LIBRARY OF CONGRESS

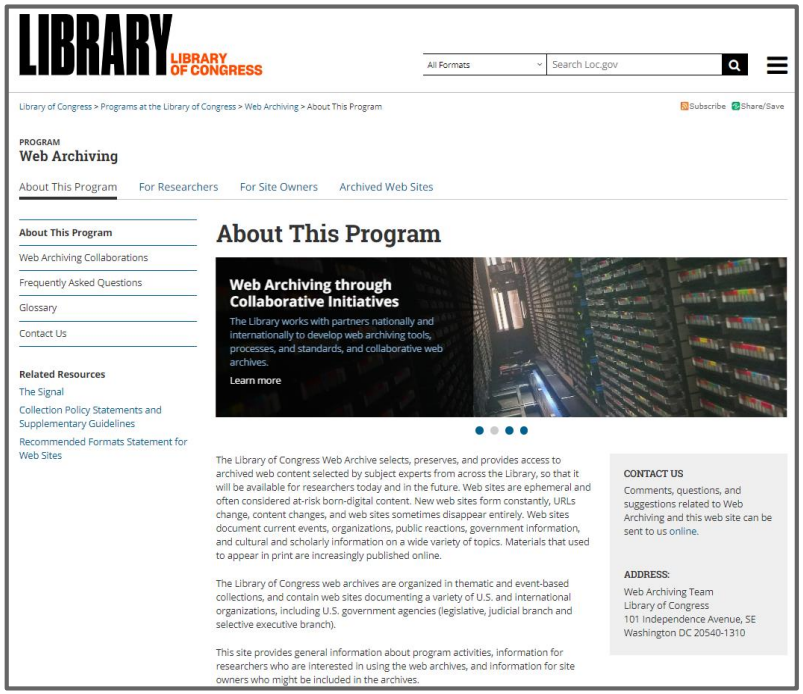
IIPC WAC June 7, 2019

Access to the Library of Congress Web Archives

Abbie Grotke
Lead Librarian, Web Archiving Team
abgr@loc.gov
@agrotke

Library of Congress Web Archiving

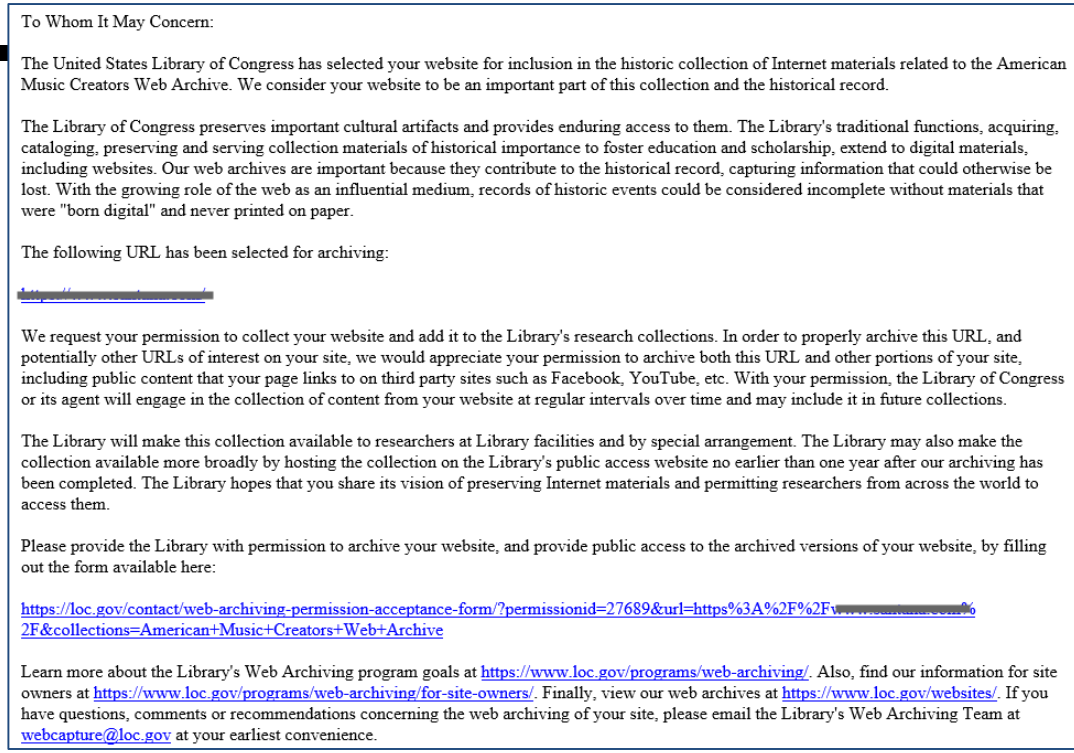
- Collecting since 2000
  - some .gov content acquired from IA back to 1996
  - 1.9 PB of content to date
- Event and thematic archives
  - Over 120 collections; about 65 active
  - Collection topics include:
    - U.S. legislative branch and select executive branch government
    - U.S and foreign elections
    - Government publications (U.S. and foreign)
    - Legal/Law Blawgs
    - Political science/public policy topics
    - Business and economics
    - Folklore and folklife
    - Comics and cartoons
    - Recent editions: music and the recording industry, Library of Congress composers, LGBTQ+ studies, food and recipes, authors, etc.



