Archive-It: From 4.8 to 5.0

How 279 institutions archive the web

Scott Reed
Partner Specialist
Archive-It

How is Archive-It different than the General Archive (www.archive.org)?

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

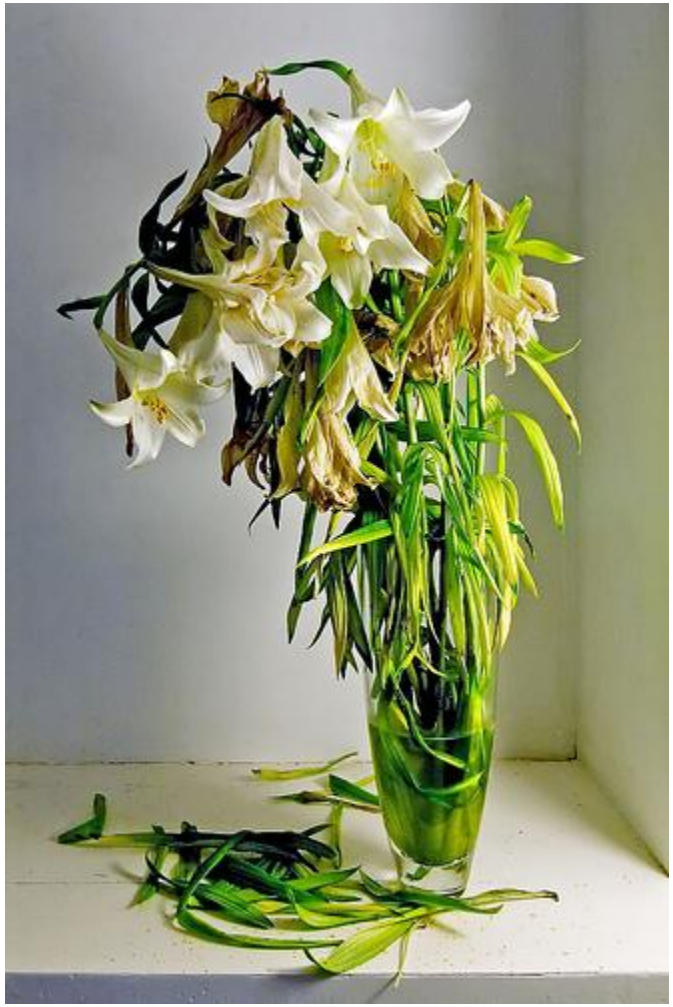
Goals for a web archive

A web archive is a collection of archived URLs grouped by theme, event, subject area, or web address.

A web archive contains as much as possible from the original resources and documents the change over time. It is a priority to recreate the same experience a user would have had if they had visited the live site on the day it was archived.

How long does a website live?

- A report in **Scientific American** claims 44 days.
- A subsequent academic study in **IEEE** suggests 75 days.
- A **Washington Post** article indicates the number is 100 days.
- A recent study by **ODU** says that after the first year of publishing, nearly 11% of social media will be lost and after that we will continue to lose 0.02% per day



About the service

Archive-It is a subscription service deployed in February 2006

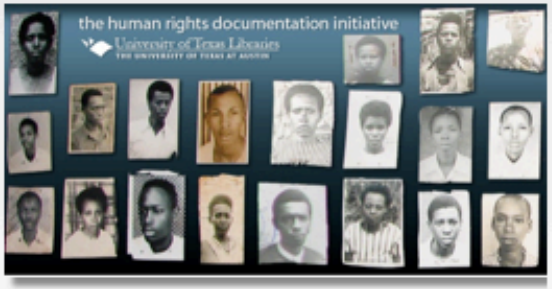
- **Web based application** that allows users to create, manage, access and store collections of digital content
- The service is a **fully hosted solution**, and includes access and storage.
- **Provides tools for selection and scoping** including cataloging with metadata
- Ability to **capture content using 10 different crawl frequencies**
- Archived content includes: html, videos, PDFs, images, + more from many site types, including social media
- **Can browse archived content 24 hours after a capture is complete**; and full text search is available within 7 days
- **Restricted access options** are available

Subscription Based Service

- Many different account levels depending on the amount of data captured and frequency of crawls.
- Only billed for current subscription year's collecting activity.
- Long term access and storage (2 copies) of archived content is not dependent on continued subscription to the service.
- We work with individual institutions as well as consortiums.
- Free Trial Accounts are offered to archive up to five websites and discuss scope of your organization's web archiving plan and budget.
- Subscription includes Partner Specialist user support and technical support.

