Apr 23rd, 1:00 PM - 4:00 PM

Web Archiving for Academic Institutions

Lori Donovan
Internet Archive

Mary Haberle
Internet Archive

Web Archiving for Academic Institutions

Presenter 1 Title
Senior Program Manager, Archive-It

Presenter 2 Title
Web Archivist

Session Type
Workshop

Abstract
With the advent of the internet, content that institutional archivists once preserved in physical formats is now web-based, and new avenues for information sharing, interaction and record-keeping are fundamentally changing how the history of the 21st century will be studied. Due to the transient nature of web content, much of this information is at risk. This half-day workshop will cover the basics of web archiving, help attendees identify content of interest to them and their communities, and give them an opportunity to interact with tools that assist with the capture and preservation of web content. Attendees will gain hands-on web archiving skills, insights into selection and collecting policies for web archives and how to apply what they've learned in the workshop to their own organizations.

Location
KIPJ Room B

Comments
Lori Donovan works with partners and the Internet Archive's web archivists and engineering team to develop the Archive-It service so that it meets the needs of memory institutions. She also serves as Program Manager for the Internet Archive's crawling with Library of Congress. She enjoys working at a mission-based organization, helping organizations fulfill their own missions by archiving the web. Lori has a Masters of Science in Information from the University of Michigan, specializing in Archives and Digital Preservation. She previously studied history and political science at Boise State University.

Mary Haberle is a Web Archivist at Archive-It where she provides partner training and support services. Her prior work experience includes processing archival collections at the Academy of Motion Picture Arts and Sciences and the University Club of New York, as well as contributing to digitization projects at the American Jewish Joint Distribution Committee and Franklin Furnace Archive. Mary earned her Master of Library and Information Studies degree from McGill University and a Digital Archives Specialist Certificate from the Society of American Archivists.

Web Archiving for Academic Institutions

Instructors:
Lori Donovan
Mary Haberle
Internet Archive

DIS 2018
April 23, 2018 University of San Diego

AGENDA

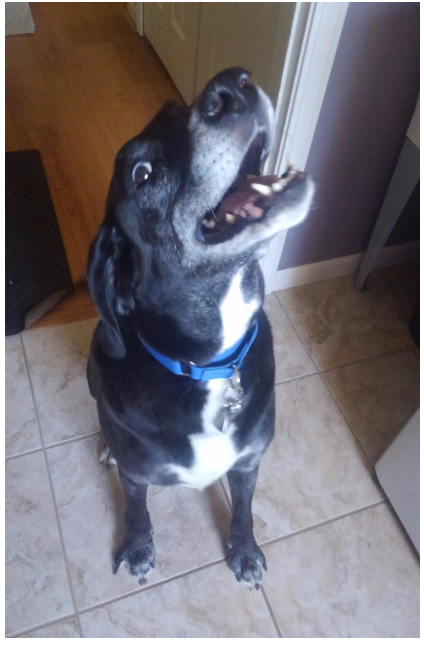
Purrsonalities by Sarah Landstrom

- ★ **Definitions and Community**
  - The what and why of web archives
  - Current challenges and initiatives
- ★ **Collection**
  - Selection and acquisition
  - Description and scoping
- ★ **Manage**
  - Challenges & Opportunities
  - Tools & Services
- ★ **Use**
  - Access & Research
- ★ **Demonstration**
- ★ **Create your own collection**

INTRODUCTIONS

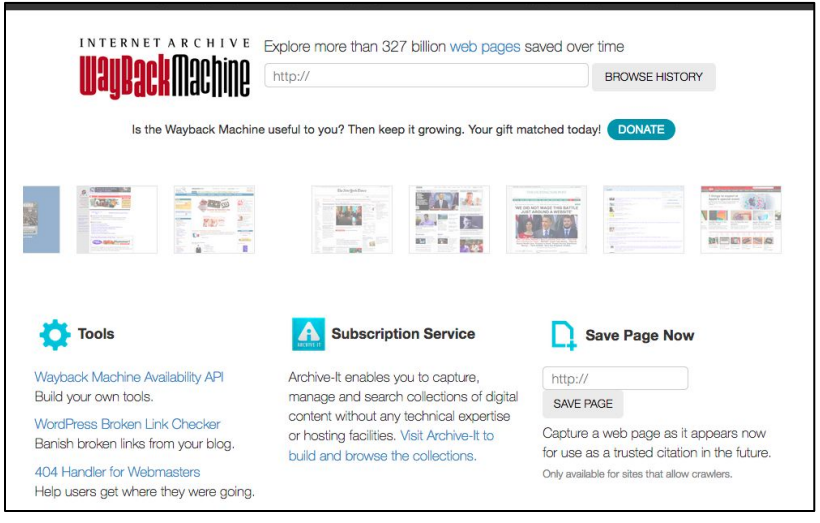
Name, organization, experience and/or interest in web archiving

Bonus points: Tell us about your pets!



- We are a non-profit Digital Library & Archive founded in 1996
- 35+PB unique data: 20PB web, 14m text, 3.6m vid, 3.5m aud, 100K soft, etc
- Developed: Open source web archiving tools, formats and standards
- Engineers, librarians/archivists, program staff

THE WAYBACK MACHINE



Online: https://archive.org/web/

The largest publicly available web archive in existence.

> 300 Billion Web pages
> 100 million websites
> 150 languages
~ 1 billion URLs added per week

WEB ARCHIVING

What is a web archive?
A collection of archived URLs grouped by theme, event, subject area, or web address and stored in WARC file format.

A web archive contains as much as possible from the original resources and documents their 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.

THE LIFESPAN OF A WEBSITE

How long does a website last?

In general, a typical web page can be expected to last **~90-100 days** before changing, moving, or disappearing completely.

> In 2013, our colleagues at Old Dominion University determined that **over 10%** of event related content posted to social media platforms is lost after one year.

> In 2014, a study by UCLA determined that **7-in-10** scholarly articles that include citations with hyperlinks suffer from *reference rot*.

WHY WEB ARCHIVE?

- **Institutional History**: Maintain a record of your institution's web presence over time.

- **Responsibility**: preserve things like course information, special exhibit information, policies, organizational reports— many documents now showing up only as digital content.

- **Research**: Many libraries are seen as authorities on a particular subject, topic or person, and collect web-based information to augment other holdings.

