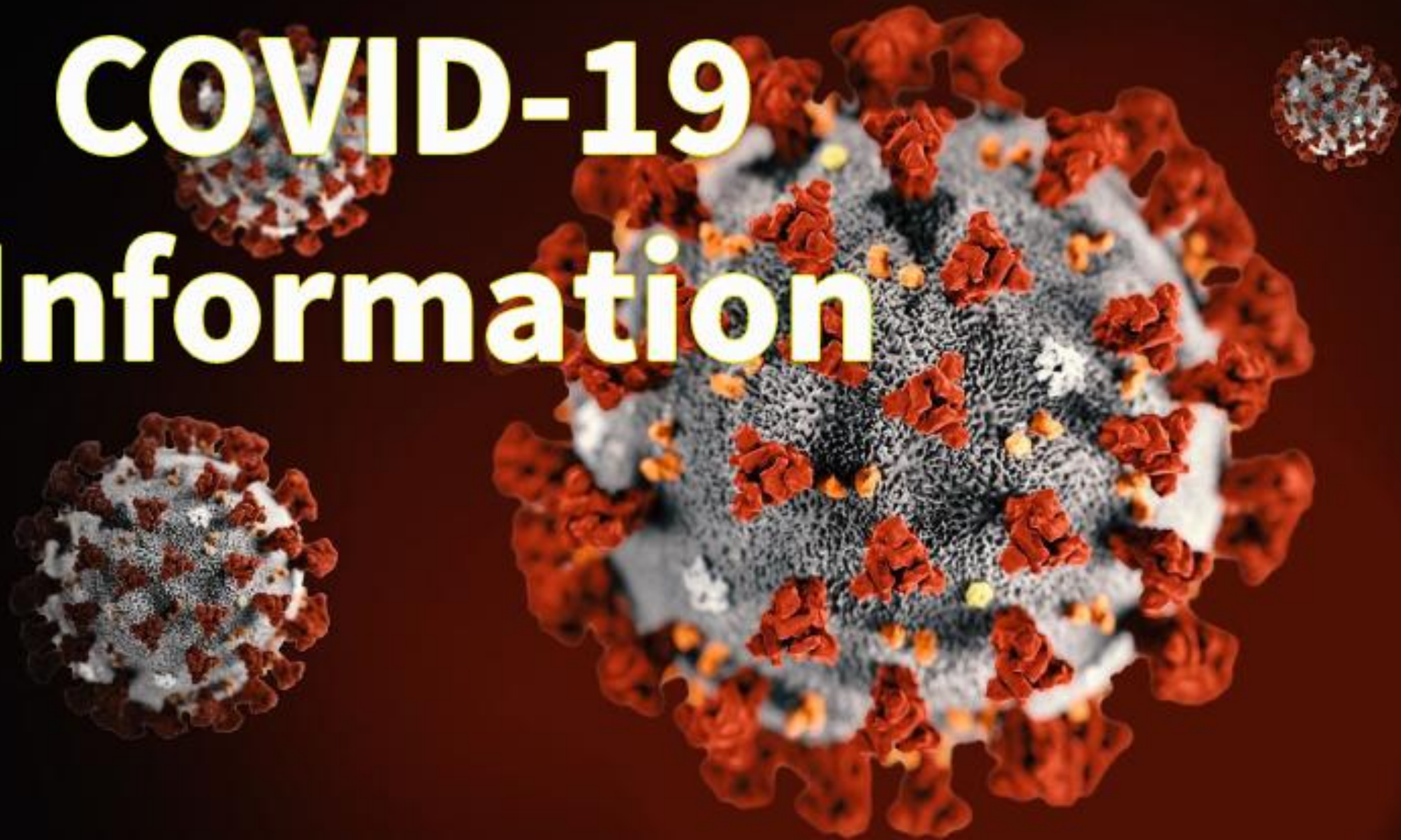
Canadian News Media, Overview

- 48 total, 34 done daily:
- 35 English, 10 French, 11 Indigenous, 2 multicultural

- Acadie Nouvelle
- Bloomberg Canada
- CBC
- City News Toronto
- CTV News
- Financial Post
- Global News
- Globe and Mail

- Huffington Post Canada
- iPolitics
- Journal de Montreal
- Journal de Quebec
- La Presse
- Le Devoir
- Le Droit
- National Post

- Rabble.ca
- Radio-Canada
- Toronto Star
- Vancouver Sun
- Whitehorse Daily Star
- Winnipeg Free Press
- Macleans - COVID-19

Sample Previous and Recent "Rapid" Events

Collection Title	#Seeds/ Crawls/Objects	Data Volume:
Lac-Megantic Rail Disaster (2013)	35 / 44 / 388k	19 GB
Shooting on Parliament Hill (2014)	351 / 17 / 545k	20 GB
Humboldt Broncos Bus Accident (2018)	253 / 220 / 787k	29 GB
Nova Scotia Mass Casualty Commission	120 / 370 / 731k	215 GB
Black Lives Matter / Indigenous Lives Matter	199 / 482 / 2.01m	187 GB
Residential Schools Grave Discoveries	45 / 55 / 60k	50 GB
COVID-19 – Truck Convoy Protests	168 / 735 / 1.46m	568 GB

