Web Harvesting the Government of Canada sites

A Collaborative Effort

Sam-chin Li
Tom Smyth
Amanda Wakaruk
Web sites are collected via software that downloads code, images, documents, and others files essential to completely and faithfully reproduce the web site at the time of capture.

International Internet Preservation consortium
The web is often the only source for government information.

Web archives keep important records of the government and ensure democracy for now and for the future.
March 2013

- Reduce Redundant, Outdated and Trivial Content (ROT Criteria)
- GoC Web Convergence Plan
- UofT starts to archive GoC web content
A URL appearing in a seed list as one of the starting addresses a web crawler uses to capture content. Also called a targeted URL.

Archive-IT glossary
Scope and Frequency

- Government of Canada Departments, Agencies, Crown Corporations, Special Operating Agencies and various affiliated organizations.
- One time only
- Daily and Semiannually
- Separate seeds for important resources
Departments and Agencies

This page provides an alphabetical listing of links to Government of Canada Departments, Agencies, Crown Corporations, Special Operating Agencies and various affiliated organizations.

If the name of the organization you are looking for is not listed, try consulting the Financial Administration Act. It contains information on all current Government of Canada departments, agencies, Crown corporations and special operating agencies, and also includes a listing of organizations that no longer exist or that have been privatized.

A

- Aboriginal Affairs and Northern Development Canada
  - First Nations Statistical Institute
  - Indian Oil and Gas Canada
  - Indian Residential Schools
- Agriculture and Agri-Food Canada
  - Canadian Biodiversity Information Facility
  - Canadian Pari-Mutuel Agency
- Assisted Human Reproduction Canada
- Atlantic Canada Opportunities Agency
We started too late!

Many contents have already disappeared from the Web!

March 2013
Departments and Agencies

This page provides an alphabetical listing of links to Government of Canada Departments, Agencies, Crown Corporations, Special Operating Agencies and various affiliated organizations.

If the name of the organization you are looking for is not listed, try consulting the Financial Administration Act. It contains information on all current Government of Canada departments, agencies, Crown corporations and special operating agencies, and also includes a listing of organizations that no longer exist or that have been privatized.

A

- Aboriginal Affairs and Northern Development Canada
  - First Nations Statistical Institute
  - Indian Oil and Gas Canada
  - Indian Residential Schools
- Agriculture and Agri-Food Canada
  - Canadian Biodiversity Information Facility
  - Canadian Pari-Mutuel Agency
- Assisted Human Reproduction Canada
- Atlantic Canada Opportunities Agency
- Atlantic Pilotage Authority
- Atlas of Canada - Natural Resources Canada
- Atomic Energy of Canada Limited
- Auditor General of Canada, Office of

Solution:
Wayback Machine
June, 2013

- Checked the gaps in the Wayback Machine GoC Harvest
- Crawled the gaps
- Purchased the content
- Moved it to UofT Archive-IT account
Purchased Wayback Machine Harvest from December 2007 – March 2013
<table>
<thead>
<tr>
<th></th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aug 9, 2008</td>
<td>Mar 25, 2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
HEROIN AND OTHER DANGEROUS DRUGS ARE BANNED FROM HEALTH CANADA'S SPECIAL ACCESS PROGRAMME - GOVERNMENT OF CANADA PUTS SAFETY AND SECURITY OF CANADIANS FIRST AND FOCUSES ON TREATMENT AND RECOVERY

Joined by doctors, health experts, Canadians who have battled with addiction and all other stakeholders, Federal Minister of Health, Leona Aglukkaq announced today that the government has made an important decision to ban heroin and other dangerous drugs from the Special Access Programme (SAP).
• Select sites
• Crawl test to identify the scope
• Schedule the frequency
• Harvesting
• Quality Assurance
• Patch crawls
• Create metadata
Seed Selection & Quality Assurance

How can we make sure the important content is archived in a such short timeframe?
URLs Recommended by subject specialists:
• 46

Libraries Volunteered their time:
• Industrial Relations & Human Resources
• OISE

Library community involvement
Seeds

- Captured: 146
- Transferred from Global Wayback: 418

http://www.archive-it.org/collections/3608

University of Toronto Archive-IT
• Sam-chin Li, Government Information Librarian
• Don McLeod, Archive-IT account coordinator at UofT
• Nicholas Worby, Graduate Student Library Assistant
“Announcing the first ever Archive-It US Government Shutdown Notice Awards!”