https://www.loc.gov/programs/web-archiving/about-this-program/

Policies Affecting Access

- No legal deposit or legal mandate
- Two-pronged permissions-based approach: **permission or notice to crawl**, **permission or notice to display offsite**
- Not all sites treated the same: mixed permissions depending on **country of publication** and **category of website**
- A few examples:
  - A **U.S. government website** would be **No Notice** (no email sent)
  - A **candidate** running for office in the **United States** would be a **Notice/Notice** (notify of intent to crawl and to provide offsite access)
  - A **journalism** site published in **Thailand** is a **Notice/Permission** (notify of intent to crawl, request permission provide offsite access)
  - An **online comic** published anywhere ("**creative expression**") would be **Permission/Permission** anywhere (Permission required to crawl and provide offsite access).
- Everything is under a one-year embargo
- Permissions managed using Digiboard (inhouse tool) and OpenWayback exclude files



To Whom It May Concern:

The United States Library of Congress has selected your website for inclusion in the historic collection of Internet materials related to the American Music Creators Web Archive. We consider your website to be an important part of this collection and the historical record.

The Library of Congress preserves important cultural artifacts and provides enduring access to them. The Library's traditional functions, acquiring, cataloging, preserving and serving collection materials of historical importance to foster education and scholarship, extend to digital materials, including websites. Our web archives are important because they contribute to the historical record, capturing information that could otherwise be lost. With the growing role of the web as an influential medium, records of historic events could be considered incomplete without materials that were "born digital" and never printed on paper.

The following URL has been selected for archiving:

[redacted]

