Web Archiving at Library and Archives Canada: Methodology, Collections and Recent Developments

OCUL Digital Curation Summit 2016

Tom Smyth
Manager, Digital Operations
Preservation Branch
tom.smyth@canada.ca
Overview

1. Program Context
2. Web Archival Curation Methodology at LAC
3. Recent Collections
4. Relaunch of the Government of Canada Web Archive
5. Looking to the Future
LAC’s Legislative Authority for Web Archiving

Library and Archives of Canada Act

OBJECTS AND POWERS

Sampling from Internet

Section 8. (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”. 
LAC’s Web Archiving Program: Background

- International Internet Preservation Consortium (IIPC) was chartered in 2003 with LAC as a founding member (Chair in 2008 and again in 2015)

- LAC’s Web Archiving Program launched in December 2005

- Between 2005-8, activities concentrated on capturing:
  - Provincial and Federal Elections
  - Summer and Winter Olympics
  - Commemoration (State funerals; Calgary Stampede, the War of 1812)

- 2013: Early thematic collections documented e.g., the development of the “Idle No More” movement, the Keystone Pipeline, and Arctic Sovereignty)
Curation Methodology
Hello

Is this the page you’re looking for?

Lionel got a hold of your page...

To rescue it try searching again or visit the AC home page >>

search

report missing link >>
Why?

*Here today, gone tomorrow*

www.pm.gc.ca

2006-02-05                               2006-02-10
Why?

Here today, gone tomorrow

www.pm.gc.ca

19th October 2015

4th November 2015
What should be collected?

Web resources that are:

• In alignment with the institution’s themes of special emphasis (mission and/or heritage value)
• Valuable as research collections, where the resources may not be available in another format
• Valuable and appropriate for diverse use as datasets (context, perspective, bias)
• (Valuable to our stakeholders)
• Tenable and sustainable for preservation and provision of access
Toward a Collection Development Strategy for Web Archives

- LAC’s stated collection development priorities:
  - Commemoration of Canadian Historical Events
  - Aboriginal peoples
  - Canada’s Military Heritage
  - Atlantic, Western, and Northern Canada
  - Women
  - French-language
  - Multicultural Canada
Collaborative Collection Development

- Thematic projects have drawn on the expertise of internal/external specialists for collection scoping and to develop a collaborative seedlist
  - Political
  - Historical
  - Military
  - Social
  - Cultural
  - Economic, etc.

(What expertise could you contribute?)
How should we collect it?

Five Main Web Archiving Activities

1. Domain crawl of the federal web presence
2. Curation of thematic research collections
3. Reactionary, events-based collections
4. Preservation archiving of resources at risk
5. Acquisitions linked to existing library collections or archival fonds
Types of Thematic Collections

1. Domain-level harvests

2. Curated thematic research collections
   A) Ongoing Collections:
      – Commemorations (State Funerals, Historical Events, 2015-)
Types of Thematic Collections (2)

2. Curated thematic research collections
   
   B) One-time Collections:
   
   • Arctic Sovereignty, Development of the Keystone Pipeline (2013)
   • Centenary of the First World War (2015-)

3. One-time website harvests
   
   A) Events-based harvesting:
   
   • Disaster at Lac-Megantic (2013)
   • Shooting Incident at Parliament Hill (2014)
3. B) One-time Website Harvests: Preservation of Sites Now Offline

- Websites of Previous PMs
- Websites of Previous GGs
- PCH: Websites on the various Royal Tours
- Royal Commissions and Commisions of Inquiry:
  - Iacobucci, Oliphant, Major, and Cohen Commissions
- INAC/AANDC’s Aboriginal Portal
- First Nations Statistical Institute
- National Aboriginal Health Organization
- National Council of Welfare
- Federal Healthcare Partnership
- Canadian Employment Insurance Financial Board
- National Round Table on the Environment & Economy
- DFAIT: Afghanistan.gc.ca
- DFAIT: G8 and G20 sites
- INAC → AANDC
- CIDA → DFAIT
- HRSDC → ESDC
- Etc. Ask me for a resource!
About Us

The National Council of Welfare was established as an advisory group to the Minister of National Health and Welfare by the Government Organization Act of 1969. The Council now reports to the Minister of Human Resources and Skills Development Canada (HRSDC).

Mandate

The mandate of the Council is to advise the Minister regarding any matter relating to social development that the Minister may refer to the Council for its consideration or that the Council considers appropriate.

Role

The Council advises the Minister on matters concerning poverty, the realities of low-income Canadians; and related programs and policies by:

- communicating directly with the Minister;
- providing information, knowledge transfer and awareness of poverty-related issues to various stakeholders and the public; and
- providing a means for people concerned with poverty, especially low-income Canadians, to make their views known to government.