1996: Internet Archive founded and starts web harvesting activities
2001: Wayback Machine launches
2002: First release of Heritrix (web crawler)
2003: International Internet Preservation Consortium (IIPC)
2006: Archive-It, one of the first web archiving services
2009: WARC file, format generated by Heritrix, becomes an ISO standard
2009: University of Alberta Libraries (UAL) starts web harvesting with Archive-IT

Adapted from Scott Reed’s slides: Wakaruk, Lau, Reed, “Community Tools and Best Practices for Harvesting and Preserving At-Risk Web Content.” Association of Canadian Archivists, June 2013
http://hdl.handle.net/10402/era.32072
Current Status (October 2013)

- 13 collections, > 67 million documents crawled (3.6 TB), ~600 active seeds, >1700 websites: ~50-60 are gc.ca
- Government web content distributed across multiple collections: e.g., Energy & Environment, Health Sciences Grey Literature, Idle No More, Circumpolar, Government Information Collection (~133 seeds, mostly Alberta)
UAL Collection Development Committee
Born Digital Working Group (2009-2014)

- Leads the Libraries’ direction related to born digital collections.
- Members: AUL, Collections Manager; Librarians: Digital Initiatives, Digital Repository Services, Metadata, Govt Info, Circumpolar, Business, Humanities and Social Sciences

Policies and Procedures

- Policy and Procedures for Born Digital Collections
- Metadata and Cataloguing Policy
- Web Archiving Guidelines (2): Selectors and Collection Managers
- Web Archiving Policy; Web Archiving Agreement
**UAL Web Archiving Policy**

(excerpts)

**ACCESS:** contents will be made publicly available

**OWNERSHIP:** remains with the website owner

**AUTHORIZATION:** we cannot authorize re-use

**TAKE-DOWN:** we will evaluate requests from content owners to remove content

**LIABILITY:** we will not be held liable for actions of website owners

[http://www.library.ualberta.ca/webarchive/](http://www.library.ualberta.ca/webarchive/)
Metadata Needs Assessed on Collection Basis

Mixed model of metadata:
- Traditional document level: full, brief, use warrant access points
- Non-traditional: AI text search/widgets, AI metadata

Ongoing:
- Exploring options for metadata creation (e.g., copy, outsourcing)
- Investigating discovery access options (e.g., EDS, AI Portal, others?)
- Developing use case studies (sub-group of BDWG)

Access = Metadata
Mis-à-jour Giant [electronic resource] : Projet d'assainissement de la mine Giant
Projet d'assainissement de la mine Giant (Canada)

Corporate Author: Projet d'assainissement de la mine Giant (Canada)
Title: Mis-à-jour Giant [electronic resource] : Projet d'assainissement de la mine Giant.

Electronic access: Free Access
Publication info: [Yellowknife] : Indian and Northern Affairs Canada, [2008]
Physical description: 1 online resource.
General Note: Distributed by the Government of Canada Depository Services Program.
General Note: "QS-Y289-005-FF-A1".
General Note: Also published under the title: The Giant update, Giant Mine Remediation Project.
Added Entry-Corporat: Canada. Indian and Northern Affairs Canada.
Added Entry-Corporat: Northwest Territories.
Govt Canada class#: R76-1/2007F-PDF
key: 6062544
Equine technician [electronic resource]
Alberta. Alberta Education.

Corporate Author: Alberta. Alberta Education.
Title: Equine technician [electronic resource].
Electronic access: Free Access
Physical description: 1 electronic text (3 p.)
Series Added Entry- Alberta curriculum guides.
Uniform title- Running title: Green Certificate Program (Senior High)
Subject term: Horses--Study and teaching (Secondary)--Alberta.
Subject term: Horsemen and horsewomen--Alberta.
Technical Details: Mode of access: Internet.
Contents Note: Equine operations and care 33 -- Equine processes and practices 33 -- Equine support systems 33.
Series Statement: [(Alberta curriculum guides)]
key: 5360314

Page 1 of 1 (1 Total Results)
Title: Equine Technician
URL: http://education.alberta.ca/media/1106757/equine.pdf
Captured 12 times between Oct 19, 2010 and Apr 26, 2013
Subject: equine technician, programs of study, program of studies, curriculum guide, senior high school, grade 10, grade 11, grade 12, division 4, division four, division IV, green certificate program
Language: English
Date: 2009
Rights: This material is provided under educational reproduction permissions included in Alberta Education's Copyright and Disclosure Statement, see terms at http://education.alberta.ca/using/copyright.aspx. This Statement requires the following identification: "The source of the material is [Alberta Education] http://www.education.alberta.ca/. The use of these materials by the end user is done without any affiliation with or endorsement by the Government of Alberta. Reliance upon the end user's use of these materials is at the risk of the end user."
March, 2013