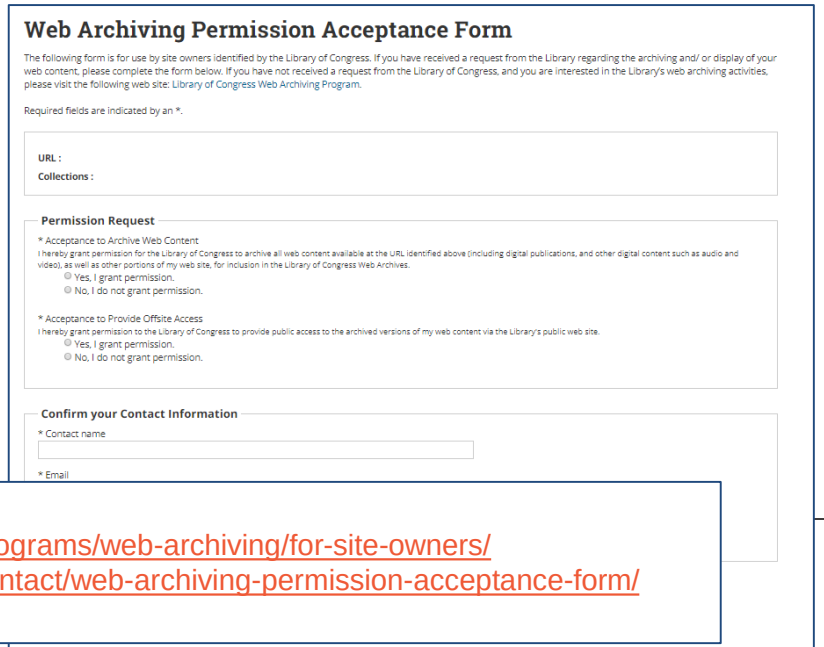
We request your permission to collect your website and add it to the Library's research collections. In order to properly archive this URL, and potentially other URLs of interest on your site, we would appreciate your permission to archive both this URL and other portions of your site, including public content that your page links to on third party sites such as Facebook, YouTube, etc. With your permission, the Library of Congress or its agent will engage in the collection of content from your website at regular intervals over time and may include it in future collections.

The Library will make this collection available to researchers at Library facilities and by special arrangement. The Library may also make the collection available more broadly by hosting the collection on the Library's public access website no earlier than one year after our archiving has been completed. The Library hopes that you share its vision of preserving Internet materials and permitting researchers from across the world to access them.

Please provide the Library with permission to archive your website, and provide public access to the archived versions of your website, by filling out the form available here:

https://loc.gov/contact/web-archiving-permission-acceptance-form/?permissionid=27689&url=https%3A%2F%2F[redacted]%2F&collections=American+Music+Creators+Web+Archive

Learn more about the Library's Web Archiving program goals at https://www.loc.gov/programs/web-archiving/. Also, find our information for site owners at https://www.loc.gov/programs/web-archiving/for-site-owners/. Finally, view our web archives at https://www.loc.gov/websites/. If you have questions, comments or recommendations concerning the web archiving of your site, please email the Library's Web Archiving Team at webcapture@loc.gov at your earliest convenience.



Web Archiving Permission Acceptance Form

The following form is for use by site owners identified by the Library of Congress. If you have received a request from the Library regarding the archiving and/ or display of your web content, please complete the form below. If you have not received a request from the Library of Congress, and you are interested in the Library's web archiving activities, please visit the following web site: Library of Congress Web Archiving Program.

Required fields are indicated by an *.

URL :
Collections :

Permission Request

* Acceptance to Archive Web Content
I hereby grant permission for the Library of Congress to archive all web content available at the URL identified above (including digital publications, and other digital content such as audio and video), as well as other portions of my web site, for inclusion in the Library of Congress Web Archives.
Yes, I grant permission.
No, I do not grant permission.

