Web Archiving at UTL

iSkills workshops
February 2018

Sam-chin Li
Reference and Government Information Librarian, UTL

Nich Worby
Government Information and Statistics Librarian, UTL
Agenda

• What is web archiving
• Why we collect?
• Current projects and efforts
• Ways to use the collections
• Issues
  • Copyright (privacy, social media)
  • Limitations
  • Sustainability
• Questions
What is 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.

IIPC: International internet preservation consortium
Erreur HTTP 404 - Non trouvé

Nous vous remercions de visiter le site Web de Bibliothèque et Archives Canada. Désolé, l'information que vous cherchez a été changée ou n’est plus disponible.

Pour en savoir davantage sur nos activités et initiatives organisationnelles, ou nos programmes et services, ainsi que notre Collection consultez notre site web.

Veuillez mettre à jour vos signets ou liens en conséquence

- Généalogie et histoire familiale
- Immigration
- Bases de données
- Patrimoine autochtone
- Patrimoine postal et philatélique
- Patrimoine militaire
  - Soldats de la Première Guerre Mondiale (CEF)

Vous pouvez aussi nous contacter pour de plus amples informations ou pour nous signaler un lien brisé.

N'hésitez pas à nous faire part de vos commentaires et suggestions.

Nous regrettons tout inconvénient que cette situation aurait pu entraîner.

Merci

HTTP Error 404 - Not Found

Thank you for visiting the Library and Archives Canada website. Sorry, the information you are looking for has changed or is no longer available.

For information on our corporate activities and initiatives, or our programs, services as well as our Collection visit our Website.

Please update your bookmarks and/or links accordingly.

- Genealogy and Family History
- Immigration
- Databases
- Aboriginal Heritage
- Postal Heritage and Philately
- Military Heritage
  - Soldiers of the First World War (CEF)

You can also contact us for more information, or to report a broken link.

We welcome your ideas and suggestions.

We apologize for any inconvenience this may have caused.

Thank you
Ephemeral nature of web content

• Lifespan of web pages
  • 44 days (Brewster Kahle, Preserving the Internet, Scientific American)
  • 75 days (Steve Lawrence, Persistence of web references in scientific research)
  • 100 days (Rich Weiss, On the Web, Research work proves ephemeral)

• News-driven and social media websites
• Overwritten to more up-to-date information
Why we collect?

- Teaching
- Research
- Cultural Heritage
- Accountability
- At Risk Content
Archive-It (Internet Archive)

- A web archiving subscription service from the Internet Archive
- Captures and provides access to archived websites
- Seed selection and management
- Scope of a crawl
- Crawl frequency
- Quality assurance
- Metadata
Canadian Government Information Collection: Background

• No print government publications were to be distributed to libraries across Canada and overseas in April 2014

• Electronic-only model
  • DSP e-collection (pdf)
  • Departmental websites (html)

• Library and Archive Canada web archiving government websites stopped at 2007

• No comprehensive, national digital preservation plan

• Numerous department libraries were closed
### UTL Web Archiving Collection (1):
Government of Canada (GoC) web site

#### Crisis

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>11/2012</td>
<td>leaked information</td>
</tr>
<tr>
<td></td>
<td>Freedom of Information request</td>
</tr>
<tr>
<td>3/2013</td>
<td>1,500 websites to 1</td>
</tr>
<tr>
<td></td>
<td>60% content will be cut</td>
</tr>
<tr>
<td>7/2013</td>
<td>Deadline</td>
</tr>
</tbody>
</table>

#### Response

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/2013</td>
<td>Contacted LAC</td>
</tr>
<tr>
<td></td>
<td>UTL crawled GOC sites</td>
</tr>
<tr>
<td>4/2013</td>
<td>Collaborated with other libraries</td>
</tr>
<tr>
<td>6/2013</td>
<td>Wayback crawl moved to UTL Archive-It account (Dec. 2007 – June 2013)</td>
</tr>
</tbody>
</table>

*Mission impossible!*
UTL Web Archiving Collection (1):
GoC web sites

- 2005 – Nov. 2007 (Library and Archives of Canada)
- Dec. 2007 – 2015 (University of Toronto Archive-It)
- 2013- present (Library and Archives Canada)
UTL Web Archiving Collection(2): City of Toronto web site

- City of Toronto
  - Content
- UTL
  - Web archiving
- Toronto Archives
  - Quality assurance
  - Metadata
UTL Web Archiving Collection(3): Toronto 2015 Pan Am & Parapan American Games

