Curation Methodology
for the Federal Elections 2019

Tom J. Smyth
Manager, Digital Integration
Digital Preservation and Migration Division
@smythbound
tom.smyth@canada.ca
What is Web Archiving?

• Involves the use of specialized software to:
  – Target and copy web resources over www
  – Download the data to the server-side
  – Emulate the original published and interactive context of the web resource via an access portal

• Is an internationally-practiced, digital preservation-based discipline that guarantees future access to resources that constitute digital documentary heritage

  …where this is otherwise precarious
Program History

- 2005 - LAC formally launched its Web Archiving Program with a crawl of the GC web domain.
- 2008 - LAC was a founding member of the International Internet Preservation Consortium (IIPC) and has held the chair twice (2007, 2016), and the vice-chair in 2018. It currently serves as the Treasurer (2019).
- 2009 – Methodology first developed for curating a thematic collection on the Vancouver 2010 Olympic Games.
- 2013 – LAC began collecting in support of TBS’ Web Renewal Initiative (WRI) and began reporting this work as part of LAC’s contribution to Open Government.
- 2014 – Moved to Internet Archives’ Archive-IT Cloud Platform. This enables staff to select and preserve a minimum of 13 terabytes per fiscal going forward.
- 2017 – Launched the Truth and Reconciliation Web Archive in partnership with National Centre for Truth and Reconciliation.
- 2018 – Passed the milestone of over 1 billion assets in the web archive.
Web archiving has its own unique authority under the *Library and Archives of Canada Act* section 8(2):

- Web archiving is conducted specifically for digital preservation purposes;
- Web Archiving includes social media, which has the Web as medium; and
- Is considered a distinct stream of acquisition at LAC.
Web Archiving Program Methodology

We conduct five main activities:

1. Comprehensive crawls of the Government web presence
2. Curation of thematic web archival and social media collections
3. Broad collection to document topics in Canadian society
4. Reactive curation to document significant Canadian events
5. Acquisition of resources upon request (nominations)
Defining the Federal Elections Collection

• Starts with a scoping document that functions also like a web archival collection development policy:
  – What are we documenting? Defines the topics for curation
  – Prioritizes topics in terms of importance to the overall collection
  – Defines degree of harvesting comprehensiveness by topic
  – Defines level of quality control to be conducted by topic
  – Scoping document capturing these details at project close-out by morphing into a “finding aid” to be published along with the collection
# Elections Topics Defined

<table>
<thead>
<tr>
<th>Topic</th>
<th>Seeds</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Political Parties</td>
<td>391</td>
<td>Captures all MP personal sites via the party domain</td>
</tr>
<tr>
<td>Civic Engagement</td>
<td>25</td>
<td>Non-partisan groups encouraging voting/democracy</td>
</tr>
<tr>
<td>Political Opinion</td>
<td>16</td>
<td>Partisan Editorial material; moving mostly to social media</td>
</tr>
<tr>
<td>Attack Ads/Sites</td>
<td>12</td>
<td>Partisan attack pages or ads</td>
</tr>
<tr>
<td>Elections Issues/Advocacy</td>
<td>43</td>
<td>Carbon tax, climate change, human rights, think tanks</td>
</tr>
<tr>
<td>Satire</td>
<td>8</td>
<td>Everyone loves political satire!</td>
</tr>
<tr>
<td>News Media</td>
<td>28</td>
<td>Subsections of news media devotes to the elections</td>
</tr>
<tr>
<td>Federal Government Resources</td>
<td>39</td>
<td>GC resources likely to change if the government does</td>
</tr>
<tr>
<td>Polls/Stats/Forecasts/Results</td>
<td>22</td>
<td>Collection of the main polls over the collection period</td>
</tr>
<tr>
<td>Debates</td>
<td>27</td>
<td>Commentary on impact of the debates</td>
</tr>
<tr>
<td>Social Media</td>
<td>228</td>
<td>Discussed in more detail in a bit….</td>
</tr>
</tbody>
</table>
Collections Methodology

• Most seeds were crawled:
  – Once a month or more between August-September
  – Two weeks+ before election on 21\textsuperscript{st} October
  – One week after the results were known
  – At the time of swearing-in on 20\textsuperscript{th} November and resulting Cabinet shuffle
  – Selected news media sites were collected in their entirety
  – Some seed groups to be collected quarterly or ongoing hereafter

• Method captures a temporal record of elections lead-up, events, results, and then the post-election state of the government
Social Media for the Elections