Function

The Council fulfills its mandate and achieves its objectives by:

- publishing a variety of reports and operating a website devoted to poverty and social policy issues;
- presenting submissions to groups such as parliamentary committees;
- communicating with the Minister on issues raised in reports and meetings of the Council;
- meeting with varied stakeholders to foster informed conversations about solutions to poverty; and
- responding to inquiries from researchers, members of the media and the public.
Δεν Δηλώνεται ο πρόγραμμα ικανοποίησης των Έβαλων για την Κομισιόν από την Κομισιόν της Ε.Ε., Υποβληθέντα την 12 Ιουλίου 2013. Δεν Δηλώνεται ο πρόγραμμα ικανοποίησης των Έβαλων για την Κομισιόν από την Κομισιόν της Ε.Ε., Υποβληθέντα την 12 Ιουλίου 2013.

- Δεν Δηλώνεται ο πρόγραμμα ικανοποίησης των Έβαλων για την Κομισιόν από την Κομισιόν της Ε.Ε., Υποβληθέντα την 12 Ιουλίου 2013.
- Δεν Δηλώνεται ο πρόγραμμα ικανοποίησης των Έβαλων για την Κομισιόν από την Κομισιόν της Ε.Ε., Υποβληθέντα την 12 Ιουλίου 2013.
- Δεν Δηλώνεται ο πρόγραμμα ικανοποίησης των Έβαλων για την Κομισιόν από την Κομισιόν της Ε.Ε., Υποβληθέντα την 12 Ιουλίου 2013.
- Δεν Δηλώνεται ο πρόγραμμα ικανοποίησης των Έβαλων για την Κομισιόν από την Κομισιόν της Ε.Ε., Υποβληθέντα την 12 Ιουλίου 2013.
Our ~26 Terabytes: Data Distribution by Web Collection

Web Data by Collection Category (TB)

- Federal Government: 15.408 TB
- Federal Politics: 3.523 TB
- Provincial Gov. and Politics: 2.752 TB
- Commemoration: 0.695 TB
- Special Requests: 0.081 TB
- Olympics/Paralympics: 0.056 TB
- Other: 0.034 TB
Recent Collections
Fort McMurray Wildfires

• An example of our events-based approach
• Building this collection involved targeting and scraping relevant news media twice a day (in French and English) from May to July 2016
  – Refined crawling using seeds of boolean searches for all topical media through Google News and RSS feeds
  – Avoids crawling individual newspapers
  – International media perspectives were also sampled
• Collection thereby documents the events, reactions and analysis as they occurred/were published.
Incendies de forêt

Début du retour des habitants à Fort McMurray

Agence France-Presse | Publié le 1er juin 2016 à 05:49 - Mis à jour le 1er juin 2016 à 17:33
First World War Commemoration

• A long-term project was initiated in January 2015 to document the First World War throughout the centenary period (until 2018)
• Collection goals:
  – Highlight the role of Canadian institutions in capturing and communicating the history of Canadians in the First World War
  – Document regional and national perspectives by collecting the resources of memory institutions; municipal, provincial, and federal government bodies; family and local historical societies; First Nations and Aboriginal groups; military units and veterans' organizations.
  – Aims for balance among institutional, individual, and media perspectives on the war itself and how the war is remembered (193 seeds to date)
• LAC curated a representative collection of 150 web resources and provided this seedlist to the IIPC to reflect Canadian perspectives on the First World War
Rough landing

Royal Flying Corps crew check the condition of a downed Allied biplane circa 1917. Crash landings were one of the many hazards faced by airmen during the war.
Truth and Reconciliation (TRC)

- LAC’s Truth and Reconciliation Collection documents:
  - The web-footprint of the commission itself over time
  - Content that records the effect and impact of the TRC on Canadian society.
  - Organizations connected with the commission either as active partners at National Events or through funding initiatives to support commemoration.
  - The outcomes of the TRC and the legacy of residential schools
  - Resources with a community focus on survivors, commemoration, healing, and reconciliation
  - Blogs and other personal websites that provide perspectives on the Indian Residential School system in Canada

- Seedlist built in collaboration with archivist subject matter experts within LAC that worked on LAC’s component of TRC (218 seeds to date)

- Will form the core of a pan-Canadian web archival collection led by the National Centre for Truth and Reconciliation and supported by the Universities of Winnipeg and Manitoba
Rio 2016 Summer Olympic and Paralympic Games

• Collection documents:
  – Social dimensions/issues surrounding the Games (Zika virus, doping)
  – News media (thousands of articles via news websites and RSS feeds)
  – Federal government involvement
  – Sports organizations and federations; athletic development
  – A total of 1.8 TB of data over the course of the project.