COVID-19
Information

COVID-19 Project Scope

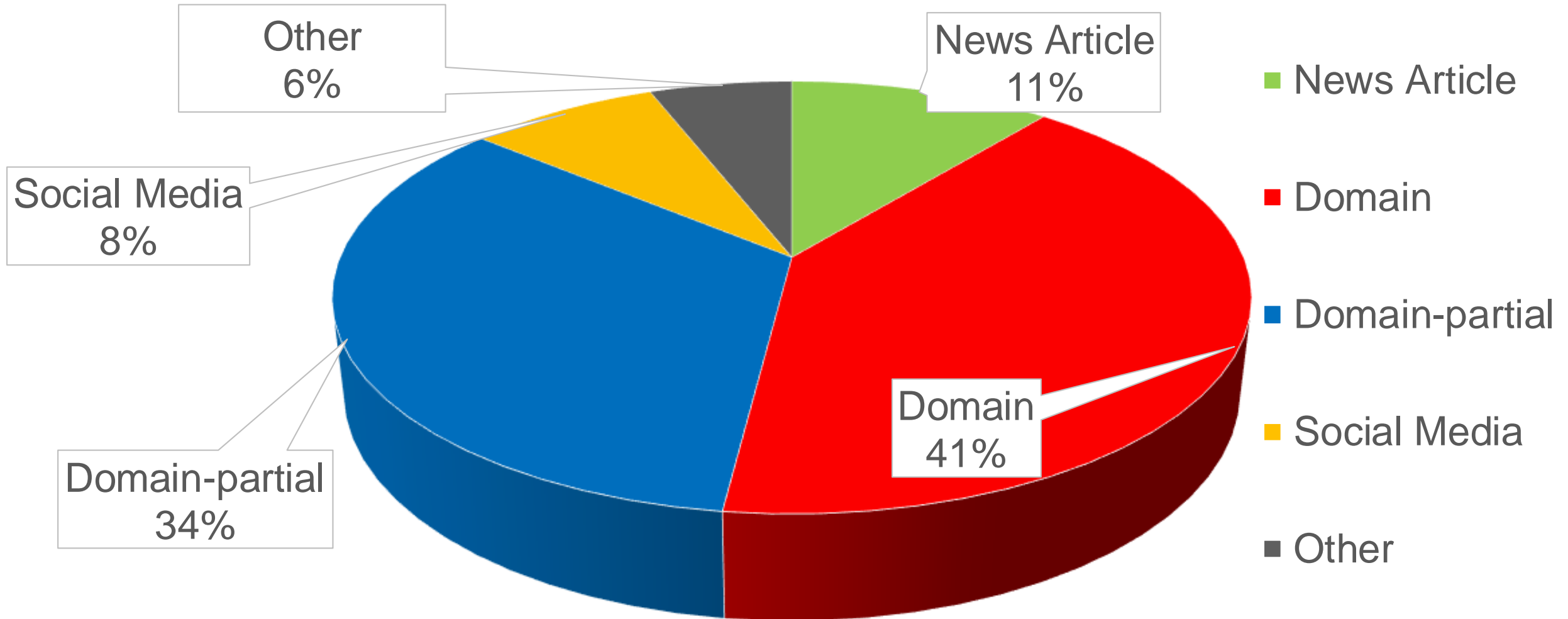| Project Scope By Theme:	#Domains
Arts & Culture	302
Business & Economy	195
Charities	209
Families & Education	115
Federal, Provincial, and Territorial Govt	257
Health Science & Academic	587
Indigenous Perspectives	35
Protests / Alternate Perspectives (e.g., Truck Convoys)	141
Regional Public Health	119
Religion	88

Thematic Twitter Data

- **~3.83 million tweets collected**
  - **COVID in Canada:** #covidcanada, #covid19canada, #canadalockdown, #canadacovid19
  - **The mask movement:** #maskupcanada, #masks4canada
  - **Day of Visibility for the Live Event Community:** #lightuplive, #eclaironslesscenes
  - **COVID by province:** #COVIDBC, #COVIDAB, #COVIDSK, #COVIDMB, #COVIDON, #COVIDQC, #COVIDNB, #COVIDNS, #COVIDPEI, #COVIDNFLD
  - **Truck Convoy Protests 2022**: #convoidelaliberte, #freedomconvoycanada, #freedomconvoy2022, #truckerconvoy2022, #freedomconvoy, #truckersforfreedom, #truckersforfreedom2022, #convoyforfreedom, #convoyforfreedom2022

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	~2015
Total digital assets collected for COVID-19, YTD	~478 mil
Total data collected for COVID-19, YTD	~16 TB
Tweets captured for COVID-19, YTD	~3.83 mil

COVID-19 Web Archive Resource Distribution

- News Article 11%
- Domain 41%
- Domain-partial 34%
- Social Media 8%
- Other 6%

(Web) Archival Finding Aids

- The 'end state' of the web archival scoping document that includes the details of how the initial scope of the project evolved over time
- Published alongside the actual collection (archival practice) to assist digital humanities researchers that use the Web as primary historical source
- Details the end-state of the collection's metadata schema, how it evolved over time to reflect the scoped-in-content, how to use the controlled vocab to search the collection
- Defines what as deliberately excluded (scope-out)
- Enables researchers to see at-a-glance whether the collection or dataset addresses their research questions (without slogging through it by browsing or search)

Program Future State

- Continue collecting information related to COVID-19 in Canada and related issues (as long as it takes…)

- Deploy a new Government of Canada Web Archive access portal in Q3 of 2022-23 (on schedule).

- Expand to Canadian domain crawls (all of .ca).

- Continue to document Canadian perspectives and news media on Ukraine.

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

Questions/comments welcome
archivesweb-webarchives@bac-lac.gc.ca