ALL ABOUT WEB ARCHIVING

Hosted by ALCTS, the association for Library Collections and Technical Services
Jillian Lohndorf
Web Archivist, Internet Archive
Agenda

Section 1: The what and how of web archiving
Section 2: What is in web archives and where are web archives found?
Section 3: Issues of arrangement and description
Internet Archive is a non-profit Digital Library founded in 1996 with 45+PB of unique data: web archives, text, video, audio, software
THE WAYBACK MACHINE

Online: [https://archive.org/web/](https://archive.org/web/)

The largest publicly available web archive in existence.

> 330 Billion webpages
~ 100 million pages captured per day
WHAT IS WEB ARCHIVING?

What is a web archive?
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 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.
Heritrix
HTTrack
Wget
Brozzler

ARC
WARC

Wayback Machine
OpenWayback
pywb (Python Wayback)
oldweb.today
WHAT DO WEB ARCHIVES LOOK LIKE?

https://www.webarchive.org.uk/wayback/archive/20100813132532/http://www.aboutolympics.co.uk/
WHAT GOES IN A WEB ARCHIVE?

**Document a subject area or web-only conversations or events**
- 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 content**
- Support an electronic records system to meet records retention requirements
- Collect publications/documents that are no longer in print form
- Construct a historical record of an institution or individual's web/social media presence

**Content about to leave the web**
- Before a site or social media goes through a major redesign
- Before a site or social media comes down from the web altogether
WHAT GOES IN A WEB ARCHIVE?
WHAT GOES IN A WEB ARCHIVE?
WHERE TO FIND WEB ARCHIVES

The Wayback Machine: https://archive.org/web/
WHERE TO FIND WEB ARCHIVES

British Library
WHERE TO FIND WEB ARCHIVES

Web Archiving services, like Archive-It: https://archive-it.org/
Custom Portals: NARA
WHERE TO FIND WEB ARCHIVES

Library of Congress > Programs at the Library of Congress > Web Archiving > Archived Web Sites

PROGRAM
Web Archiving

About This Program  For Researchers  For Site Owners  Archived Web Sites

Archived Web Sites

Related Resources
The Signal
Collection Policy Statements and Supplementary Guidelines
Recommended Formats Statement for Web Sites

Archived Web Sites

Browse archived web sites by collection topic. Additional web archives may be searched by visiting Search Archived Web Sites.

COLLECTION
Papal Transition 2005 Web Archive

Collection Period: April 5, 2005 to November 30, 2005. On April 2, 2005, Pope John Paul II died at the age of 84; seventeen days later, on April 19, 2005, Pope Benedict XVI was elected the 266th pope of the Roman Catholic Church. The death of a presiding pope of the Catholic Church is an event that affects more than one billion Catholics worldwide...

Contributor: Library of Congress
Date: 2008
Collection Items: View 209 Items

Library of Congress
WHERE TO FIND WEB ARCHIVES

End of Term Web Archive
US Federal Web Domain at Presidential Transitions


End of Term at a Glance:
Crawl dates:
September 2008 to May 2009
September 2012 to March 2013
September 2016 to March 2017
Number of websites captured:
2008: 3,305 websites
2012: 3,273 websites
2016: 53,024 websites

Archive Scope
The End of Term Web Archive contains federal government websites (.gov, .mil, etc.) in the Legislative, Executive, or Judicial branches of the government. Websites that were at risk of changing (i.e., whitehouse.gov) or disappearing altogether during government transitions were captured. Local government websites, or any other site not part of the federal government domain were out of scope.
WHERE TO FIND WEB ARCHIVES

California State Government : ca.gov web archive

Author: California Digital Library; University of California (System). Regents.
Publisher: [Oakland, CA?] : Regents of the University of California
Summary: Provides archives of content published by the State of California, and provides access to California state agencies' web sites.
Rating: (not yet rated) 0 with reviews - Be the first.
Subjects: Web archives -- California, California State Archives, Web archives.

More like this: Similar Items

Find a copy in the library
WHERE TO FIND WEB ARCHIVES

Catalogs, like NYARC (with full text search catalog integration)
WHERE TO FIND WEB ARCHIVES

- Integrations
  - Catalog
  - Website
  - Institutional Repository
  - Worldcat
  - Custom (can you use an API?)

ArchivesSpace- Archive-It integration
QUESTIONS?

St. Cloud State University
Metadata

n. ~ Information that characterizes or describes an information artifact, record, or archival resource.

[SAA Dictionary, forthcoming]
DESCRIBING WEB ARCHIVES

How do web archives fit into digital records?

How is metadata for web archives different than other kinds of metadata?

Libraries principally (but far from exclusively) catalog individual resources at the title level, and many resources are expected to, in effect, describe themselves via features such as a title page and verso (or CD label, sheet music cover or film credits) from which data elements can be transcribed. The nature of the content is revealed principally by the title (if it is descriptive) and subject terms, though notes describing content have become more common over time. Digital content that is openly accessible is considered published.

