"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

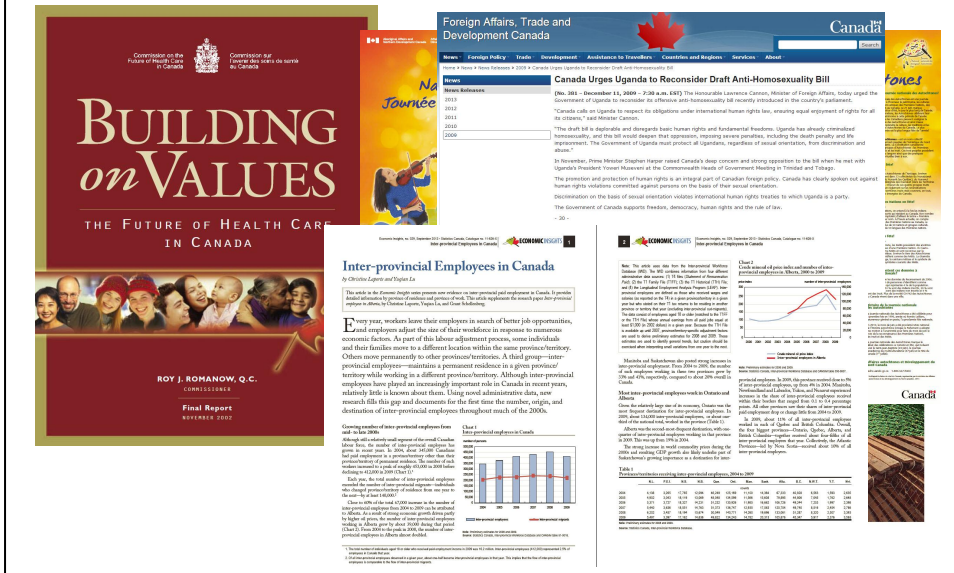## Government Records and Publications

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;

Government Publications:
Depository Services Program (DSP)

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.

See TBS Procedures for Publishing for more information: http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?section=text&id=27167

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)

- 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

Speaking Notes for the Honourable Tony Clement, President of the Treasury Board of Canada - "Using Technology to Challenge the Status Quo in Government Operations" - references the Web Renewal Action Plan: http://www.tbs-sct.gc.ca/media/ps-dp/2012/1106a-eng.asp

CLA GIN Blog: http://agiig.wordpress.com/

Recent Timeline of Cdn Govt Pubs

March 2013 (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
http://fipa.bc.ca/library/Government%20Documents/GoC_web_plan_Part1.pdf
http://fipa.bc.ca/library/Government%20Documents/GoC_web_plan_Part2.pdf

ROT Criteria
http://www.tbs-sct.gc.ca/ws-nw/wu-fe/rot-rid/index-eng.asp

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
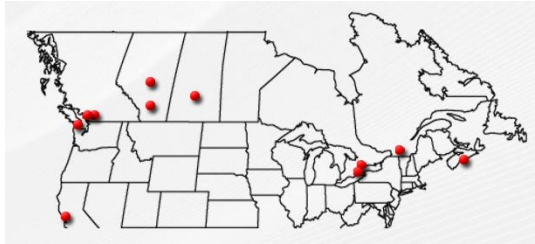# Private LOCKSS Network (CGI-PLN)
http://plnwiki.lockss.org/wiki/index.php/CGI_network

Policies
- Governance
- Collection and Access

Committees
- Steering
- Technical

Technical Overview

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

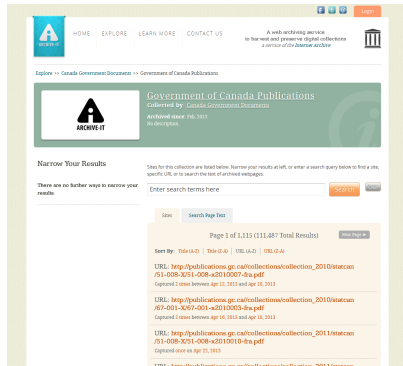# Member Technical Requirements

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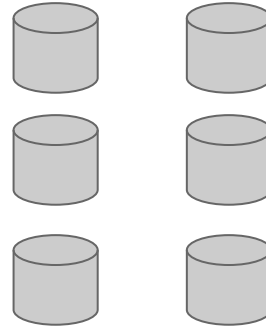
Preservation Architectures: Criteria

- Access front end
- Access / proxy integration
- LOCKSS plugin exists
- Pros
- Cons
- Sustainability
- Unknowns

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

Archive-It

LOCKSS PLN

WARCs

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.

