Harvesting the Government of Canada Web Presence

At
Library and Archives Canada

Tom Smyth
Lead, Digital Ingest Projects
Acquisition and Evaluation Branch
tom.smyth@bac-lac.gc.ca

Legislative Context: Mandate

Library and Archives of Canada Act (S.C. 2004, c. 11)

Preamble

WHEREAS it is necessary that

(a) the documentary heritage of Canada be preserved for the benefit of present and future generations;

(b) Canada be served by an institution that is a source of enduring knowledge accessible to all, contributing to the cultural, social and economic advancement of Canada as a free and democratic society;

(c) that institution facilitate in Canada cooperation among the communities involved in the acquisition, preservation and diffusion of knowledge; and

(d) that institution serve as the continuing memory of the government of Canada and its institutions;

Legislative Context: Authorities for Collection

Multiple authorities exist within the *LAC Act* that empower it to collect government information:

- *Section 10: Legal Deposit*
  - – *Covers all publications that are published in Canada, including those from the GC.*
- *Section 12 & 13: Government and Ministerial Records*
  - – *Covers the disposition, transfer, and right of access to government records.*

Legislative Context:
Harvesting at LAC

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 Harvesting: Background

- LAC began harvesting the Government of Canada web presence in 2005
  - In total LAC has collected it three times (2005, 2006, 2007)
- LAC’s means of making this harvested data accessible is the *Government of Canada Web Archive* (GCWA)
- LAC suspended GC domain harvests in 2008
  - A period of acquisition and resource evaluation policy renewal
  - Harvesting of individual domains or sites has continued by official request:
    - Since July 2008, LAC has processed 132 requests for preservation harvesting of GC websites

