A brief overview of the Canadian Government Information ~~Private LOCKSS~~ Digital Preservation Network

About Us

Carla Graebner / SFU Research Data Services & Government Information Librarian

Nicholas Worby / UofT Government Information and Statistics Librarian

About CGI-DPN

- Canadian Government Information Digital Preservation Network
- Started October 2012; 11 institutions
- The mission of the CGI DPN is to preserve digital collections of government information.
    - Government of Canada Publications Catalogue
    - Pre-election Government of Canada and candidate websites
    - Provincial, thematic, event specific collections
- All Time Total Data: 4.6 TB
- All Time Total Docs: 57,976,853

What web archives are …

"**Web archiving** is the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use."

International Internet Preservation Consortium, www.netpreserve.org

What web archives are not ...

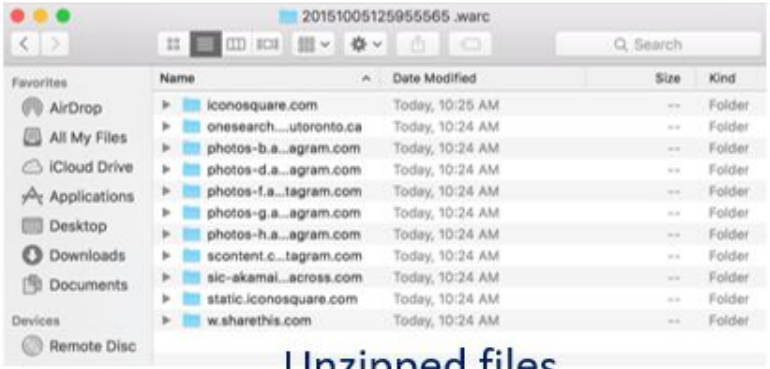
- Not exact replicas of what was on the web
- Not fully functional (e.g. user interaction, dynamic features)
- Not objective or fully automated (humans make decisions)
- Not complete

What's in a WARC?

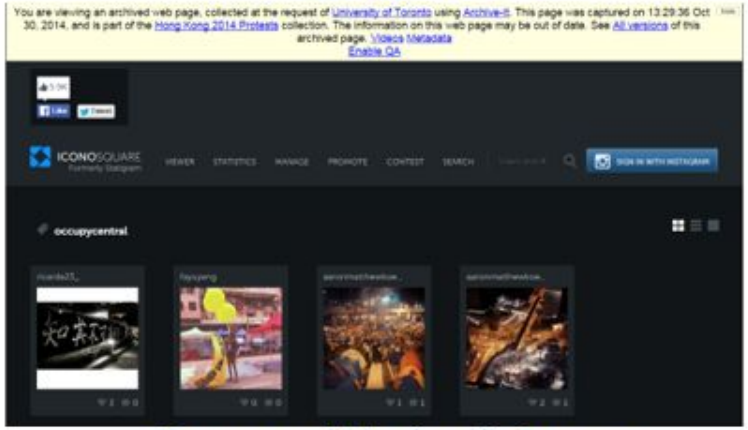
- ISO standard format (28500:2009)
- Container for all content captured in a crawl (e.g. HTML, CSS, .JS, .PDF, .JPG)
- Playable in a browser via the Wayback Machine or Open Wayback
- Includes crawl metadata like software used, HTTP response codes, file metadata if available, IP addresses, URIs, time of harvest, MIME types, etc.