Accessing Content

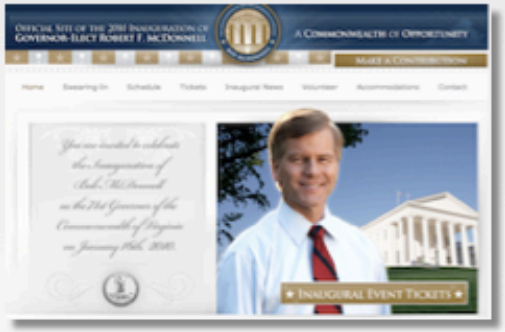
Explore Collections | Find a Collection by Name | Search | *Show All Collections*



Human Rights Documentation Initiative

By University of Texas at Austin Libraries, Human Rights Documentation Initiative

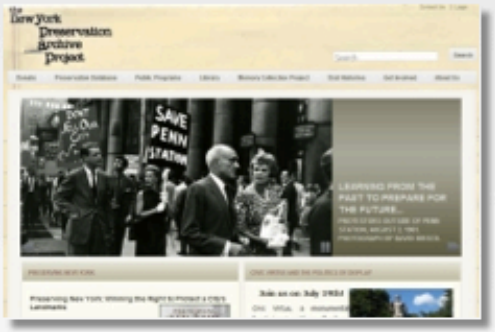
The University of Texas Libraries' Human Rights Documentation Initiative Collection features fragile websites containing human rights documentation and related...



Virginia's Political Landscape, 2010

By Library of Virginia

A collection of Web sites that document Virginia's 2010 Congressional elections (primary and general). All 11 members of Virginia's Congressional...



Avery Library

By Columbia University Libraries

A collection of websites chosen by subject specialists from the Avery Architectural and Fine Arts Library at Columbia University.

Accessing Content

UNIVERSITY OF WATERLOO

Administration, Faculties, and Services

Collected by: University of Waterloo

Archived since: Sep, 2010

Description: University administration, the six university faculties, and major services

Subject: Universities & Libraries

Narrow Your Results

Group Sort By: Count | (A-Z)

a) Home Page and Anniversary Pages (4)
b) Senior Administration (10)
c) Office of the Registrar (4)
d) Development and Alumni Affairs (3)
e) Student Success (5)

More ▼

Creator Sort By: Count | (A-Z)

Communications and Public Affairs (7)
University of Waterloo (4)
Faculty of Arts (3)
Faculty of Engineering (3)
Office of Student Success (3)

More ▼

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here | Search | Clear

Sites | Search Page Text

Page 1 of 1 (58 Total Results)

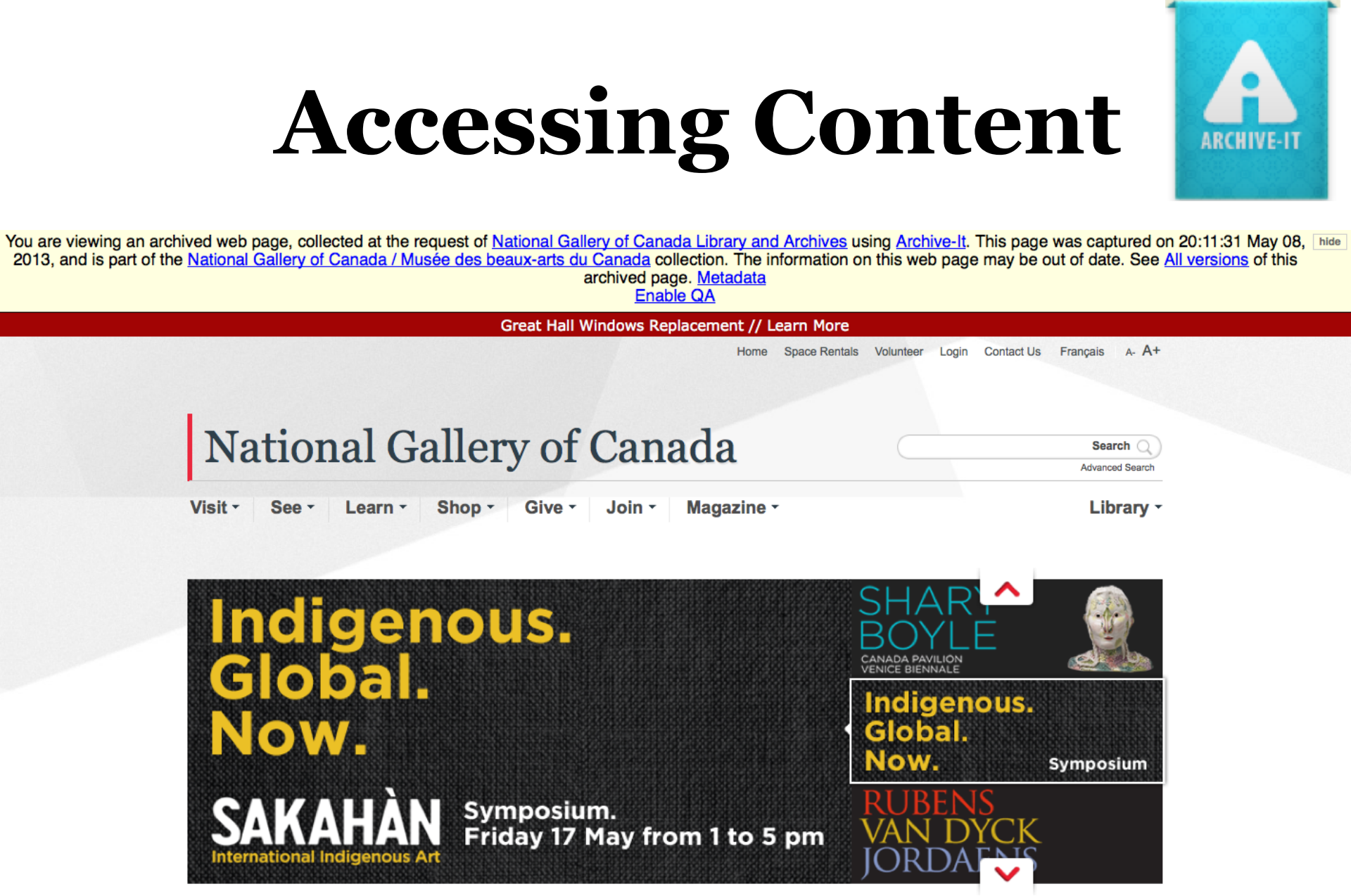
Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Office of Alumni Affairs