Li
Ca

Français

Home > Go

Introduct

Search

Basic Sea

Advanced

Departme

URL List

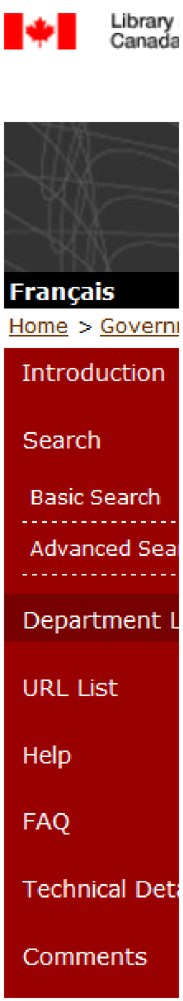
Help

FAQ

Technical

Comment

Date Createc
Date Modifie



Library
Canada

Français

Home > Governi

Introduction

Search

Basic Search

Advanced Sea

Department L

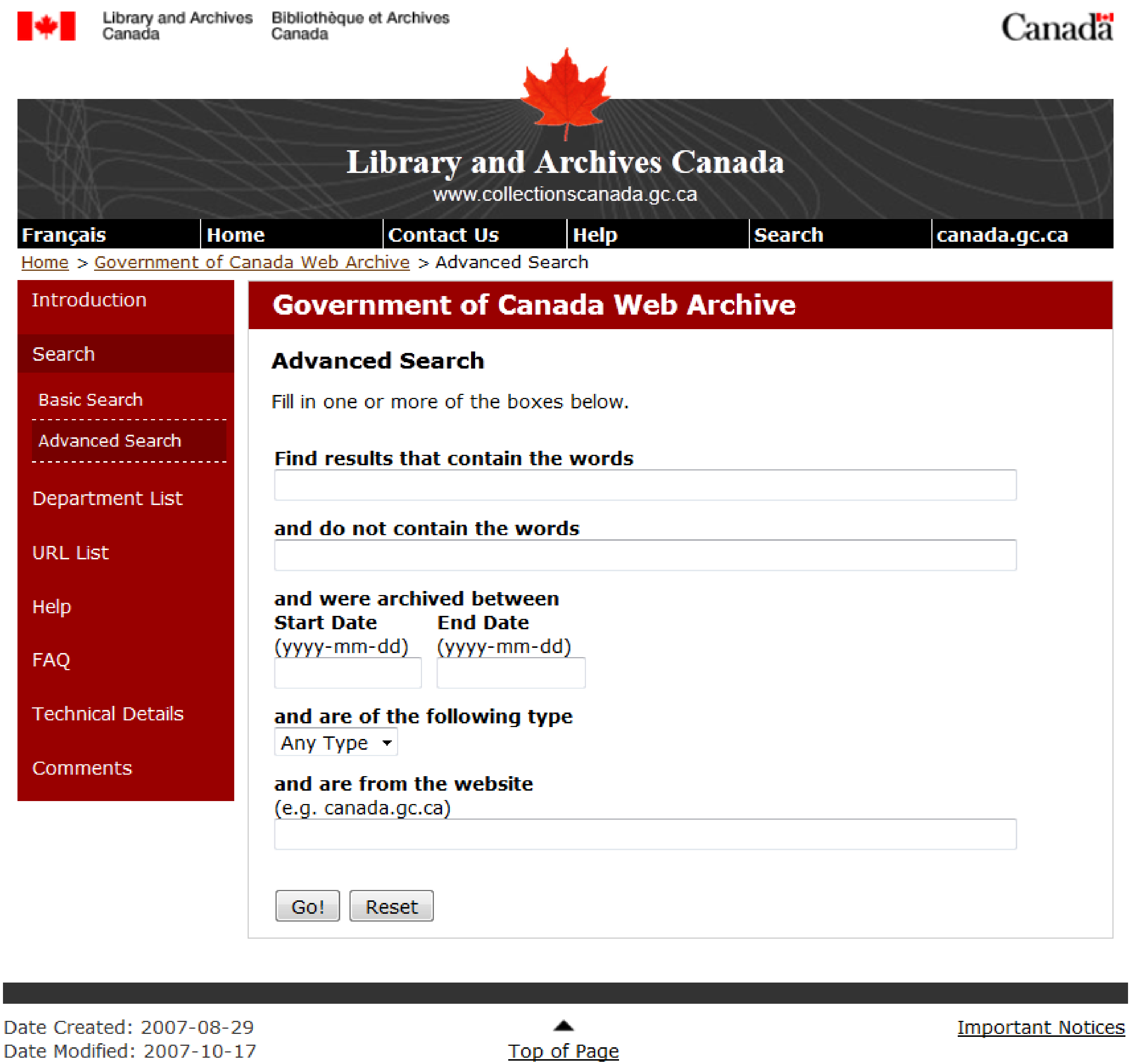
URL List

Help

FAQ

Technical Det:

Comments



Library and Archives Canada
Bibliothèque et Archives Canada

Canada

Library and Archives Canada
www.collectionscanada.gc.ca

Français | Home | Contact Us | Help | Search | canada.gc.ca

Home > Government of Canada Web Archive > Advanced Search

- Introduction
- Search
  - Basic Search
  - Advanced Search
- Department List
- URL List
- Help
- FAQ
- Technical Details
- Comments

Government of Canada Web Archive

Advanced Search

Fill in one or more of the boxes below.

Find results that contain the words

and do not contain the words

and were archived between

Start Date (yyyy-mm-dd) **End Date** (yyyy-mm-dd)

and are of the following type

Any Type

and are from the website
(e.g. canada.gc.ca)