Journal Title: Archive-It Collections
Plugin: Archive-It Plugin
Access Type: OpenAccess
Content Size: 132,101,778,647
Disk Usage (MB): 125985.1
Repository: /cache0/gamma/cache/b/
Status: 100.00% Agreement
Publisher: Canadian Government Information Private LOCKSS Network
Available From Publisher: Yes
Created: 10:43:57 07/31/13
Crawl Pool: org.lockss.plugin.archiveit.ArchiveItPlugin
Last Completed Crawl: 01:03:41 09/12/13
Last Crawl: 01:03:32 09/12/13
Last Crawl Result: Successful
Last Completed Poll: 03:09:36 08/02/13
Last Poll: 01:10:50 08/01/13
Last Poll Result: Complete
AU configuration
Repair candidates
List: URLs, Files

Node Url	Version	Size	Tree Size	Children
lockssau	-	-	132101778647	1
https://partner.archive-it.org	-	-	132101778647	2
https://partner.archive-it.org/cgi-bin	-	-	132101776469	2
https://partner.archive-it.org/cgi-bin/getarcs.pl	-	-	132101733486	139
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-ANNUAL-16495-20130611180035117-00000-wbgrp-crawl066.us.archive.org-6441.warc.gz	1	24977	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-ANNUAL-8507-20130322204108582-00000-wbgrp-crawl057.us.archive.org-6440.warc.gz	1	1793262	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-19302-20130605213057460-00000-wbgrp-crawl067.us.archive.org-6441.warc.gz	1	1015868466	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-19302-20130605234535820-00001-wbgrp-crawl067.us.archive.org-6441.warc.gz	1	1014987164	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-19302-20130606015441449-00002-wbgrp-crawl067.us.archive.org-6441.warc.gz	1	1010677585	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-19302-20130606041526730-00003-wbgrp-crawl067.us.archive.org-6441.warc.gz	1	483698751	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-4108-20130314225806766-00000-wbgrp-crawl058.us.archive.org-6442.warc.gz	1	1016104577	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-4108-20130315015001493-00001-wbgrp-crawl058.us.archive.org-6442.warc.gz	1	1009227257	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-4108-20130315031143958-00002-wbgrp-crawl058.us.archive.org-6442.warc.gz	1	1002921913	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-4108-20130315044140139-00003-wbgrp-crawl058.us.archive.org-6442.warc.gz	1	1150312648	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-4108-20130315055053487-00004-wbgrp-crawl058.us.archive.org-6442.warc.gz	1	1008237902	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-4108-20130315072648732-00005-wbgrp-crawl058.us.archive.org-6442.warc.gz	1	1003660153	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-4108-20130315100559226-00006-wbgrp-crawl058.us.archive.org-6442.warc.gz	1	1000135134	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-4108-20130315113334351-00007-wbgrp-crawl058.us.archive.org-6442.warc.gz	1	1001118189	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-4108-20130315130941588-00008-wbgrp-crawl058.us.archive.org-6442.warc.gz	1	1000051856	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-4108-20130315142531706-00009-wbgrp-crawl058.us.archive.org-6442.warc.gz	1	1009493504	-	-
https://partner.archive-it.org/cgi-bin/getarcs.pl/ARCHIVEIT-3572-NONE-4108-20130315161654600-00010-wbgrp-crawl058.us.archive.org-6442.warc.gz	1	1000208677	-	-

Over 111k PDFs at this point

WARC/1.0
WARC-Type: response
WARC-Target-URI: http://publications.gc.ca/collections/collection_2013/rncan-nrcan/M114-32-2005-eng.pdf
WARC-Date: 2013-06-05T21:31:08Z
WARC-Payload-Digest: sha1:ZEXFKMKRYEIQJX6MKYZ4Y6MRLHVNQYYT
WARC-IP-Address: 205.193.152.47
WARC-Record-ID: <urn:uuid:85f81067-fec1-45dd-ad8e-ef0ad7ac4ef3>
Content-Type: application/http; msgtype=response
Content-Length: 341516

HTTP/1.1 200 OK
Date: Wed, 05 Jun 2013 21:31:08 GMT
Server: Apache/2.2.3 (Linux/SUSE)
Last-Modified: Fri, 18 Jan 2013 14:08:33 GMT
ETag: "b74028-534ff-4d390a58bda40"
Accept-Ranges: bytes
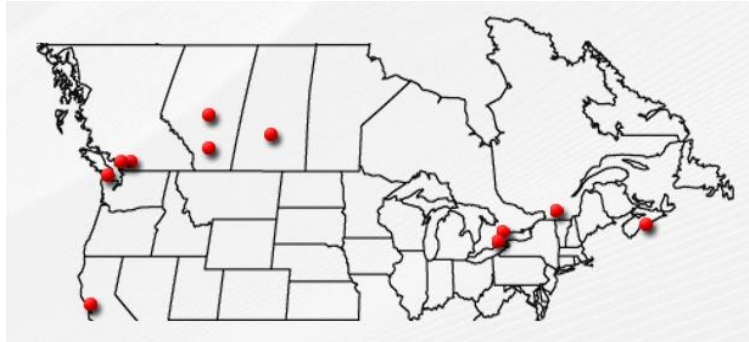Content-Length: 341247
Connection: close
Content-Type: application/pdf

        Linearized 1/L 341247/O 922/E 80152/N 21/T 322839/H [527 471]>>
xref
918 11
0000000016 00000 n
0000001197 00000 n
0000000527 00000 n
0000001490 00000 n
0000001631 00000 n
0000001873 00000 n
0000001909 00000 n
0000002084 00000 n
0000002161 00000 n
0000004831 00000 n
0000000998 00000 n

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)

Canadian Government Information Private LOCKSS Network

University of Victoria
University of British Columbia
Simon Fraser University
Stanford University
University of Alberta
University of Calgary

University of Saskatchewan
University of Toronto / Scholars Portal
McGill University
Dalhousie University

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/