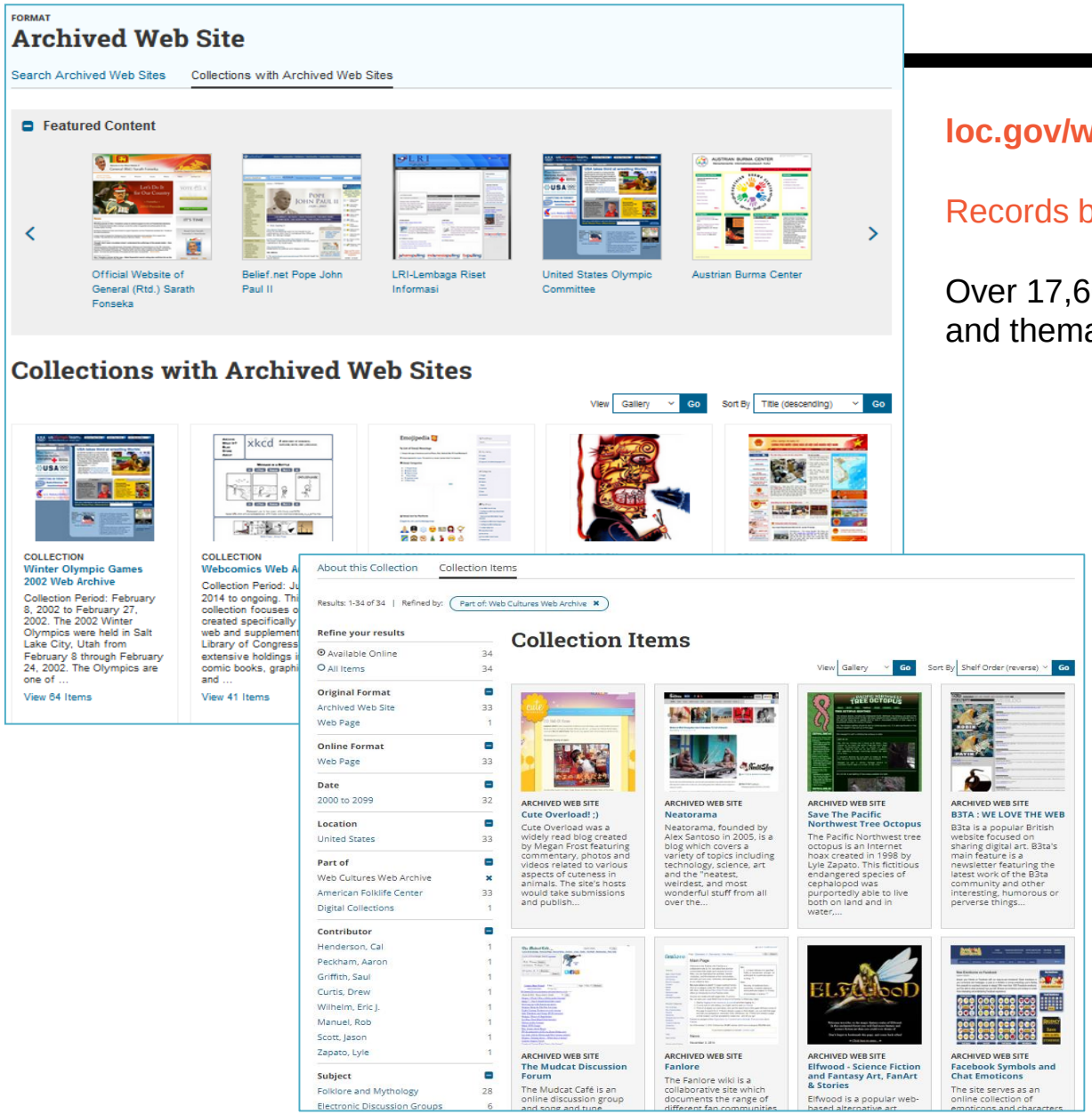
* Acceptance to Provide Offsite Access
I hereby grant permission to the Library of Congress to provide public access to the archived versions of my web content via the Library's public web site.
Yes, I grant permission.
No, I do not grant permission.

Confirm your Contact Information

* Contact name

* Email

https://www.loc.gov/programs/web-archiving/for-site-owners/
https://www.loc.gov/contact/web-archiving-permission-acceptance-form/