- About 400 seeds
- Content disappearing on a daily basis
- Deadline: July 31, 2013

Mission Impossible!
Collaboration

UofA and UofT
COPPUL members
LAC
<table>
<thead>
<tr>
<th></th>
<th>Departments/Agencies</th>
<th>URLs tried</th>
<th>URL 1</th>
<th>Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td><strong>Aboriginal Affairs and Northern Development Canada</strong></td>
<td><a href="http://www.aadnc-aandc.gc.ca/">http://www.aadnc-aandc.gc.ca/</a></td>
<td></td>
<td>U of A</td>
</tr>
<tr>
<td>3</td>
<td><strong>British Columbia's Land Claims Agreement</strong></td>
<td><a href="http://www.ainc-inac.gc.ca/al/lcd/ccl/fagr/bc-eng.asp">http://www.ainc-inac.gc.ca/al/lcd/ccl/fagr/bc-eng.asp</a></td>
<td></td>
<td>U of A</td>
</tr>
<tr>
<td>4</td>
<td><strong>Canada's Northern Strategy</strong></td>
<td><a href="http://www.northernstrategy.gc.ca/">http://www.northernstrategy.gc.ca/</a></td>
<td></td>
<td>U of A</td>
</tr>
<tr>
<td>5</td>
<td><strong>First Nations Statistical Institute</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td><strong>Indian Oil and Gas Canada</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td><strong>Indian Residential Schools</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td><strong>Labrador Inuit Land Claims Agreement</strong></td>
<td><a href="http://www.ainc-inac.gc.ca/al/lcd/ccl/fagr/lrl-eng.asp">http://www.ainc-inac.gc.ca/al/lcd/ccl/fagr/lrl-eng.asp</a></td>
<td></td>
<td>U of A</td>
</tr>
<tr>
<td>9</td>
<td><strong>Nunavut Land Claims Agreement</strong></td>
<td><a href="http://www.ainc-inac.gc.ca/al/lcd/ccl/fagr/nl-eng.asp">http://www.ainc-inac.gc.ca/al/lcd/ccl/fagr/nl-eng.asp</a></td>
<td></td>
<td>U of A</td>
</tr>
<tr>
<td>10</td>
<td><strong>Northwest Territories Land Claims Agreement</strong></td>
<td><a href="http://www.ainc-inac.gc.ca/al/lcd/ccl/fagr/wt-eng.asp">http://www.ainc-inac.gc.ca/al/lcd/ccl/fagr/wt-eng.asp</a></td>
<td></td>
<td>U of A</td>
</tr>
<tr>
<td>11</td>
<td><strong>Quebec Land Claims Agreement</strong></td>
<td><a href="http://www.ainc-inac.gc.ca/al/lcd/ccl/fagr/qc-eng.asp">http://www.ainc-inac.gc.ca/al/lcd/ccl/fagr/qc-eng.asp</a></td>
<td></td>
<td>U of A</td>
</tr>
<tr>
<td>13</td>
<td><strong>Yukon Land Claims Agreement</strong></td>
<td><a href="http://www.ainc-inac.gc.ca/al/lcd/ccl/fagr/yk-eng.asp">http://www.ainc-inac.gc.ca/al/lcd/ccl/fagr/yk-eng.asp</a></td>
<td></td>
<td>U of A</td>
</tr>
<tr>
<td>14</td>
<td><strong>Aboriginal Portal of Canada</strong></td>
<td><a href="http://www.aboriginalcanada.gc.ca">http://www.aboriginalcanada.gc.ca</a></td>
<td></td>
<td>U of A</td>
</tr>
<tr>
<td>15</td>
<td><strong>pdfs</strong></td>
<td><a href="https://www.google.ca/search?as_q=&amp;as_eq=&amp;as_oq=&amp;as_eq=&amp;as_n10=1">https://www.google.ca/search?as_q=&amp;as_eq=&amp;as_oq=&amp;as_eq=&amp;as_n10=1</a></td>
<td></td>
<td>U of A</td>
</tr>
<tr>
<td>16</td>
<td><strong>Agriculture and Agr-Food Canada</strong></td>
<td><a href="http://agr.gc.ca">http://agr.gc.ca</a></td>
<td></td>
<td>UoF</td>
</tr>
<tr>
<td>17</td>
<td><strong>Canadian Biodiversity Information Facility</strong></td>
<td><a href="http://www.cbif.gc.ca/">http://www.cbif.gc.ca/</a></td>
<td></td>
<td>UoF</td>
</tr>
<tr>
<td>18</td>
<td><strong>Canadian Pari-Mutuel Agency</strong></td>
<td></td>
<td></td>
<td>UoF</td>
</tr>
<tr>
<td>19</td>
<td><strong>Assisted Human Reproduction Canada</strong></td>
<td></td>
<td></td>
<td>UoF</td>
</tr>
<tr>
<td>20</td>
<td><strong>Atlantic Canada Opportunities Agency</strong></td>
<td></td>
<td></td>
<td>UoF</td>
</tr>
<tr>
<td>21</td>
<td><strong>Atlantic Pilotage Authority</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td><strong>Atlas of Canada - Natural Resources Canada</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**March, 2013**