Documentary history of hosting the games across several Ontario municipalities
UTL Web Archiving Collection(4): Hong Kong Politics

- Social media
- Blogs
- Government sites
Other UTL Web Archiving Collections

- Canadian Political Parties and Political Interest Groups (2005+)
- Ontario Government website (2014+)
- Canadian Labour Unions (2006+)
- Federal Election Candidate Sites 2015
- Global Summitry Archive
- Snowden Archive
- Toronto Mayoral Election 2014
- Ontario Election 2018 (upcoming project)
- University of Toronto Archives (2006+)
Web Archiving Partnership Collection (1)
Canadian provincial & territorial government web sites

• Partner: Canadian Government Information Digital Preservation Network (CGI DPN)
• Background (2016):
  • Out of Library and Archive Canada (LAC) mandate
• Workflow:
  • Environmental scan
  • Dividing the provincial / territorial sites among members
  • Check copyright status
  • Ministries first, then agencies and commissions
Web Archiving Partnership Collection (2): Opioid Crisis in Canada

- Partner: CGI DPN
- Background:
  - Out of LAC mandate
- Workflow:
  - Vancouver: UBC, Victoria University
  - Ontario: UTL
Web Archiving:
Selection criteria (CGI DPN)

• Risk
• Availability
• OA service with / without digital preservation plan
• Fee-based service
• Size / budget: storage space, frequency
• Accommodations
• Copyright / privacy issues
• Content type: dynamically generated, database
• Partnership opportunity
• Significance of this content and how it fits the collection priorities
Web Archiving efforts:

Guerrilla Archiving

Saving Environmental Data from Trump

- Data Refuge project
- Environmental Data & Governance Initiative (EDGI)
- End of Term Web Archive
- Internet Archive
- U of T
  - December 17, 2016
  - Led by the Technoscience Research Unit
  - 150 academics and volunteers (including 6 UTL librarians)
  - Media interest
  - Fed 3,142 URLs into the Internet Archive web crawler
  - Data transmitted to Penn Data Refuge Project
  - Tool kit
Guerrilla Archiving: Why we participated

Donald J. Trump
@realDonaldTrump

The concept of global warming was created by and for the Chinese in order to make U.S. manufacturing non-competitive.

11:15 AM ~ 6 Nov 2012
Ways to use the collections

• Wayback Machine browser access
• Data/text mining
• Mapping and visualization
• Research tool: data analysis
Ways to use the collections (1):
Archived-It browse access
Enhanced access:
Metadata and search capabilities

Site search

Full text advanced search
Approaches to web archiving

- Server-side archiving
  - Downloading content directly from server

- Transactional archiving
  - Archiving user requests: SiteStory

- Client-side archiving
  - Use of crawlers remotely: Heritrix, Wget, Archive-It

Transactional archiving

Architecture Overview

Regular user of live version of site

Image: https://mementoweb.github.io/SiteStory/
Approaches to web archiving

- **Server-side archiving**
  - Downloading content directly from server

- **Transactional archiving**
  - Archiving user requests: SiteStory

- **Client-side archiving**
  - Use of crawlers remotely: Heritrix, Wget, Archive-It

Web crawlers

HTTP Request

Seed URL: www.utoronto.ca

HTTP Response

Host server(s)

Crawl parameters:
- Scoping rules
- Time limits
- Data limits
- # of seed URLs within crawl
- Robots.txt

Parse, open, copy, repeat...

Copy

Job

WARC file(s)

URLs

- .../about.html
- .../programs.html
- .../graduate.html
- .../news.html
- .../research.html

Copy

- .../about1.html
- .../about2.html
- .../photo.jpg
- .../style.css
- ...etc

Copy

- .../arts.html
- .../science.html
- .../photo.jpg
- .../style.css
- ...etc
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, URLs, time of harvest, MIME types, etc.
Viewing WARC