U.S. WEB ARCHIVING STATISTICS

National Digital Stewardship Alliance (NDSA) 2016 Survey (PDF)

- ★ 94% of respondents use an external web archiving service like Archive-It
- ★ 71% of organizations devote one half FTE or less to web archiving
- ★ 60% started programs between 2011 and 2015
- ★ 60% rely on other organizations' or community-generated policies in the creation of their own
- ★ Principle concerns include ability to archive social media (70%), video (69%), and databases (62%)

WEB ARCHIVING COMMUNITY

[International Internet Preservation Consortium]

[SAA's Web Archiving Section]

[Archive-It] ([Blog] / [Twitter])

@KristaOldham 6 May 2015

[Mid Atlantic Archive-It User Group]

USE CASES

Document a subject area or an event

- Often related to traditional collecting activity around the same topical focus
- Capture spontaneous events
- Document different perspectives and social commentaries

Fulfill a mandate to capture and preserve web history

- Support electronic records system to meet records retention requirements
- Collect publications/documents that are no longer in print form
- Historical record of an institution or individual's web/social media presence

Closure crawls

- Document a website before it changes, is redesigned, or closes

UNIVERSITY OF TEXAS AT AUSTIN:
LATIN AMERICAN GOVERNMENT DOCUMENTS COLLECTION

Use Case:

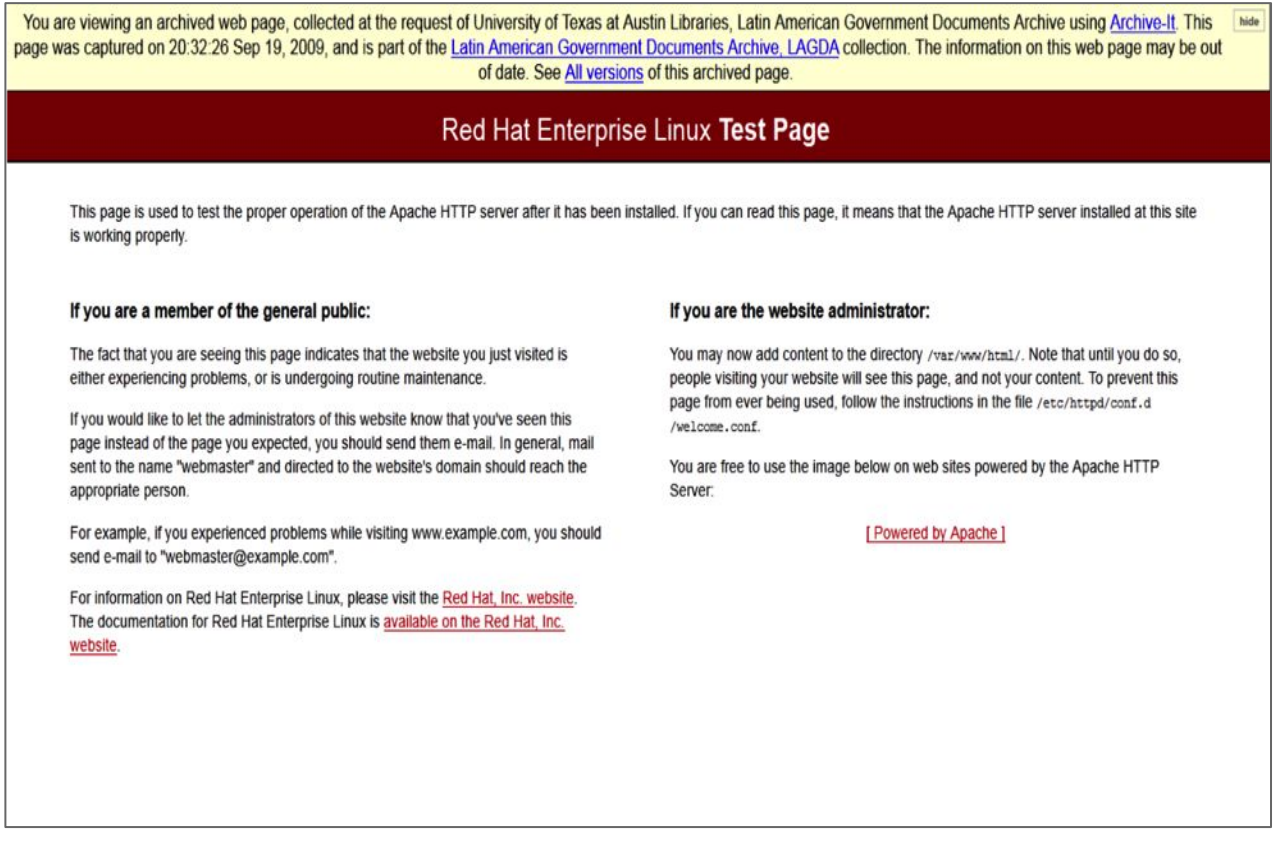
Archive government documents from 18 different countries in Latin America

Content includes:

- Full-text archives of official documents
- Original video and audio recordings of key regional leaders
- Thousands of annual and "state of the nation" reports
- Collections of Latin American elections and political parties

Honduras presidential website, 2008 **(before coup)**

UNIVERSITY OF TEXAS AT AUSTIN:
LATIN AMERICAN GOVERNMENT DOCUMENTS ARCHIVE



Honduras presidential website, 2009 **(during coup)**

Honduras presidential website, 2010 **(after coup)**

COLUMBIA UNIVERSITY

Use Case 1:

> Archive the university's web presence in order to meet required records retention mandates.

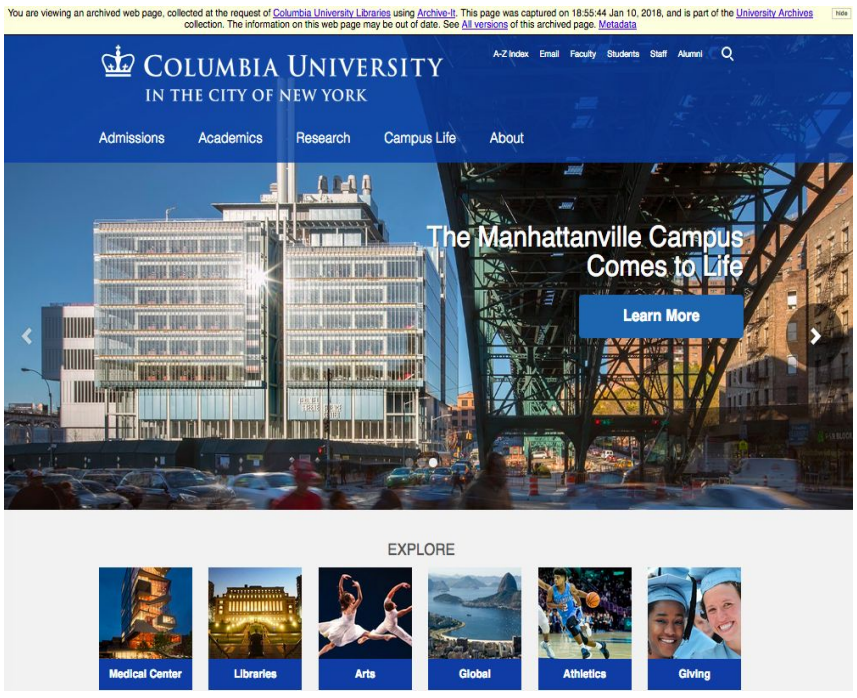
Use Case 2:

> Collect in subject areas selected by academics, librarians, curators throughout the university.