• Supplied our seedlist to the IIPC's Olympics/Paralympics collection
  – IIPC and its partner institutions together preserved 4,817 seeds from 125 countries

• Twitter capture via Twarc (https://github.com/edsu/twarc)
  – Twitter data from several hashtags, including #TeamCanada, #EquipeCanada, #GoCanadaGo, #LetsGoCanada, #Paratough, #Parafort/s was collected to document the discourse of Canadians on the games.
  – Tweet IDs will be made available via Dataverse soon, so that interested researchers can create their own local collections.
Rio 2016: Jour 9

Rio 2016: Jour 9

Rio 2016 : Jour 10

Rio 2016 : Soirée en bronze pour Andre de Grasse

Rio 2016 : Jacqueline Simoneau et Karine Thomas en action

Rio 2016: Jour 8
Treasury Board of Canada Secretariat’s Web Renewal Initiative

Part of:

Government of Canada

Gouvernement du Canada

Home → Open Government → About Open Government → Canada's Action Plan on Open Government

Follow: Twitter RSS Google+ LinkedIn

Canada's Action Plan on Open Government
Context of Web Renewal

- Treasury Board Secretariat’s Web Renewal Initiative seeks to consolidate the GC’s Web presence from ~1,500 domains down to six or less, with the majority migrating to Canada.ca by 2017
  - Initiative linked to Open Government, its Action Plan, Open Government, Virtual Library/”Open Information Portal”

- LAC began collecting in the 2014-15 fiscal year and activities will be ongoing into 2017-2018
Government of Canada Web Archive

Government of Canada Web Archive - websites archived by Library and Archives Canada. Forms, search boxes and external links may not function within this archived page.

[New Search] [View other versions of this page]

 Origins of the RCMP

Sir John A. Macdonald | The N.W.M.P. | The March West | The American Sioux | Settlement | Policing the Railway | Servants of the Public | The North West Rebellion | Free Land | Klondike Gold Rush | North to the Arctic Shores | The Royal Canadian Mounted Police

Historical Notes, Uniform and Dress, Games...
National security and defence

Learn about the security and defence of Canada. Find out about the Canadian Armed Forces and the investments being made to support them. Learn how border, transportation and cyber security help keep Canadians safe and secure.

Topics

- National security
  Discover how we protect the security of Canadians through counter-terrorism, security screenings and the protection of Canada’s critical infrastructure.

- Canadian Armed Forces
  Find out who the Canadian Armed Forces are and what they do. Learn about the various operations in Canada and around the world. Find out about military history, honours and remembrance or find a job in the Forces.
As of September 2014, we have collected:

- ~100 million documents
- 7.7 terabytes of data
- 1,300+ major domains
- 3,300 seeds

~90% of all our federal data (December 2005 to December 2015) is currently available via the Government of Canada Web Archive.
Relaunch of the Government of Canada Web Archive (GCWA)

http://webarchive.bac-lac.gc.ca
http://archivesduweb.bac-lac.gc.ca
Federal Government Web

A

- Advisory Council on Science and Technology
- Agriculture and Agri-Food Canada
- Assisted Human Reproduction Canada
- Association of Professional Executives of the Public Service of Canada
- Atlantic Canada Opportunities Agency
- Atlantic Pilotage Authority Canada
- Atomic Energy of Canada Limited
- Auditor General of Canada, Office of the
Agriculture and Agri-Food Canada

- www.agr.gc.ca
- www.beefexportcanada.gc.ca
- www.cbif.gc.ca
- www.coop.gc.ca
- www.cpma-acpm.gc.ca
- www.dairyinfo.gc.ca
- www.infolait.gc.ca

Prime Minister of Canada, Office of the

- www.pm.gc.ca
- www.premierministre.gc.ca
- www.primeminister.gc.ca
Looking to the Future

• Major updates to LAC’s web archiving technical infrastructure
  – Development and deployment full-text search
  – Providing access to historical collections from the open Canadian web that have been dark-archived to date
    • All non-federal holdings, starting around Q3 of FY 2016-17
    • Approximately 16+ TB of additional web archival holdings including the recent Elections
  – Rebranding the “Canadian Web Archives”
  – Working on the digital preservation of LAC’s web archival holdings (this fiscal)
A Call for Collaboration

• We want to collaborate with our partners in the Canadian research universities
  
  – Collaborative policy, software, collections development
  – Broader web archival collection scopes
  – Cover off more twitter hashtags on projects/subjects of mutual interest
  – Identify and prevent overlaps in effort
  – Exchange technical best practices
404 PAGE NOT FOUND

GOOD
Tom Smyth
Manager, Digital Operations and Program Lead, Web Archiving
tom.smyth@canada.ca
@smythbound

Questions?
Email me!