WARC/1.0
WARC-Type: warcinfo
WARC-Date: 2015-10-05T16:14:23Z
WARC-Filename: ARCHIVEIT-4809-CRAWL_SELECTED_SHEDS-20151005142385-00000.warc.gz
WARC-Record-ID: <urn:uuid:dd60010d7-2153-483f-a3ef-ee133d4da42>
Content-Type: application/warc-fields
Content-Length: 76

software: Heritrix/3.8.0-SDAPSHOT-20150926-2355
http://crawler.archive.org
ip: 207.241.239.72
hostname: aiding407.us.archive.org
filename: WARC File Format 1.0
isVersion: 4809-20150805142385
description: jobID=176976, recurrence=NOW, maxDuration=600, maxDocumentCount=2500, $TESTCH
robots: obey
http-header-user-agent: Mozilla/5.0 (compatible, special_archives/Archive-T; http://archive

WARC/1.0
WARC-Type: response
WARC-Target-URI: dns:iconosquare.com
WARC-Date: 2015-10-05T16:14:23Z
WARC-IP-Address: 207.241.239.252
WARC-Record-ID: <urn:uuid:dd60010d7-2153-483f-a3ef-ee133d4da42>
Content-Type: text/dns
Content-Length: 56

20151005161423
iconosquare.com: 3000 IN A 77.57.105.221

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-Response-Strict-Transfer-Encoding: chunked
WARC-Response-Content-Digest: sha1:3841794F710DAF9F3D45D723E4F625A3F900-B0-3AE805AD5E6F
Content-Type: application/http; msgtype=response
Content-Length: 546

HTTP/1.1 200 OK
Last-Modified: Tue, 13 Sep 2015 11:29:36 GMT

Unzipped files

Browser Wayback Access
Crawler limitations

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

Crawlers: Heritrix, webrecorder.io, WARCreate, Wget & more

Access: Open Wayback, Web Archive Player

Storage

Servers

Bundled programs: WAIL

Curation: Web Curator Tool, NetArchive Suite & more

Staffing

Open Source

Software as service platforms

- Archive-It (Internet Archive)
- ArchivetheNet
- Hanzo Archives

& more

Subscription

Staffing
What is WAIL

A desktop based graphical user interface that includes:

• Heritrix
• PyWB (a Python based version of Open Wayback)
• GUI for collections & managing WARCs
Navigating WAIL

View your captures in PyWB in your browser
Quick examples

<table>
<thead>
<tr>
<th>Capture</th>
<th>Status</th>
<th>Original Url</th>
<th>Archive File</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fri, 12 Jan 2018 20:54:02 GMT</td>
<td>200</td>
<td><a href="https://www.youtube.com/user/iSchoolUofT">https://www.youtube.com/user/iSchoolUofT</a></td>
<td>iSkills-1515790424329.warc</td>
</tr>
</tbody>
</table>

* Unique captures are bold. Other captures are duplicates of a previous capture.
Seeds

1. http://semaphore.utoronto.ca/ (with Page + Same Domain Links)


Web archiving workflow (with Archive-It)

- Selection
- Capture
- Quality Assurance
- Metadata
- Provide access
- Storage: Internet Archive

Troubleshoot
Web archiving workflow (with Archive-It)

**Selection**
- Collaborative selection of seed list with collection development & liaison librarians, as well as stakeholder groups & faculty
- Establish scope of collection, e.g. which URLs to capture & not capture, frequency, etc.
- Document decisions!

**Capture**
- Add crawl parameters (e.g. scoping rules, time limits, data limits, ignore file types/robots.txt) & run crawl in Archive-It dashboard
Web archiving workflow (with Archive-It)

- Review post crawl reports
- Manual inspection
- Review crawl parameters & scope
- Patch crawl or re-crawl with new parameters
- Contact Archive-It tech support
Web archiving workflow (with Archive-It)

- Bulk uploads of Dublin Core metadata via a spreadsheet or individual in Archive-It dashboard
- Metadata at seed or event document level
- OAI PMH feed opt-in
- Can download converted MARC records from OCLC

- Content flagged as public added to general Internet Archive Wayback Machine as well as individual Archive-It collection pages
- Full text searchable access via collection pages
- Access to derivative files (e.g. WATs, WANEs, LGEs) through Archive-It Researcher Services
Web archiving workflow (with Archive-It)

- Storage & on-going maintenance done by the Internet Archive
- Data resides on U.S. servers
- Data subject to Internet Archive’s removal policies
- WARCs can be downloaded & hosted locally
Preservation issues

• Format obsolescence within WARCs
• Changing web technologies: crawlers always a step behind
• Few transparent collection development & selection policies
• How to communicate provenance?
• Institutions & sustainability of projects
Discussions of results

• What captured well?
• What couldn’t you capture? Why?
• What alternatives strategies could you use?