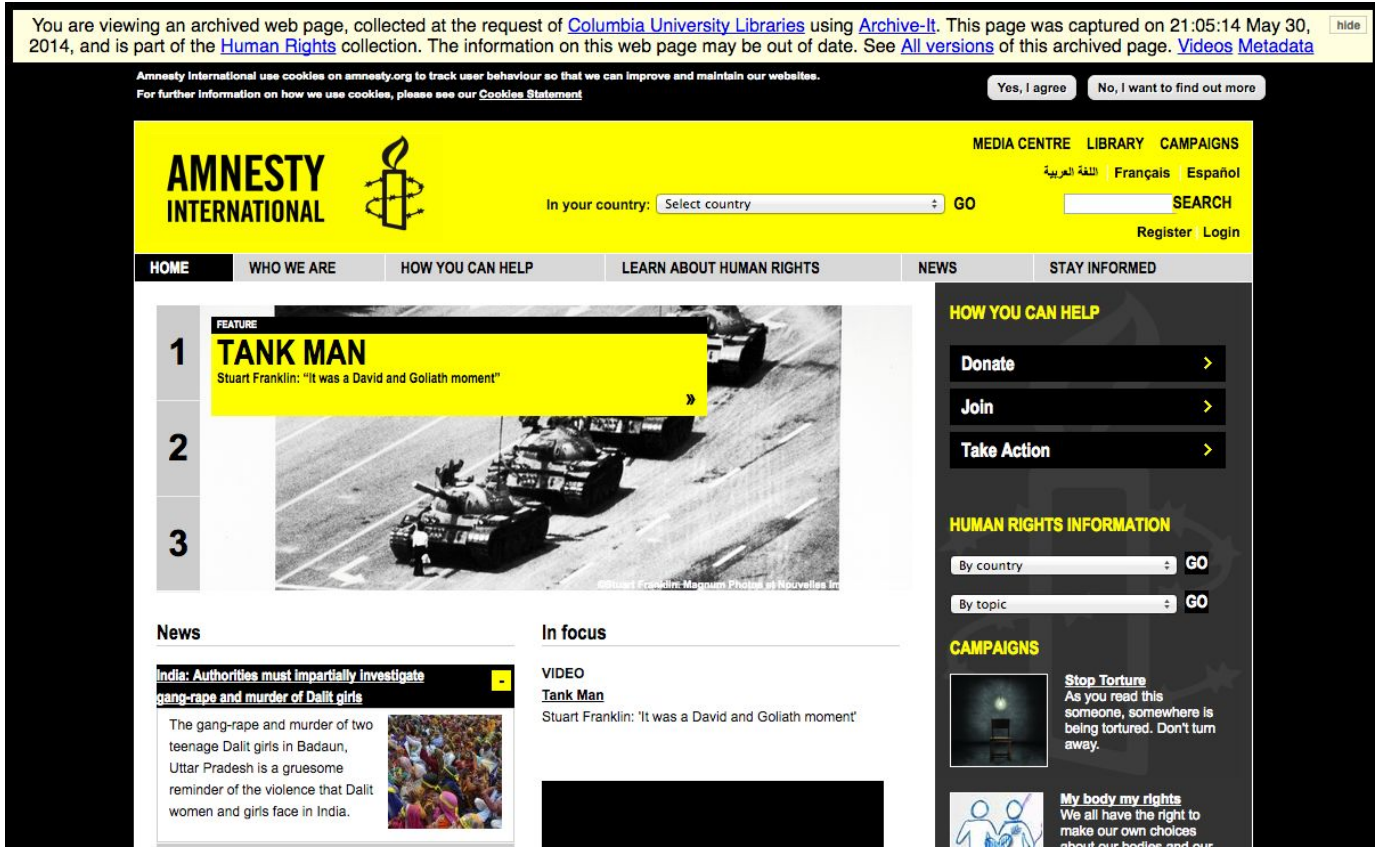
COLUMBIA UNIVERSITY: UNIVERSITY ARCHIVES

December 2011

January 2018

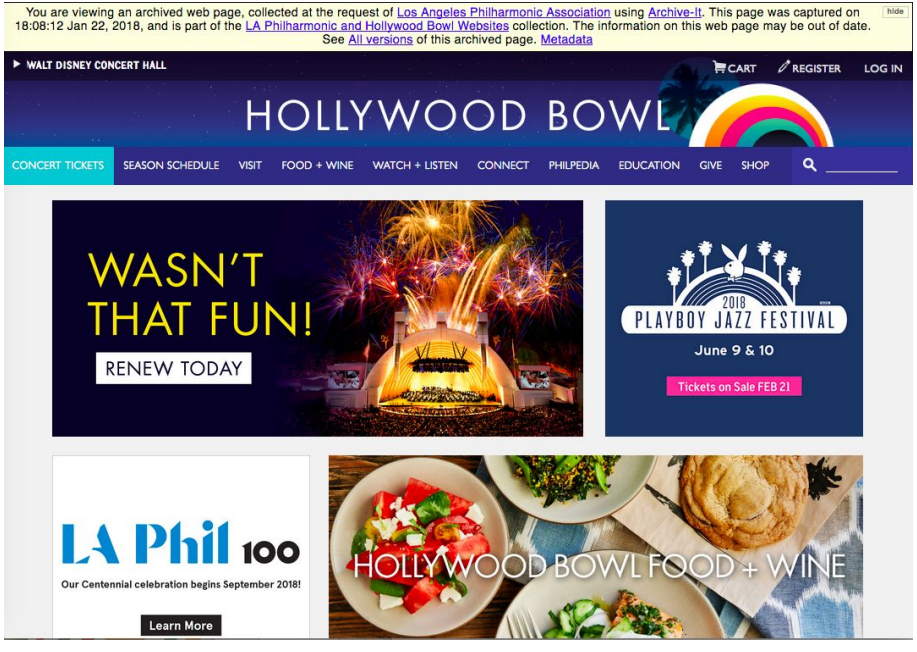


COLUMBIA UNIVERSITY:
CENTER FOR HUMAN RIGHTS DOCUMENTATION



LOS ANGELES PHILHARMONIC: HOLLYWOOD BOWL CLOSURE CRAWL

Closure crawl capture



Redesigned site on live web

LIBRARY AND ARCHIVES CANADA: TRUTH AND RECONCILIATION COMMISSION ARCHIVE

Government of Canada / Gouvernement du Canada

Canada

National Centre for Truth and Reconciliation
UNIVERSITY OF MANITOBA

THE UNIVERSITY OF WINNIPEG

UNIVERSITY OF MANITOBA

Use Case:

Collaborative collection building

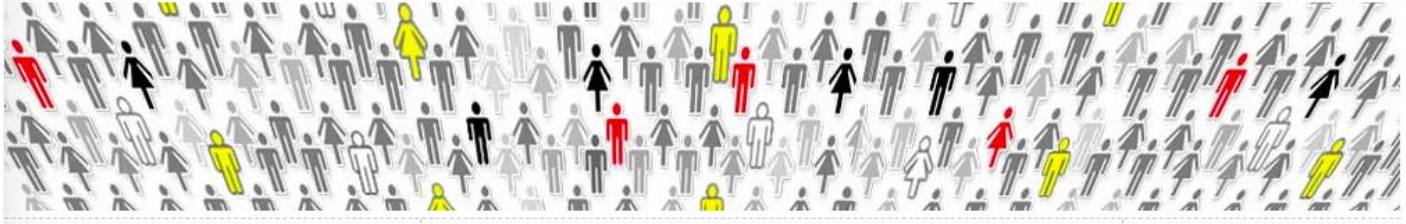
Content includes:

> Collaborative collections curated with other Canadian organizations

1000 conversations

1000 Conversations
Across Canada on Reconciliation

RSS Subscription



Archives
- March 2012
- September 2011
- August 2011
- July 2011
- June 2011
- May 2011
- March 2011
- February 2011
- October 2010
- July 2010

#285 Wetaskiwin, Alberta

- **Attendes:** Less than 50 people
- **Place:** Wetaskiwin, Alberta

Please tell us (in detail) about your event/conversation:

We downloaded the 1000 Conversations booklet and watched a few of the videos before discussing what we thought about residential schools. It was very sad and informative.

