“Why is this link dead? Aren’t government publications all online?”

Preserving digital federal content with the Canadian Government Information Private LOCKSS Network (CGI-PLN)

Mark Jordan, Simon Fraser University
Amanda Wakaruk, University of Alberta
Access Conference
September 26, 2013
Background

Government Records and Publications

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

- “record” means any documentary material other than a publication, regardless of medium or form
- “government record” means a record that is under the control of a government institution
- “publication” means any library matter that is made available in multiple copies or at multiple locations, whether without charge or otherwise, to the public generally or to qualifying members of the public by subscription or otherwise. Publications may be made available through any medium and may be in any form, including printed material, on-line items or recordings

(Section 7) The objects of the Library and Archives of Canada are

(c) to be the permanent repository of publications of the Government of Canada and of government and ministerial records that are of historical or archival value;

(d) to facilitate the management of information by government institutions;
The DSP does not collect publications from all federal government agencies, nor does it collect Statistics Canada publications or items that fall outside the definition of “publication” which includes news releases and backgrounders.

Recent Timeline of Cdn Govt Pubs

2008
- Common Look and Feel Internet Standard (late 1990s, lost pdfs, databases)
- Library and Archives Canada (LAC) stops web archiving programs

2010
- Crown Copyright Licensing Program allows non-commercial reuse of pubs
- Cessation of print parliamentary committee documents (no digital preservation plan)

2010 - December 2012
- Budget reduction (e.g., CISTI 70%) and closure of numerous federal libraries
- Virtual Library announced as part of Canada’s Action Plan on Open Government

April 2012 ([http://freegovinfo.info/node/3677](http://freegovinfo.info/node/3677))
- Depository Services Program Deficit Reduction Action Plan
- Library and Archives Canada loses major funding, 450 jobs affected (215 lost)
Recent Timeline of Cdn Govt Pubs

May 2012
● CLA Government Information Network meeting: call for interest in CGI-PLN

September 2012
● Cessation of the distribution of print parliamentary publications (no digital preservation plan)

November 2012
● President of the Treasury Board Secretariat announces Web Renewal Action Plan

November 2012 - present
● GC.CA web content disappears (see CLA GIN blog for partial list)

January 2013
● confirmation that goal is to have no/limited web content older than 2-3 years old on federal government web sites
● ALA IDTF resolution asking TBS to harvest gc.ca domain before removing content


CLA GIN Blog: http://agiig.wordpress.com/
Recent Timeline of Cdn Govt Pubs

March 2013 ([http://freegovinfo.info/node/3893](http://freegovinfo.info/node/3893))
- British Columbia Freedom of Information and Privacy Association releases Web Renewal Action Plan (obtained via Freedom of Information legislation); ROT criteria (no information about offline archiving or access)
- Wayback Machine crawl (Internet Archive pro bono)

June 2013
- Revised Communications Policy and Publishing Procedures (TBS)

August 2013
- DSP e-archive of over 111,000 pdfs added to the CGI-PLN LOCKSS boxes; ingested via Archive-IT so content is also available to the public
- confirmation that Virtual Library will only point to existing content on author agency websites (i.e., it will *not* be a repository) – TBS internal consultation ongoing
- LAC confirms that they will resume web harvesting program using Archive-IT but will not make content available to the public immediately

Web Renewal Action Plan

ROT Criteria
LOCKSS Program

LOCKSS = Lots of Copies Keeps Stuff Safe
- tamper evident, distributed preservation system
- based at Stanford; http://www.lockss.org/

Five Principles
1. libraries have local control of assets
2. perpetual access is guaranteed
3. preserve original version
4. decentralized, distributed preservation
5. affordable

LOCKSS for US Documents
- replicates Federal Depository Library Program in the digital environment
- 36 libraries and GPO participating (at least two Canadian partners)
Canadian Government Information
Private LOCKSS Network (CGI-PLN)
http://plnwiki.lockss.org/wiki/index.php/CGI_network

Policies
● Governance
● Collection and Access

Committees
● Steering
● Technical
Technical Overview

- Member requirements
- Overview of preservation architectures
- Metadata
- Disaster Recovery Plan
Member Technical Requirements

- LOCKSS box
  - 2 TB storage (for now)
- Ongoing technical administration
  - Enabling new content
- Technical Subcommittee
Preservation Architectures: Criteria

- Access front end
- Access / proxy integration
- LOCKSS plugin exists
- Pros
- Cons
- Sustainability
- Unknowns
<table>
<thead>
<tr>
<th>Name</th>
<th>Archive type</th>
<th>Access</th>
<th>LOCKSS ready</th>
<th>Sustainability</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archive-It</td>
<td>Dark</td>
<td>Archive-It front end</td>
<td>Archive-It plugin available</td>
<td>Financial model needs to be developed; ready for new collections / content</td>
<td>Cost of Archive-It account</td>
</tr>
<tr>
<td>Central portal</td>
<td>Bright</td>
<td>APLIC portal</td>
<td>Some work required</td>
<td>Technical sustainability problematic; costly in staff</td>
<td>APLIC portal does not support link resolvers; who creates metadata?</td>
</tr>
<tr>
<td>Local hosting</td>
<td>Bright</td>
<td>CONTENTdm, etc. front end</td>
<td>CONTENTdm plugin available</td>
<td>Financial model needs to be developed</td>
<td>Who hosts? Who pays? Who creates metadata?</td>
</tr>
<tr>
<td>DSP direct</td>
<td>Dark</td>
<td>DSP interface</td>
<td>No plugin available</td>
<td>Technical sustainability problematic; costly in staff</td>
<td>Content is still on gov’t servers</td>
</tr>
</tbody>
</table>
**Metadata**

- Depository Services Program Catalogue files, quarterly updates
- MARC records for local ILS/discovery layer loads
  - Links to both the DSP original and the Archive-It URL
- Linking preserved document and metadata
Disaster Recovery

No plan so far. But we have all the WARCs.
Over 111k PDFs at this point
Moving Forward

● other collections?
  ○ provincial? municipal? IGO?
  ○ LAC WARC files of GoC web harvests?
  ○ coordination of web harvesting (Government Information Day, November 1, Toronto)

● new members (contact amanda.wakaruk@ualberta.ca)
Questions?

Mark Jordan, Head of Library Systems, Simon Fraser University
mjordan@sfu.ca @jordanheit

Amanda Wakaruk, Government Information Librarian, University of Alberta Libraries
amanda.wakaruk@ualberta.ca @awakaruk
https://sites.google.com/a/ualberta.ca/wakaruk/