Go! Reset

Date Created: 2007-08-29
Date Modified: 2007-10-17

Top of Page

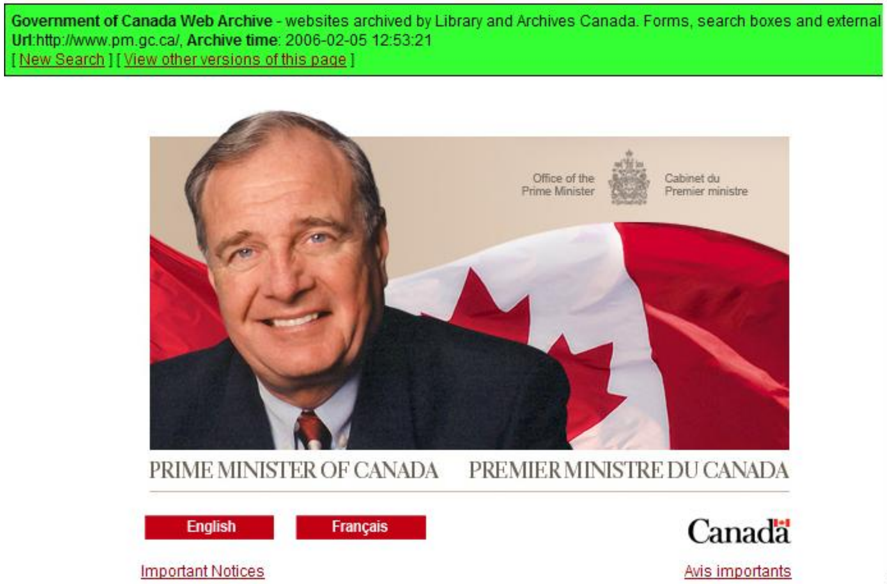
Important Notices

Collection Overview

- LAC’s web archival collection is ~15 TB
  - 7.5+ TB (183 million objects collected in 3 domain crawls) is available for consultation through the GCWA
  - ~8 TB have been collected since 2008, but this data is not yet accessible for technical reasons
    - About half of this data is non-governmental
  - LAC has web archival holdings of the Electronic Depository Services Program Checklists as of 1995:
    - http://epe.lac-bac.gc.ca/100/201/301/liste_hebdomadaire/
    - http://epe.lac-bac.gc.ca/100/201/301/weekly_checklist/

Here today, gone tomorrow pm.gc.ca

2006-02-05



Government of Canada Web Archive - websites archived by Library and Archives Canada. Forms, search boxes and external
Url:http://www.pm.gc.ca/, **Archive time**: 2006-02-05 12:53:21
[New Search] [View other versions of this page]

Office of the Prime Minister

Cabinet du Premier ministre

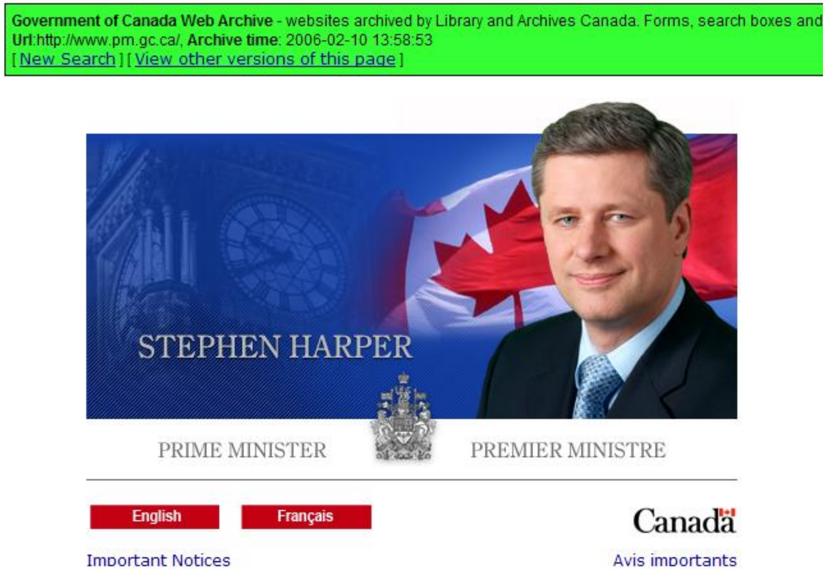
PRIME MINISTER OF CANADA

PREMIER MINISTRE DU CANADA

Important Notices

Avis importants

2006-02-10



Government of Canada Web Archive - websites archived by Library and Archives Canada. Forms, search boxes and
Url:http://www.pm.gc.ca/, **Archive time**: 2006-02-10 13:58:53
[New Search] [View other versions of this page]