5 Latest Conversations
- #285 Wetaskiwin, Alberta
- #284 Ajagutaq/Katiniit Committee
- #283 Teralba Park Stolen Generation Commemorative Site Support Group
- #282 Community Wellness Centre
- #281 Igloolik Health Centre

QUESTIONS?

STARTING A COLLECTION

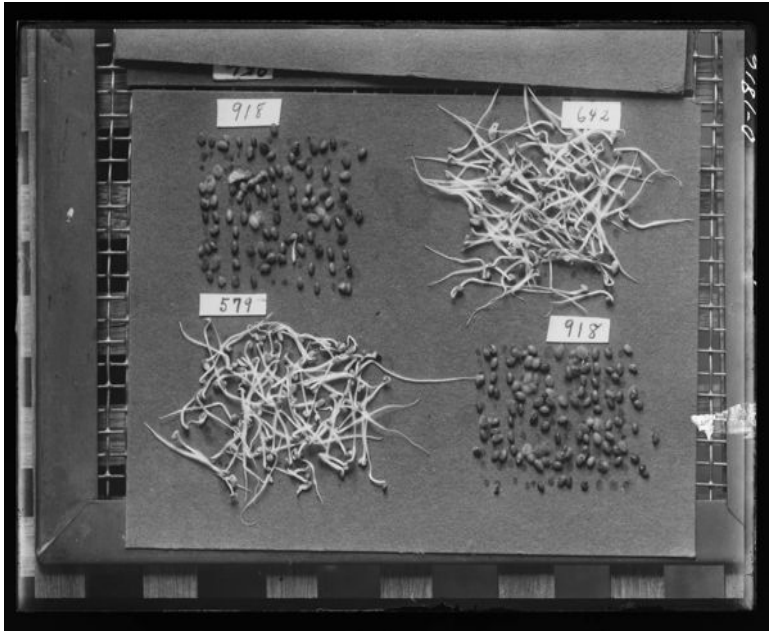
New York Public Library

Collection:
A group of URLs curated around a common theme, topic, or domain.

Ask Yourself…

- What is the **topic** of this collection?
- **Which websites** should I archive as part of this collection?

COLLECTIONS START WITH SEEDS



University of Kentucky

Seed:
The starting point URL for the crawler. The crawler will follow linked pages from the seed URL and archive them if they are “in scope.”

Document:
Any file with a distinct URL.
html, image, PDF, video, etc…

CRAWLERS AND SPIDERS AND ROBOTS, OH MY!

Département évangélique français d'action apostolique (Défap)

Crawlers are pieces of software that visit websites and index the information and files that construct them.

Scope:
What the crawler will capture and what it won't

Scoping:
Options for telling the crawler how much or how little of a seed to capture

HOW THE CRAWLER WORKS

Département évangélique français d'action apostolique (Défap)

1. Starts with seed URL(s)
2. Checks if those URLs are reachable, and archives them
3. Looks for embedded content – what does it need to render the page? *CSS, Javascript, Images, etc…*
4. Looks for links to other pages
5. Checks if those pages are "in scope" and archives them, if they are

The crawler will continue until it cannot locate any more links that are in scope or it hits a limit set for the crawl (time, data or document limits).

CRAWL SCOPING

How does the crawler know which links to archive and which to ignore?

> The seeds you add to your collection will determine the "scope" of your crawls.

> How you format your seed URLs can have an impact on the "scope" of your crawl.

University of Southern California Libraries
California Historical Society

ARCHIVE-IT SCOPING

University of Southern California Libraries
California Historical Society

Seeds can limit the crawl to a single directory of a site.

> example: www.archive.org/about/

> a / at the end of your url can have a big effect on scope

> Parts of the site not included in your seed directory will NOT be archived

Example seed: **www.archive.org/about/**

> *Link:* www.archive.org/webarchive.html **IS NOT** in scope

Example seed: **www.archive.org/about**

> *Link:* www.archive.org/webarchive.html **IS** in scope

ROBOTS.TXT BLOCKS

Los Angeles Public Library

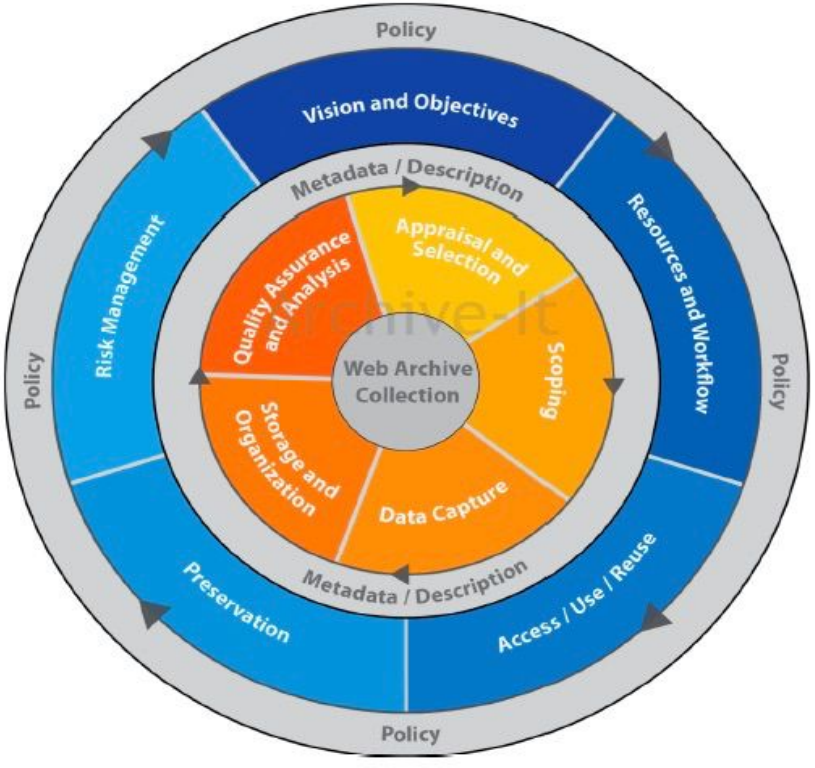
By default, our crawler *respects* all **robots.txt** files. Partners can check post-crawl reports for blocked seeds, hosts, or documents.