- Twitter hashtag analysis was conducted:
  - LAC began collecting #cdnpoli on an ongoing basis in Feb 2018.
  - #canpoli and #polcan represent the next highest volume hashes on Canadian politics, so were added for ongoing collection in May 2019.
  - The three hashes were collected in a single dataset to isolate elections commentary starting May 2019 (the basis of stats to follow).
  - #elxn43 was selected as the highest volume hashtag related to the 43rd federal election (LAC collected over 2 million #elxn42 tweets in 2015).
  - Hashes receive bilingual tweets (#polcan mostly French) so this also captures the election dialogues in both official languages.
# Elections Twitter Breakdown

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>#ELXN43</th>
<th>#CANPOLI, CDNPOLI, POLCAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection Period</td>
<td>19&lt;sup&gt;th&lt;/sup&gt; July to 21&lt;sup&gt;st&lt;/sup&gt; November</td>
<td>24&lt;sup&gt;th&lt;/sup&gt; May to 21&lt;sup&gt;st&lt;/sup&gt; November</td>
</tr>
<tr>
<td>Total tweets</td>
<td>4,436,631</td>
<td>9,634,884</td>
</tr>
<tr>
<td>Unique users</td>
<td>322,722</td>
<td>545,292</td>
</tr>
<tr>
<td>Tweets sent on 21&lt;sup&gt;st&lt;/sup&gt; Oct</td>
<td>209,918</td>
<td>158,301</td>
</tr>
</tbody>
</table>

- Total: 14,071,515 tweets collected
- Throughout the period of collection, 868,014 unique users commented
- These users sent 368,219 tweets on election day
Highlights 2019: What did we get?

• Captured all the sites of the sitting MPs in the House
  – Where not present at the party, we got the Legisinfo or Facebook
• Periodic capture of all party sites registered with Elections
• Seedlist of core GC resources that would be altered by a change of government (GG, PM, PCO, TBS, PARLinfo, etc.)
• **839 total seeds** were selected
• Over **14 million tweets** between May and November
• **3 TB** of data collected
• By comparison, **1.7 TB** was collected for the 2015 federal election
Observations Across Federal Elections

- Observed decline in MP’s use of official federal party domains to convey campaign-related information. Some official MP sites offer only very basic information even about the candidate.

- Parties structure their websites differently; for many, it was possible to get all party candidates by targeting the primary domains (candidate.party.ca; party.ca/candidate or /riding#. Others used “candidate.ca”.

- Unsurprisingly, social media has become the preferred medium for campaign messaging and announcements so we greatly expanded our SM strategy.

- Social and news media are particularly data intensive in the 2019 Federal Election Collection. Combined they account for around 1.25 TB of data or 40% of the total federal elections data.
## Other Recent Acquisitions, FY 2019-20

<table>
<thead>
<tr>
<th>Comprehensive</th>
<th>Social Media?</th>
<th>Curated</th>
<th>Social Media?</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC Web Presence</td>
<td>No</td>
<td>National Inquiry into Missing and Murdered Indigenous Women and Girls</td>
<td>Yes, curated</td>
</tr>
<tr>
<td>GC Official Publications</td>
<td>No</td>
<td>Legalization of Cannabis in Canada</td>
<td>Yes, curated</td>
</tr>
<tr>
<td>Tourism in Canada</td>
<td>No</td>
<td>Overlord D+75 - 75th anniversary of D-Day</td>
<td>Yes, curated</td>
</tr>
<tr>
<td>Provincial Domains</td>
<td>No</td>
<td>100th anniversary of CN Railway</td>
<td>Yes, curated</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Labour Unions in Canada</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100th anniversary of the Winnipeg General Strike / 125th of Labour Day</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Anniversaries of the birth of Louis Riel and Red River Resistance</td>
<td>TBD in Q4</td>
</tr>
</tbody>
</table>
Government of Canada Web Archive

Since 2005, Library and Archives Canada (LAC) has collected federal and non-federal web resources in the context of its Web Archiving Program. This website, the Government of Canada Web Archive (GCWA), provides access to archived federal websites.

What the GCWA contains

- Federal government information as originally published on the Web
- Federal web pages that are no longer available to the public
- Indices to explore our collection by organization name and by URL

Search

Browse by Institution  Browse by URL
LAC welcomes and pro-actively seeks out collaboration and internal/external expertise for web archival selection.

- Get in touch with project ideas! Planning starting up for FY20-21

- Anyone is welcome to nominate web resources by submitting the request via our program email address (web-archives-Web@bac-lac.gc.ca).

- Requests are honoured whenever possible!
Questions / comments welcome!
Tom J. Smyth
Manager, Digital Integration
Digital Preservation and Migration Division
@smythbound
tom.smyth@canada.ca

Library and Archives Canada
www.bac-lac.gc.ca

Telephone: 613-996-5115 or 1-866-578-7777  |  TTY: 613-992-6969 or 1-866-299-1699