**Seed List #1: UofA and UoF**
What does yellow mean?

Yellow highlighting was used by U of T and U of A to indicate institutions that deemed critical to capture and preserve

<table>
<thead>
<tr>
<th>No.</th>
<th>Department/Agency</th>
<th>Library</th>
<th>URL 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>169</td>
<td>Registry of the Specific Claims Tribunal of Canada</td>
<td>UVIC</td>
<td><a href="http://www.sst-tp.ca/hist/hist-e.htm">http://www.sst-tp.ca/hist/hist-e.htm</a></td>
</tr>
<tr>
<td>171</td>
<td>Species at Risk Public Registry - Environment Canada</td>
<td>UVIC</td>
<td><a href="http://www.sararegistry.gc.ca/default_e.cfm">http://www.sararegistry.gc.ca/default_e.cfm</a></td>
</tr>
<tr>
<td>172</td>
<td>Staff of the Non-Public Funds, Canadian Forces - National Defence and the Canadian Forces</td>
<td>UVIC</td>
<td><a href="https://www.cfpsa.com/en/Pages/default.aspx">https://www.cfpsa.com/en/Pages/default.aspx</a></td>
</tr>
<tr>
<td>173</td>
<td>Statistical Survey Operations - Statistics Canada</td>
<td>UVIC</td>
<td></td>
</tr>
<tr>
<td>174</td>
<td>Veterans Affairs Canada</td>
<td>UVIC</td>
<td></td>
</tr>
<tr>
<td>175</td>
<td>Veterans Review and Appeal Board Canada</td>
<td>UVIC</td>
<td><a href="http://www.vrab-tacra.gc.ca/Home-accueil-eng.cfm">http://www.vrab-tacra.gc.ca/Home-accueil-eng.cfm</a></td>
</tr>
<tr>
<td>176</td>
<td>Pacific Fisheries Resource Conservation Council</td>
<td>UVIC</td>
<td><a href="http://www.fish.bc.ca/">http://www.fish.bc.ca/</a></td>
</tr>
<tr>
<td>177</td>
<td>Canada Research Chairs - Industry Canada</td>
<td>UVIC/U of T</td>
<td>Part of the Industry Canada crawl done by U of T</td>
</tr>
</tbody>
</table>

**How to use this spreadsheet:**

Please enter your institution's name in Column B next to the agencies that you will *intend* to harvest with Archive-IT. Many of us do not have final approval to purchase ArchiveIT so this is just our first draft.

Please locate the main URL for the sites that you intend to harvest and add to URL 1 column. Use the notes field as needed. At this time the file sizes of most of these agencies isn't known.

This checklist was adapted from the UofT and UofA Archive-IT working document, please see the following URL for the updated crawl list

https://docs.google.com/spreadsheet/ccc?key=0AvIhesVQk7rrbdHktUuM2UTRRYkdaTnVhEhpYJixMVE#gid=0

May, 2013

Spreadsheet #2: UofA, UofT, some COPPUL members
March, 2013
• Sent letters to LAC for meeting

May, 2013
• Teleconference (UofA, UofT and LAC)
• Lobbied for final harvesting of the entire GoC domain before the proposed GC web convergence plan is implemented
• Lobbied for future harvesting of the GoC domain
• Share and maintain seed list

July, 2013
• Questions for LAC about recommencement of web harvesting gc.ca domain

August, 2013
• Teleconference (UofA, UofT and LAC)
• LAC confirmed the harvesting of the GoC domain

November, 2013
• Presentation at the Gov Info Day (UofT, UofA, LAC)
Government Information Day Roundtable

We’d love to hear your suggestions

Next Steps