If your site is blocked…

> Contact the site owner and ask if they will unblock our crawler specifically – **archive.org_bot**

> Some institutions choose to utilize a tool to ignore robots.txt blocks for specific cases

WEB ARCHIVING LIFE CYCLE



- Highlights the policy and workflows of 6 partner institutions: Columbia, University of Alberta, Montana State Library, State Library of North Carolina, North Carolina State Archives and Creighton University

- Covers issues, including:
  - Policy
  - Vision and Objectives
  - Workflows
  - Access
  - Preservation

https://archive-it.org/static/files/archiveit_life_cycle_model.pdf

GROUP ACTIVITY

We will complete this activity in breakout groups (5 groups of 6-7 people per group)

Approximately 35 minutes have been allotted

Learning Objective: review key sections of the *The Web Archiving Life Cycle Model* and understand the areas necessary to consider when developing a web archiving program at your institution


QUESTIONS?
ASK THEM MEOW
imgflip.com

PART II AGENDA

TOOLS & STANDARDS

CHALLENGES

NEW TECHNOLOGIES

RESEARCH & ACCESS

WEB ARCHIVING TOOLS AT INTERNET ARCHIVE

Heritrix

Web crawler – crawls and captures web pages

Umbra

Assists the crawler to access social media and other sites in the same way a browser would

Wayback

Access tool for rendering and viewing pages - surf the web as it was

ElasticSearch & **SOLR**

Full-text search indexing engine & metadata search software

Brozzler

Browser + crawler= *Brozzler*!

WARC

ISO standard for storing web archives

WARC (Web ARChive) FORMAT

- ISO 28500:2009
- Combines multiple digital resources into an aggregate archival file together with related information
- Container file
- Written by crawlers
- Concatenated raw content
- For long-term storage and preservation

CHALLENGES: CONTENT

> **Social Media** – always improving tools for archiving Facebook, Twitter, Instagram and more.

> **Dynamic content** – some implementations can be difficult to capture and replay.

> **Streaming & Downloadable Media**

> **Password-protected Sites** - new feature in Beta

> **Forms and Databases -** alternatives may include a sitemap or direct links to content

WHAT MAKES A SITE ARCHIVABLE?

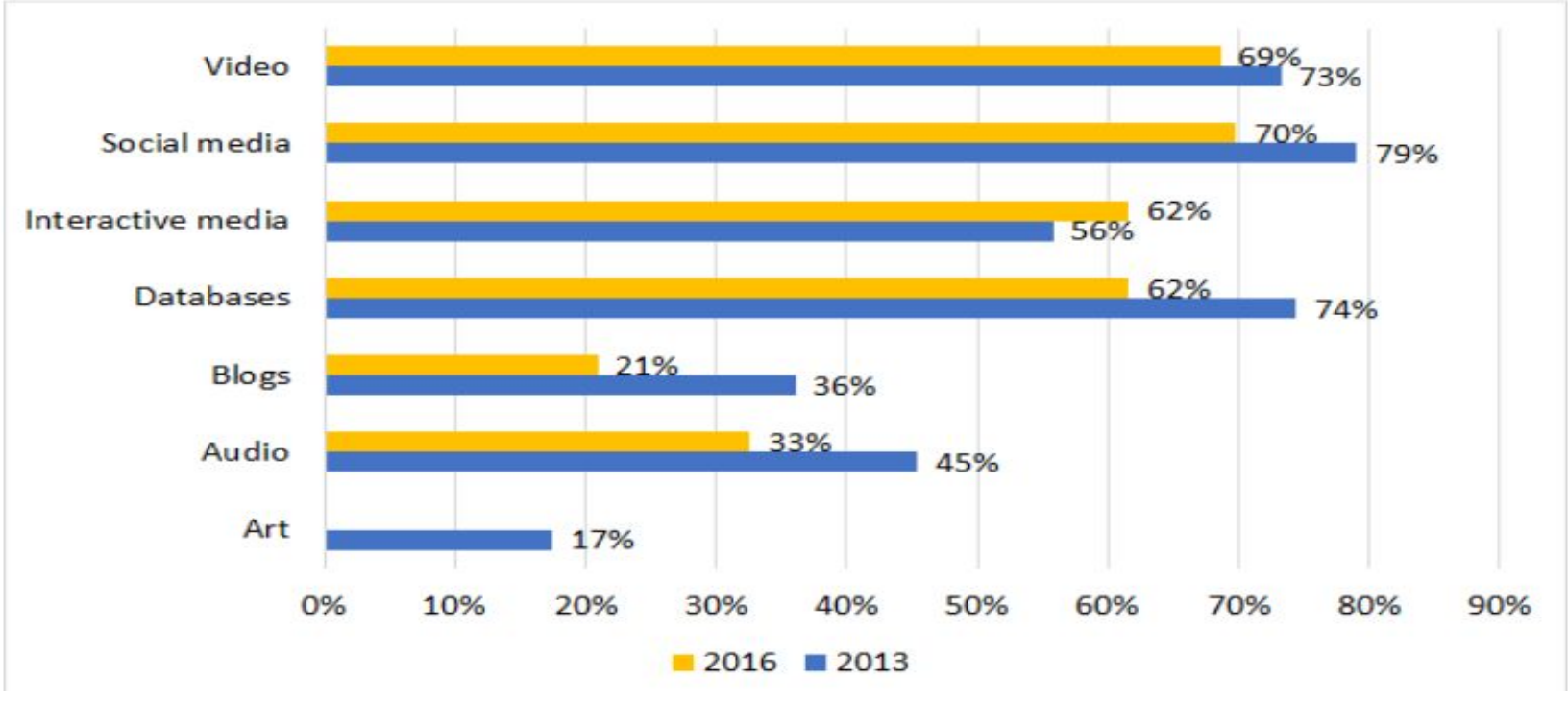
Make links transparent

Be careful with robots directives

Return reliable response codes

Many more suggestions in this Archive-It community resources post!

CHALLENGES: CONTENT



Type of content provoking concern over capacity to archive.

From *Web Archiving in the United States: A 2016 Survey, report from the National Digital Stewardship Alliance*

ARCHIVE-IT CRAWLING TECH TIMELINE

2003 - 2014 **Heritrix** → 2014 – Present **Heritrix + Umbra** → 2017 - Present **Brozzler**

Traditional web crawler

Scoping, capture, deduplication, WARC creation in one process

Less adept at triggering and capturing client side script and Javascript

ARCHIVE-IT CRAWLING TECH TIMELINE

2003 - 2014 | 2014 – Present | 2017 - Present

Heritrix → Heritrix + Umbra → Brozzler

Runs alongside Heritrix

Mimics the way a browser would access a page

Executes client side scripts so previously undetectable URLs could be accessed by Heritrix