Viewing WARCs

```
WARC/1.0
WARC-Type: warcinfo
WARC-Date: 2015-10-05T16:14:23Z
WARC-Filename: ARCHIVEIT-4893-CRAWL_SELECTED_SEEDS-JOB176974-20151005161423785-00000.warc.gz
WARC-Record-ID: <urn:uuid:6b2bab3c-31b4-4cc6-a0e3-e70c298e269c>
Content-Type: application/warc-fields
Content-Length: 764

software: Heritrix/3.3.0-SNAPSHOT-20150924-2355 http://crawler.archive.org
ip: 207.241.239.72
hostname: aidata407.us.archive.org
format: WARC File Format 1.0
conformsTo: http://bibnum.bnf.fr/WARC/WARC_ISO_28500_version1_latestdraft.pdf
isPartOf: 4893-20151005161419014
description: jobId=176974, recurrence=NONE, maxDuration=600, maxDocumentCount=null, isTestCraw
robots: obey
http-header-user-agent: Mozilla/5.0 (compatible; special_archiver; Archive-It; +http://archive


WARC/1.0
WARC-Type: response
WARC-Target-URI: dns:iconosquare.com
WARC-Date: 2015-10-05T16:14:22Z
WARC-IP-Address: 207.241.239.252
WARC-Record-ID: <urn:uuid:d68018d7-2155-493f-a3ef-ea133d94da42>
Content-Type: text/dns
Content-Length: 56

20151005161422
iconosquare.com.    3600    IN  A   77.87.108.221


WARC/1.0
WARC-Type: response
WARC-Target-URI: http://iconosquare.com/robots.txt
WARC-Date: 2015-10-05T16:14:24Z
WARC-IP-Address: 77.87.108.221
WARC-Payload-Digest: sha1:O3D4J7RMKK7WPF7EFRLHJ3KFQIRGIKFC
WARC-Record-ID: <urn:uuid:6f568506-9558-4490-a3b8-59a959dc5e6f>
Content-Type: application/http; msgtype=response
Content-Length: 566
```



Unzipped files



Browser Wayback Access

Crawler limitations

- Can only capture what the crawler can find
- Dynamically generated URLs & content
- Temporal coherence
- Not suitable for rapidly updating content→ updates faster than crawler
- Broader academic questions about context, original order

Preservation challenges

- WARCs are massive
- No exact one-to-one relationship between URL captured and single WARC or even crawl to single WARC
- URLs and content cross reference each other for purposes of deduplication which makes it difficult to split files apart
- Playback reliant on technical infrastructure many institutions can’t create on their own (e.g. Open Wayback)
    - Note: Desktop versions of web archive players exist, but files still HUGE!
- Format obsolescence within captures, future need for browser emulation

Additional Challenges

- Technology
  - LOCKSS
  - Archive-It
- Capacity
- RoI

Benefits

- Collection development, preservation and access of key resources
- Distributed work
- Co-learning

Our Collections:

- Government of Canada Publications (DSP)
- Provincial & Territorial collections
- Thematic collections

Activities

- End of Parliament Session Crawl
- Federal Election Crawl
- Territorial and Provincial crawls
- Risk Management Strategy
- Recruitment & outreach

Building a community

CGI DPN Members:

- David Boudinot, University of Victoria
- Graeme Campbell, Queen's University
- Katie Cuyler, Past Chair, University of Alberta
- Corey Davis, University of Victoria
- Eamon Duffy, McGill University
- Carla Graebner, Chair, Simon Fraser University
- Sheila Laroque, University of Saskatchewan
- Susan Paterson, University of British Columbia
- Nicholas Worby, Co-Chair, University of Toronto
- ***YOUR NAME HERE***

Past members:

- Amanda Wakaruk
- Sam-chin Li
- Steve Marks
- Mark Jordan
- Umar Qasem
- Renee Saucier
- Geoff Brown
- John Wright
- Alex Burdett
- Caron Rollins
- Dani Pahule
- Karim Tharani
- ….

Growing the community: Federal Election Crawl Volunteers

- Elizabeth-Anne Johnson
- Tanya Ulmer
- Janice Banser
- Sarah Lake
- Renee Saucier
- Snowden Becker
- Tayler Stefanovic
- Heather Wilson
- Elizabeth Walker
- Yoo Young Lee
- Geoff Brown
- James MacGregor
- Elizabeth Walker
- Lisa Smith
- Hailey Siracky
- Francine May
- Giovanna Badia
- Veronica Bergsten
- David Greene
- Kiara McNaught

Resources

How to’s:

- Awesome Web Archiving
- IIPC Training Materials
- Web Archiving At UBC Library
- Archive-It Video Curriculum

Tools:

- Archive-it
- Conifer
- Browsertrix