Getting Started with Archive-IT Services

Andrea Mills

Booksgroup Collections Specialist
Internet Archive

- Micro History
- Text Archive Update
- Archive-IT Services
1996 – The Internet Archive is created, with the goal to archive and preserve the World Wide Web

www.archive.org
2004-- Book digitization begins at University of Toronto Libraries

2006--Archive-IT begins targeted web archiving services
OpenLibrary, TVNews, Audio and Video, Computer Games and Software
Updates

10 Years of Digitization
A Decade of Collecting

- 2.3 million eBooks
- 1250 Contributing Institutions
- 400 Sponsors
- 2450 unique texts collections
- More than 150 digitization projects currently underway
Government Publications

Canadian Government Publications

Internet Archive Canada, with advice and assistance from government and university librarians across Canada, have digitized more than 20,000 Canadian Government publications and made them freely available online. This page is meant to serve as a portal for discovering publications digitized by Internet Archive Canada, to be a reference for future digitization initiatives and provide incentive to grow the list of titles that have been preserved and dedicated to the public domain.

Beginning in 2004, Internet Archive Canada began to digitize material from the Library and Archives of Canada. This work has continued steadily over the last
Social Media

Twitter
@internetarchive
@IABooksGlobal

Instagram
http://instagram.com/iabookscanada

Flickr
www.flickr.com/photos/internetarchivebookimages
Getting Started with Archive-IT Services

https://archive-it.org
Archive-IT.org
Web Archiving

The process of collecting portions of web content, preserving the collections, and then providing access to the archives - for use and re-use.
## Archive-IT vs. Wayback Machine

<table>
<thead>
<tr>
<th></th>
<th>Archive-It</th>
<th>General Archive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focused collections</td>
<td>One collection</td>
<td></td>
</tr>
<tr>
<td>Control over scope and frequency</td>
<td>Snap shot every 2 months</td>
<td></td>
</tr>
<tr>
<td>Technical support</td>
<td>Automated</td>
<td></td>
</tr>
<tr>
<td>Content indexed for full text search</td>
<td>Search not available</td>
<td></td>
</tr>
<tr>
<td>Content cataloged with metadata</td>
<td>Cataloging not available</td>
<td></td>
</tr>
<tr>
<td>Archived data can be shipped</td>
<td>Shipping not available</td>
<td></td>
</tr>
<tr>
<td>Restricted access options</td>
<td>Public access only</td>
<td></td>
</tr>
<tr>
<td>Access archived data 24 hours later</td>
<td>Access archived data 30-60 days later</td>
<td></td>
</tr>
</tbody>
</table>
Archive-IT Services

• Web based application and fully hosted solution; includes access and storage (2 copies)
• Tools for selection, scoping and metadata creation—Scope-IT
• Capture content using 10 different frequencies
Types of Content

• HTML, text, video, audio, social media, PDF, images, password-protected content, static databases, newspapers
• Social Media: Flickr, Twitter, Instagram, Vimeo and Facebook—only with Archive-IT
Features

• Different levels of access for users
• Browse collections by both URL, Full text search (basic and advanced) and metadata search
• 9 post crawl reports for Analysis
• Online Help Section, Partner Specialists and Tech Support
How does it Work?

Heritrix: Web Crawler

Umbra: Assists/provides flexibility for the crawler to access sites as a browser does

Wayback Machine: Access tool for rendering and the viewing pages - the web as it was.

NutchWAX: Search engine – Full-text search

SOLR: Metadata search
Starting to Collect
Big Questions

• Do you have a Mission/Mandate to Collect?
• What are the Goals and Objectives for the Collection?
• Vision for the Collection?
Mandate to Collect... What now?

- Institutional
- Collection
- Web Content
Goals and Objectives

• Why is this web archive important?
• Short-term Vision (3 yrs.)
• Long Term Vision (10 yrs.)
Vision for Collection

• What will it look like?
• How will it be used?
• How will it be managed and maintained?
Broad to Specific

As of today, Archive-It has collected 8,961,536,030 URLs for 2,643 public collections!
Canadian Government Information—collected by University of Toronto has 605 seeds
Prairie Provinces Politics & Economics—collected by University of Alberta has 393 seeds
Site Closures

Aboriginal Canada Portal—Closed February 12, 2013
10 Years on Mars:
Collected by University of Michigan

Capture public perception of the Mars Rovers on their 10th anniversary, and to preserve and provide access to that information for the future.

1. Official government documents
2. Popular news and Science media
3. Fringe (conspiracy theorizing, alien spotting...)
Current Events

Ebola Virus Disease—Collected by University of Manitoba has 13 seeds
Test Account and Practise

https://archive-it.org/contact-us
Test Account

• Create a collection, capture content and view the results
  • Start with Five (5) URLs
    • 1 crawl
  • Archive up to 250,000 webpages
Is your seed already in the Wayback Machine?

Search both keywords and URLs

https://archive-it.org/explore
Is the Site Archived Elsewhere?

• Ask your Colleagues
• LISTSERVs
• Registry options?
Valuable Experience

• Attempt to capture all or part of your proposed collection in your test crawl
• This will help determine Scope, Frequency, QA needs and Subscription level
Start Collecting

• Refer back to Mission, Goals and Vision for collection
• Repeat
Learn More

https://archive-it.org/learn-more

Download our white paper on the web archiving life cycle

Check out our blog:

https://archive-it.org/blog