STEPHEN HARPER

PRIME MINISTER

PREMIER MINISTRE

Important Notices

Avis importants

Gone today, but preserved @ LAC:
A short listing since 2008

- **Governor General's websites and blogs**
- **PCH: Websites on the various Royal Tours**
- **State funerals: Roméo Leblanc & Jack Layton**
- **The 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**
- **INAC → AANDC**
- **CIDA → DFAIT**
- **HRSDC → ESDC, etc.**

LAC’s Web Harvesting: Current Status

- Since 2008, LAC has curated several thematic web collections for major cultural, political, and historical events
  - These collections include but are not limited to GC web resources
  - Project topics include, for example:
    - Canada’s participation in the Olympic games
      - Beijing 2008, Vancouver 2010 (1.5+ TB), London 2012, Sochi 2014
    - 100th Anniversary of the Calgary Stampede
    - War of 1812 commemoration
    - Keystone Pipeline Development
    - Canadian Arctic Sovereignty
    - Lac Mégantic Rail Disaster

LAC's Web Harvesting: Current Status

- LAC began a 4th crawl of the Government of Canada web domain in Sept 2013:
  - *Official Languages Act*; TBS Directive on Web Accessibility
  - Data collection outsourced to Internet Archive's "Archive-It" service
  - Data will be returned to LAC and made accessible via an upgraded GCWA

2013 GC Domain Harvest: Preliminary Results

- Some GC websites successfully captured as of October 28th 2013:
  - Governor General
  - Prime Minister's Office
  - Privy Council Office (PCO)
  - Treasury Board Secretariat (TBS)
  - Finance
  - Canada Revenue Agency
  - Auditor General
  - Justice and most of the Commissioners and Ombudspersons
  - Parliament (PARLinfo and LEGISinfo)
  - Supreme Court, Federal Court, Federal Court of Appeal
  - StatsCan
  - Public Safety, RCMP
  - DFAIT (and now CIDA)
  - DND and subsidiaries
  - PWGSC, Service, HRSDC
  - Citizenship and Immigration
  - Industry and its subsidiaries
  - Heritage
  - AANDC and the suite of Northern websites
  - EC, DFO, Climate, Weather, Canadian Environmental Assessment Agency
  - Canada.gc.ca
  - Canada Gazette
  - Publications.gc.ca

2013 GC Domain Harvest: Preliminary Results (2)

As of October 28th 2013, we had collected:

- 271 seeds (all major domains, e.g., gg.ca; pm.gc.ca)
- 11,536,589 digital objects
- 879 gigabytes of data

Expected completion of project:
Early November 2013

Next Steps

- LAC is currently defining its long term business strategy and technical requirements for a renewed Web Harvesting Program
- LAC's web harvesting infrastructure will be updated to modern (IIPC) standards
- The GCWA will be updated to provide access to all of LAC's web archival holdings (~15 TB)
  - Migration of legacy ARCs to ISO standard WARCs
  - GCWA will be migrated to the "blue" WCAG compliant GC template

Next Steps

- Complete the project and communicate to clients, stakeholders and the public (once sites are added to the public Archive)
- In consultation with stakeholders, departments and central agencies:
  - Develop a long-term, robust GC web harvesting strategy
  - Align go-forward strategy with GC-wide initiatives including Web Renewal, Open Government, and other IM initiatives as they arise.

Web Harvesting Team @ LAC

Susan Haigh
Director, Acquisitions
susan.haigh@bac-lac.gc.ca

Tom Smyth
Lead, Digital Ingest Projects
tom.smyth@bac-lac.gc.ca

Patricia Klambauer
Web Harvesting Technician
web-archives-web@bac-lac.gc.ca

Library and Archives Canada

550 de la Cité Boulevard
Gatineau, Quebec
K1A 0N4
Canada

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

Fax: 613-995-6274

www.bac-lac.gc.ca
www.collectionscanada.gc.ca