In contrast, archivists generally describe a collection of unpublished materials that are related by provenance in a single metadata record or multilevel finding aid. They routinely take descriptive information from sources both internal and external to the collection (such as biographical or historical secondary material). Titles are devised rather than transcribed. Extensive free-text notes are routinely used to describe both content and context of the material. Archival content that is born-digital or digitized is not considered to be published when made available online.
At the macro level...

OCLC Web Archiving Metadata Working Group (WAM) has recommended metadata standards: https://tinyurl.com/OCLC-WAM

- Based on Dublin Core
  - Structure standard- not a content standard
  - Does not include guidance for devising the content of elements
  - “...the name and meaning of eight of the 15 DC elements match the WAM data dictionary: Contributor, Creator, Date, Description, Language, Relation, Subject and Title.” -WAM
  - No controlled vocabulary
- Meant to work with RDA/MARC, DACS, EAD, MODS, schema.org
**Collector:** The organization responsible for curation and stewardship of an archived website or collection

**Contributor:** An organization or person secondarily responsible for the content of an archived website or collection

**Creator:** An organization or person principally responsible for creating the intellectual content of an archived website or collection

**Date:** A single date or span of dates associated with an event in the lifecycle of an archived website or collection

**Description:** One or more notes explaining the content, context and other aspects of an archived website or collection

**Extent:** An indication of the size of an archived website or collection

**Genre/form:** A term specifying the type of content in an archived website or collection

**Language:** The language(s) of the archived content, including visual and audio resources with language components

**Relation:** Used to express part/whole relationships between a single archived website and any collection to which it belongs

**Rights:** Statements of legal rights and permissions granted by intellectual property law or other legal agreements

**Source of description:** Information about the gathering or creation of the metadata itself, such as sources of data or the date on which source data was obtained

**Subject:** Primary topic(s) describing the content of an archived website or collection

**URL:** Internet address for an archived website or collection
Where to add metadata
Account level
Collection level
Seed level
Document level

A Crawl is the primary unit of collection activity

- A Crawl is logically equivalent to an archival accession
- Each Crawl contains one or more Seeds
- All Seeds in a crawl must be from the same Collection
- Each Crawl is assigned a unique identifying number by Archive-It
- A Crawl is primarily defined by:
  - Date and time of start
  - Seeds included
  - Any constraints or expansions in place during
<table>
<thead>
<tr>
<th>Metadata</th>
<th>Collection Metadata</th>
<th>Bulk Seed Metadata</th>
<th>Bulk Document Metadata</th>
<th>Single Document Metadata</th>
</tr>
</thead>
</table>

### Collection Metadata

**Title**: A collection of websites from the professional digital preservation community.

**Publisher**: Add

**Creator**: Add

**Subject**: Add

**Description**: A collection of websites from the professional digital preservation community.

**Language**: English

**Rights**: Add

**Collector**: Add

**Language**: Add
California Legislature (152)
California Legislature - Assembly (175)
California Legislature - Senate (88)
California Public Colleges and Universities (10)
California State Agencies (314)

Subject
California Legislature, Assembly (77)
California Legislature, Senate (34)
California Energy Commission (18)
Climatic changes (15)
Legislators (10)

Creator
California Legislature, Assembly (43)
California State Assembly Democratic Caucus (38)
California Climate Change Center (27)
California Energy Commission (24)
Senate Republican Caucus (Calif.) (24)

Title: California Commission for Jobs and Economic Growth

URL: http://4cajobs.com/

Description: The Commission works side-by-side with state and local government, business associations, unions, educational institutions and California’s network of regional and local Economic Development Corporations (EDCs) to: Identify, illustrate and suggest ways to remove barriers to increasing employment and doing business in California. Identify and assist employers interested in expanding in California or at risk of leaving the state. Help market California products, services and destinations to national and international audiences.

Captured 9 times between Oct 26, 2008 and Sep 28, 2011

Subject: California Commission for Jobs and Economic Growth
Creator: California Commission for Jobs and Economic Growth
Language: English
Coverage: California

Note: Two most recently archived pages from 10/7/2010 and 9/28/2011 were captured during scheduled maintenance. Full text of the captured page reads: "CA Jobs

Subject: Job creation
Subject: Economic development
Some of the factors that go into metadata decisions:

- Other metadata practices
- If/how web archives are integrated into other systems or services
- Overall access goals
  - Cataloging outside of a web archiving service
- Staff time
FUTURE CONSIDERATIONS

Where can we go from here?

● Controlled vocabulary
● Automating metadata creation
  ○ Archive-It’s grab title functionality
  ○ Zotero the citation service, to grab metadata, then edit (like copycataloging!)
● More options, integrations, customization
LEARN MORE

Follow our Blog: https://blog.archive.org/
https://archive-it.org/blog

Follow us on Twitter: @internetarchive
@archiveitorg

Like us on Facebook: https://www.facebook.com/internetarchive/
https://www.facebook.com/ArchivelT

Questions: ait@archive.org
jillian@archive.org