Clicking or hovering to execute Javascript

Allows for dynamic scrolling

ARCHIVE-IT CRAWLING TECH TIMELINE

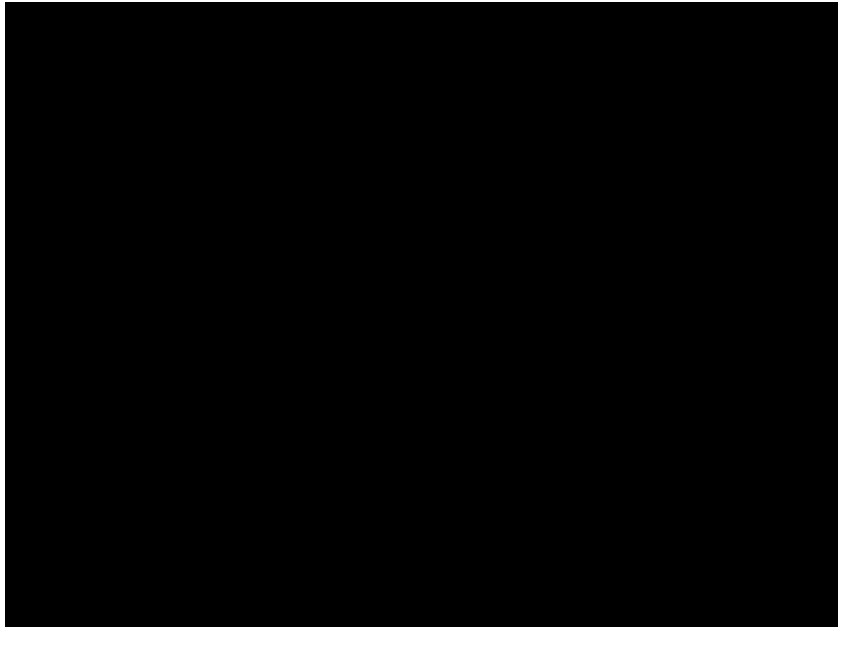
2003 - 2014 | 2014 – Present | 2017 - Present

Heritrix → Heritrix + Umbra → Brozzler

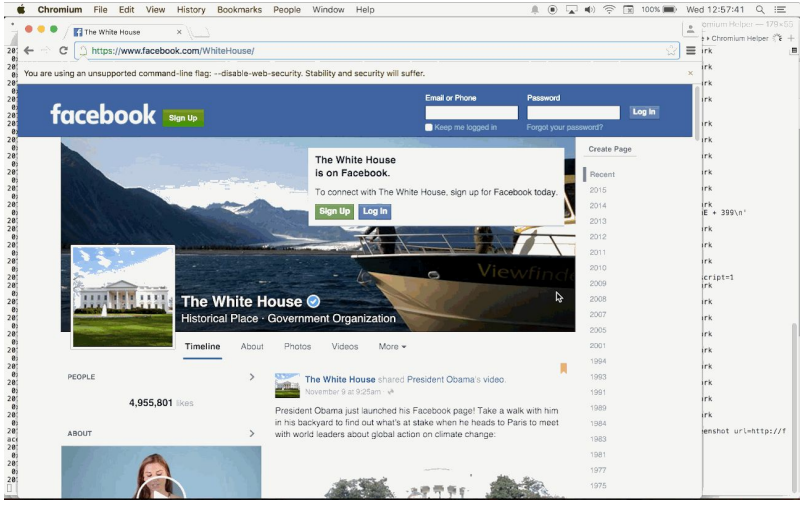
Captures http traffic as it's loaded

Uses a real browser to fetch pages and embedded URLs, and to extract links

Works with youtubedl to improve media capture



BROZZLER



"browser"|"crawler" = BROZZLER

Runs on an instance of chromium browser

Opens page in the browser, takes a screenshot, sends to warcprox, written as a WARC file

Runs a javascript behavior and finds a@href outlinks

WEB ARCHIVING TOOLS & SERVICES

- Services & Tools
  - Archive-It
  - Internet Memory Foundation
  - Commercial: Hanzo, Pagefreezer, Mirrorweb
  - Tools: Webrecorder, WAIL
  - API based: twarc, Social Feed Manager

- Access:
  - Oldweb.today, Memento, Webrecorder

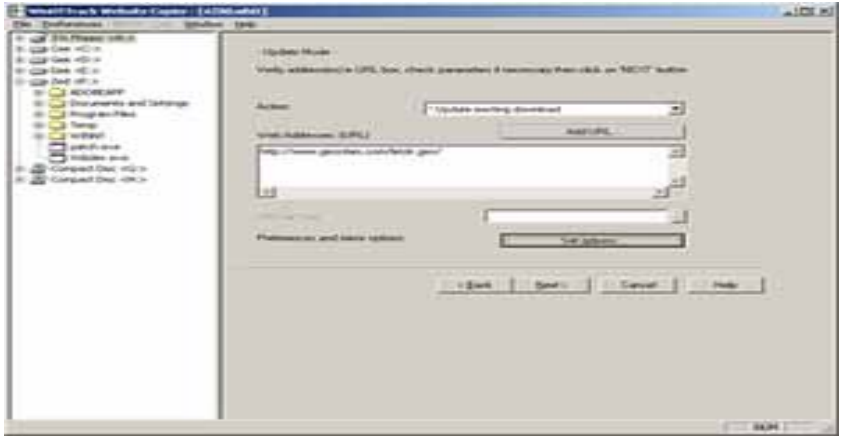
For a full list: http://netpreserve.org/web-archiving/tools-and-software/

WEB ARCHIVING TOOLS: HTTRACK & WGET

HTTRACK

http://www.httrack.com/

Download a site from the Internet to a local directory



WGET

> Terminal tool

> Allows you to download directories via the command line

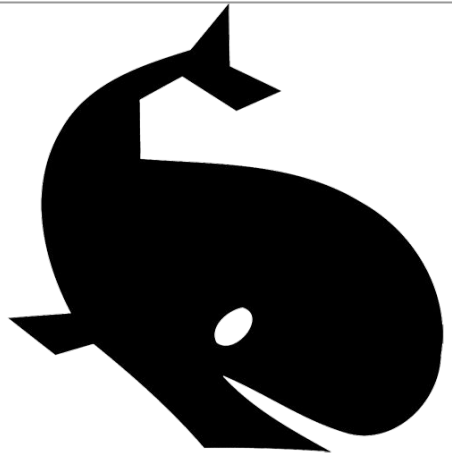
> http://ftp.gnu.org/gnu/wget/

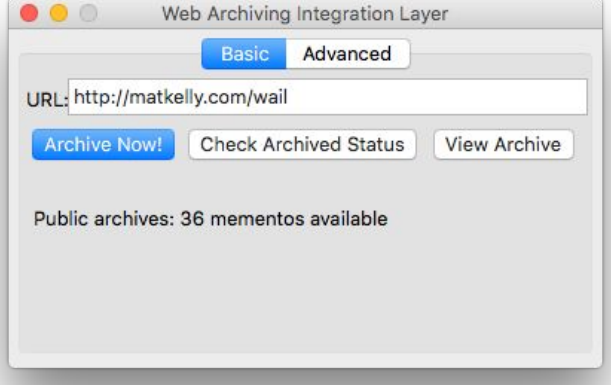
WEB ARCHIVING TOOLS: WAIL

"Web Archiving Integration Layer (WAIL) is a graphical user interface (GUI) atop multiple web archiving tools intended to be used as an easy way for anyone to preserve and replay web pages."