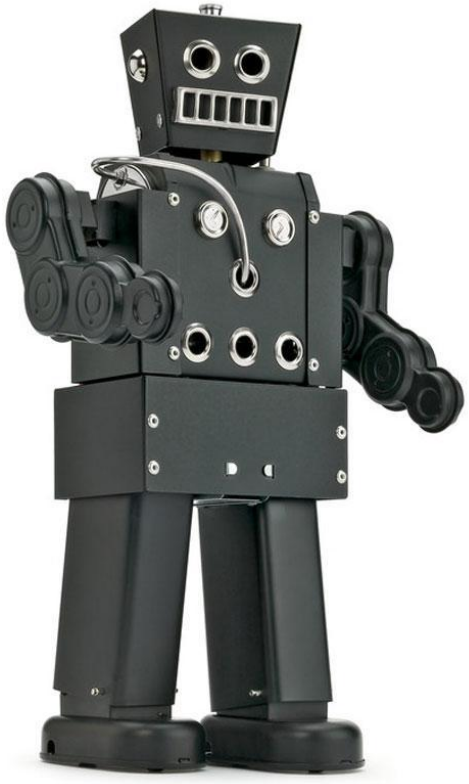
URL: http://alumni.uwaterloo.ca/

Captured 31 times between May 4, 2011 and Dec 17, 2012
Videos: 1 Videos Captured
Group: d) Development and Alumni Affairs
Creator: Office of Alumni Affairs

Accessing Content



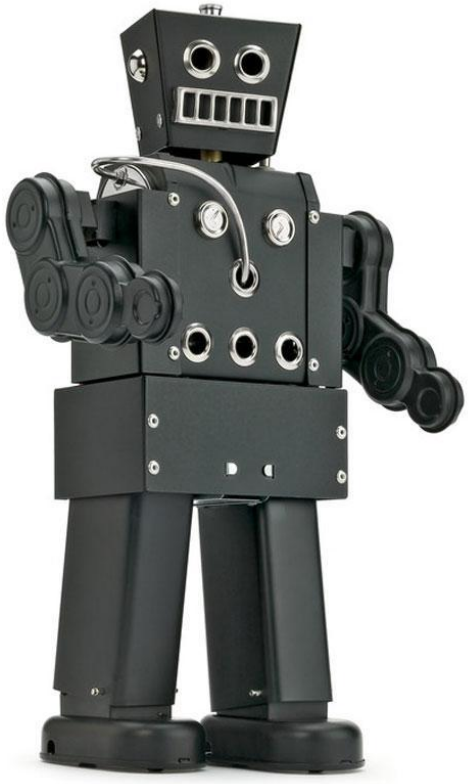
How Archive-It Works



- Crawlers (also known as spiders or robots) are pieces of software that visit websites and index the information included therein (think of Google – it works because of crawlers).

- To archive the Web, Archive-It crawls URLs and captures a copy of the information and files displayed on selected websites.

How Archive-It Works



1. Starts with your seed URL(s)
2. Checks if those URLs are reachable, and archives them
3. Check embedded content – what does it need to render the page? (CSS, Javascript, Images, etc.)
4. Look for links to other pages, checks if those pages are 'in scope', and archives them
5. Keeps going until either:
   - Cannot locate any more links that are in scope, or
   - Hits the maximum time or data limit for the crawl

Archive-It 4.8

May 2013

- Ability to import metadata
- Restrict access by IP address
- Quality Assurance while browsing (Wayback QA)
- Alpha feature: password protected content

Archive-It 5.0

Release: 2014

- Complete re-design of the user interface
- Overhaul of web technologies (capture mechanisms, search platform, display options)
- Increased ways for patrons to access archived content including integration with internal systems and catalogs
- Focus on media rich sites and social media (Facebook, Twitter, Flickr Youtube, etc)
- Researcher tools visualizations and analytics

+ more Archive-it Partner requested features derived from user support groups, usability testing, surveys

The Web Archiving Life Cycle Model