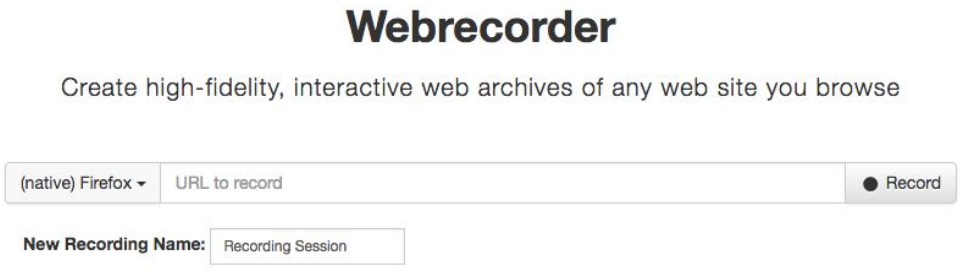
Uses Heritrix 3.2.0 and OpenWayback 2.3.0

Developed by Mat Kelly at the Web Science and Digital Libraries research group at Old Dominion University





WEB ARCHIVING TOOLS: WEBRECORDER



Available at webrecorder.io/

Developed by Rhizome

Focus on dynamic web content such as embedded video and complex javascript

Relies on a human user browsing the live web

Provides for both capture and access

SOCIAL MEDIA WEB ARCHIVING TOOLS & SERVICES

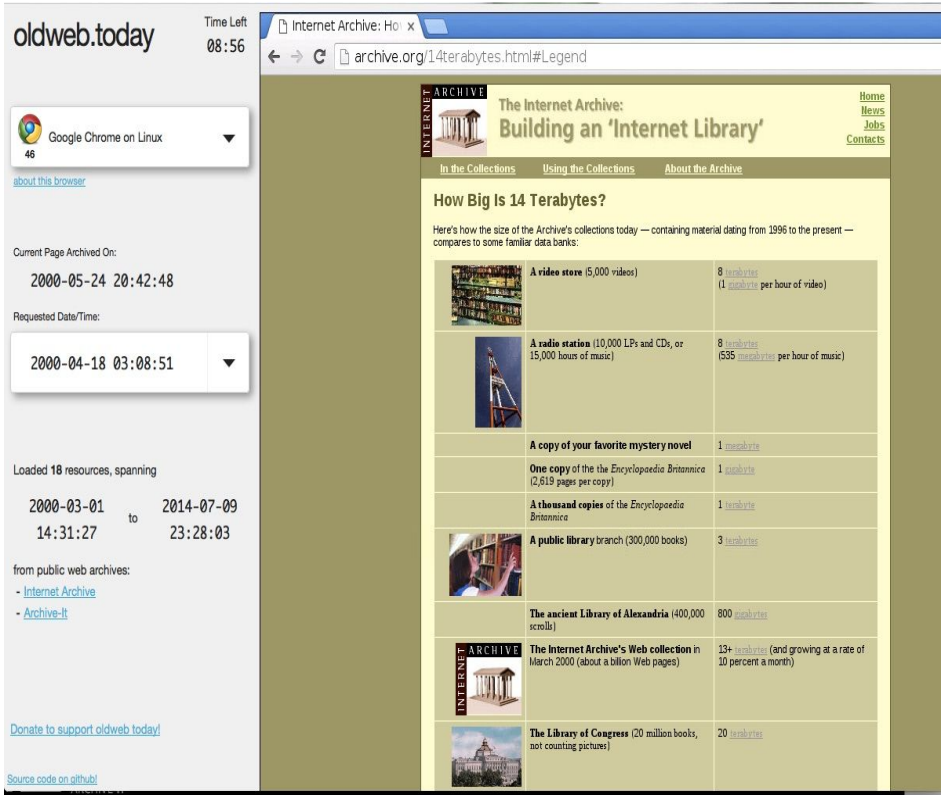
Social Feed Manager

- ★ Developed at George Washington University Libraries
- ★ Collects Twitter data in bulk using the Twitter API
- ★ Open source; available on github

twarc

- ★ Developed by Ed Summers at Maryland Institute for Technology in the Humanities
- ★ A command line tool for archiving Twitter JSON data
- ★ Also uses the Twitter API
- ★ Useful for running searches on terms to collect all tweets mentioning a keyword

ACCESS: OLDWEB.TODAY



- ★ Browser emulator to access publicly available archived sites using virtual versions of old browsers
- ★ Focus is on playing back the site as it would have been originally experienced
- ★ Developed at Rhizome

ACCESS:
MEMENTO/TIME TRAVEL SERVICE

http://timetravel.mementoweb.org/

Chrome plug in allows you to navigate between the present web and the web of the past

archive.today, Archive-It, Arquivo.pt: the Portuguese Web Archive, Bibliotheca Alexandrina Web Archive, DBpedia archive, DBpedia Triple Pattern Fragments archive, Canadian Government Web Archive, Croatian Web Archive, Estonian Web Archive, Icelandic web archive, Internet Archive, Library of Congress Web Archive, NARA Web Archive, National Library of Ireland Web Archive, perma.cc, PRONI Web Archive, Slovenian Web Archive, Stanford Web Archive, UK Government Web Archive, UK Parliament's Web Archive, UK Web Archive, Web Archive Singapore, WebCite, Bayerische Staatsbibliothek

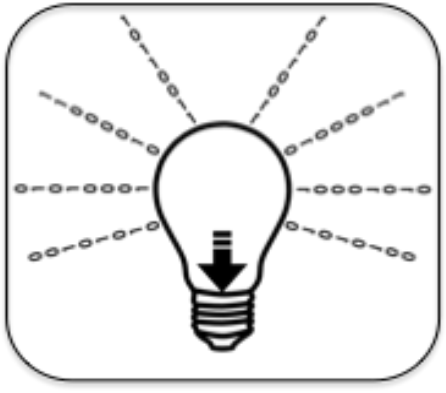
RESEARCH SERVICES & WEB ARCHIVE DATASETS

TYPOLOGY OF RESEARCHER INTERESTS

- ★ **Documentary:** Evidentiary, Attestation, Legal discovery/claim
- ★ **Social/Political Scientists:** Communications, Politics/Government, Social Anthropology
- ★ **Web Science:** Technology Systems and Protocols
- ★ **(Digital) Humanities:** Historians and humanities disciplines, networks, collection building
- ★ **Computer Science:** Information Retrieval, Data Processing and Indexing, Infrastructure and tools
- ★ **Data Analysts:** Mining/Training, language processing, trend analysis

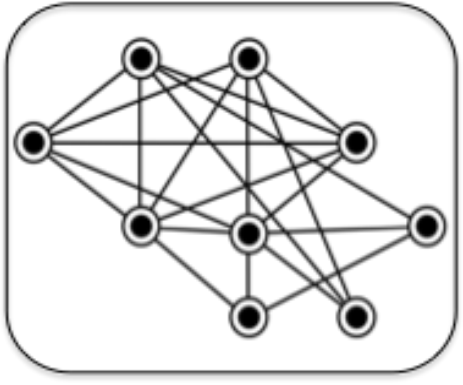
RESEARCH SERVICES & WEB ARCHIVE DATASETS

Web Archive Datasets



WAT Datasets

(Web Archive Transformation)
Key Metadata from Every Resource



LGA Datasets

(Longitudinal Graph Analysis)
What Links to What over Time

WANE Datasets

(Web Archive Named Entities)
Names of People, Places, Organizations

RESEARCH SERVICES & WEB ARCHIVE DATASETS

LESSONS LEARNED SUPPORTING RESEARCHERS

- ★ Researchers don't always know what they want
- ★ Researchers default to wanting access to raw/all data
- ★ Researchers will have varying levels of technical resources or support
- ★ Address upfront issues of technical proficiency, non-archive technical support and/or methodological stuff
- ★ Will require reference/resources to explain and contextualize web archive tools and processes
- ★ More data doesn't equal better analysis

RESEARCH SERVICES & WEB ARCHIVE DATASETS

STRATEGIC APPROACHES TO SUPPORTING RESEARCHERS

- ★ Focus on derivation, portability, and access
- ★ Focus on scalable partnerships & decentralization
- ★ Research support expectations often != with available resources or services
- ★ Research methodologies (conceptual, practical, technical) often != with data, collecting, tools
- ★ Service models or death (though yet to emerge for most data-driven LAM-ish research)

RECAP AND REVIEW

- ★ **Definitions and Community**
  - We defined terms, practices, and the current landscape
- ★ **Collection**
  - We discussed the whys and hows of creating web archives
- ★ **Manage**
  - We outlined management, tools, and services
- ★ **Use**
  - We looked at formats, archival replay, and data mining

- ★ **Now We'll Demo It All!**
- ★ **Then It's Your Turn To Archive!**

READING LIST

Davis, Corey. "Archiving the Web: A Case Study from the University of Victoria." Code4Lib Issue 26, 2014-10-21 (2014). http://journal.code4lib.org/articles/10015

Keeping Collections: More Podcast Less Process. Episode 007. The Web Archivist Are Present. http://keepingcollections.org/more-podcast-less-process-episode-007/

D-Lib Magazine. Special Issue on Web Archives. March/April 2012. http://www.dlib.org/dlib/march12/03contents.html

Web Archiving In The United States: A 2016 Survey. NDSA, 2017. Web. 10 May 2017. Results Of A Survey Of Organizations Preserving Web Content, http://ndsa.org/documents/WebArchivingintheUnitedStates_A2016Survey.pdf

Summers, Ed. 'The Web As A Preservation Medium'. *inkdroid*. N.p., 2013. http://inkdroid.org/journal/2013/11/26/the-web-as-a-preservation-medium/

Pennock, M. (2013, March). Web-archiving (DPC Technology Watch Report 13-01). Digital Preservation Coalition. Retrieve from http://dx.doi.org/10.7207/twr13-01

Taylor, Nicholas. 'Anatomy Of A Web Archive | The Signal: Digital Preservation'. N.p., 2013. Web. 30 Jan. 2015. http://blogs.loc.gov/digitalpreservation/2013/11/anatomy-of-a-web-archive/

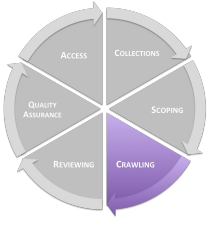
ADDITIONAL RESOURCES: support.archive-it.org

Collections:

Collection Management Overview
Managing Metadata
Managing Seeds

Scoping:

How the crawler determines "Scope"
Seed Types
Seed vs. collection level scoping
Identify and avoid crawler traps
Scoping for specific types of sites

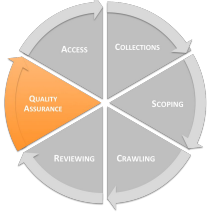


Crawling:

Scheduling recurring crawls
Starting One-Time or Test crawls
Saving Test crawls

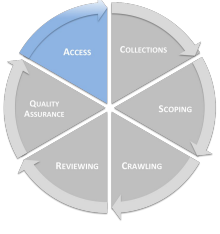
Reviewing:

Reviewing captures
Reading your crawl report



Quality Assurance:

Wayback QA
Quality assurance from the Host Report
Using Proxy mode



Access:

Through your Archive-It account
Through Archive-it.org
Through other domains
Downloading WARC files


I HAZ A QUESTION

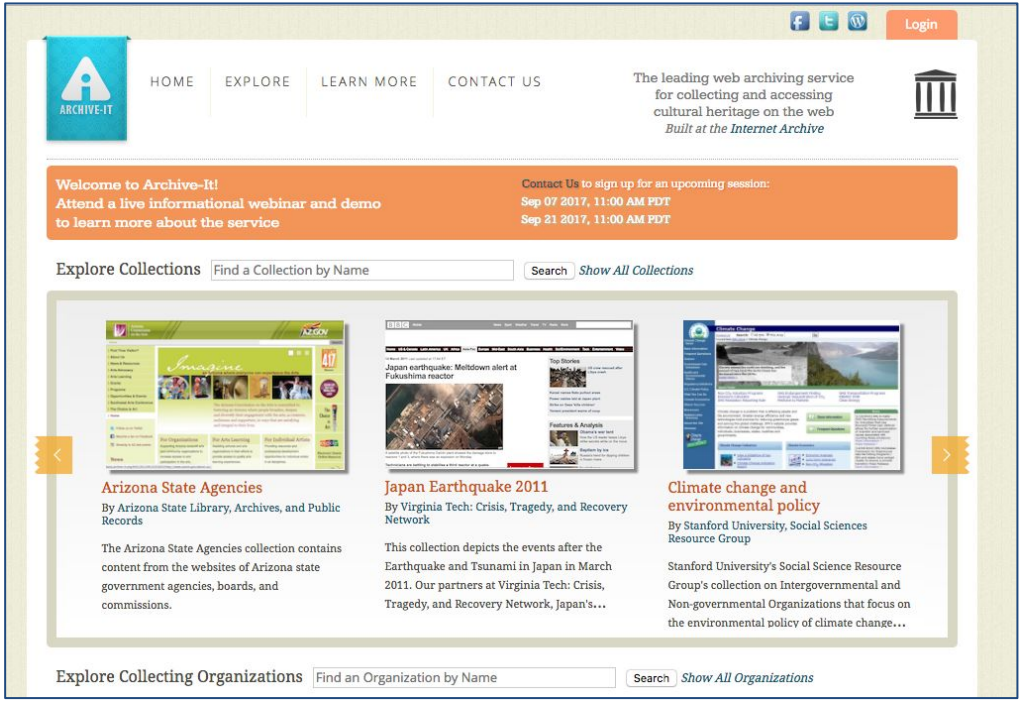
ARCHIVE-IT DEMO

Archive-It
https://archive-it.org/

Click “Login” in upper right

Login details:

- Username: disworkshop
- Password: dis2018



Lori Donovan, Senior Program Manager, Web Archiving | lori@archive.org
Mary Haberle, Web Archivist | mhaberle@archive.org
Internet Archive & Archive-It | @internetarchive & @archiveitorg