loc.gov/websites/collections

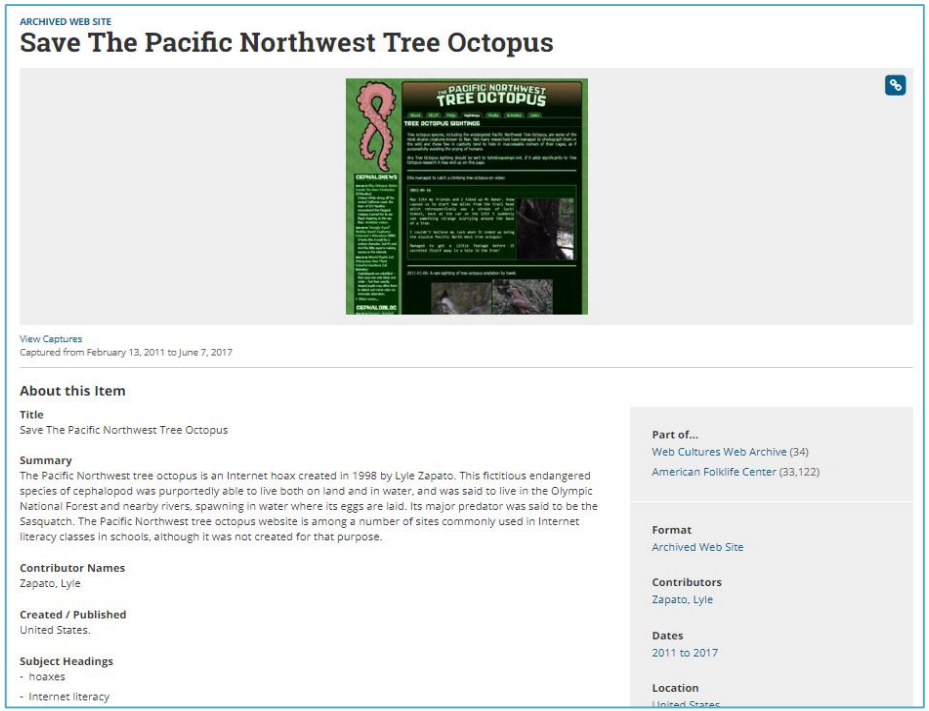
Records browse and search: **loc.gov/websites**

Over 17,600 web archives from over 60 event and thematic collections

Tackling the backlog:

Only 33 collections with contextual material at the collection level.

By September 2019 should have rolling updates of records coming out of embargo.

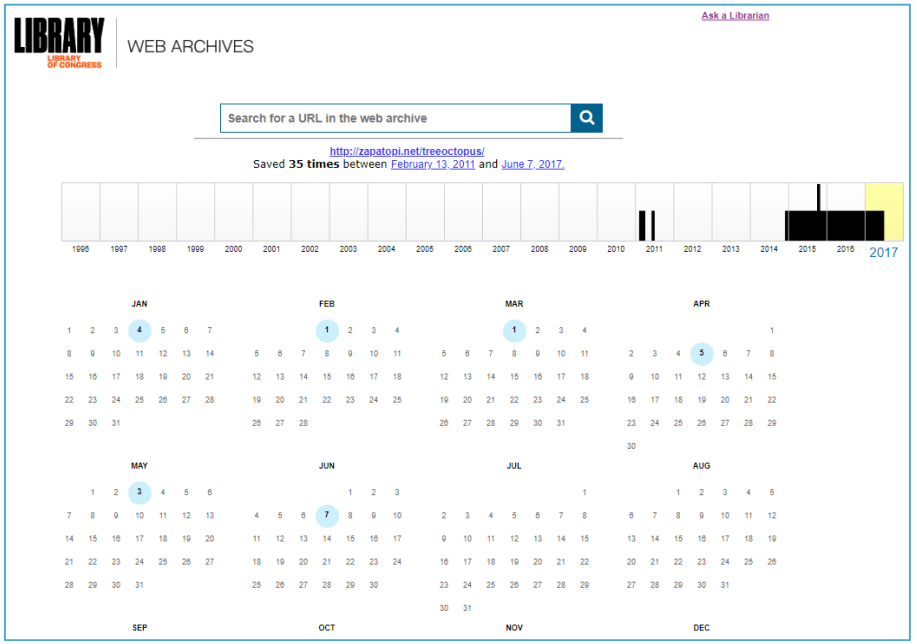


Record search at loc.gov

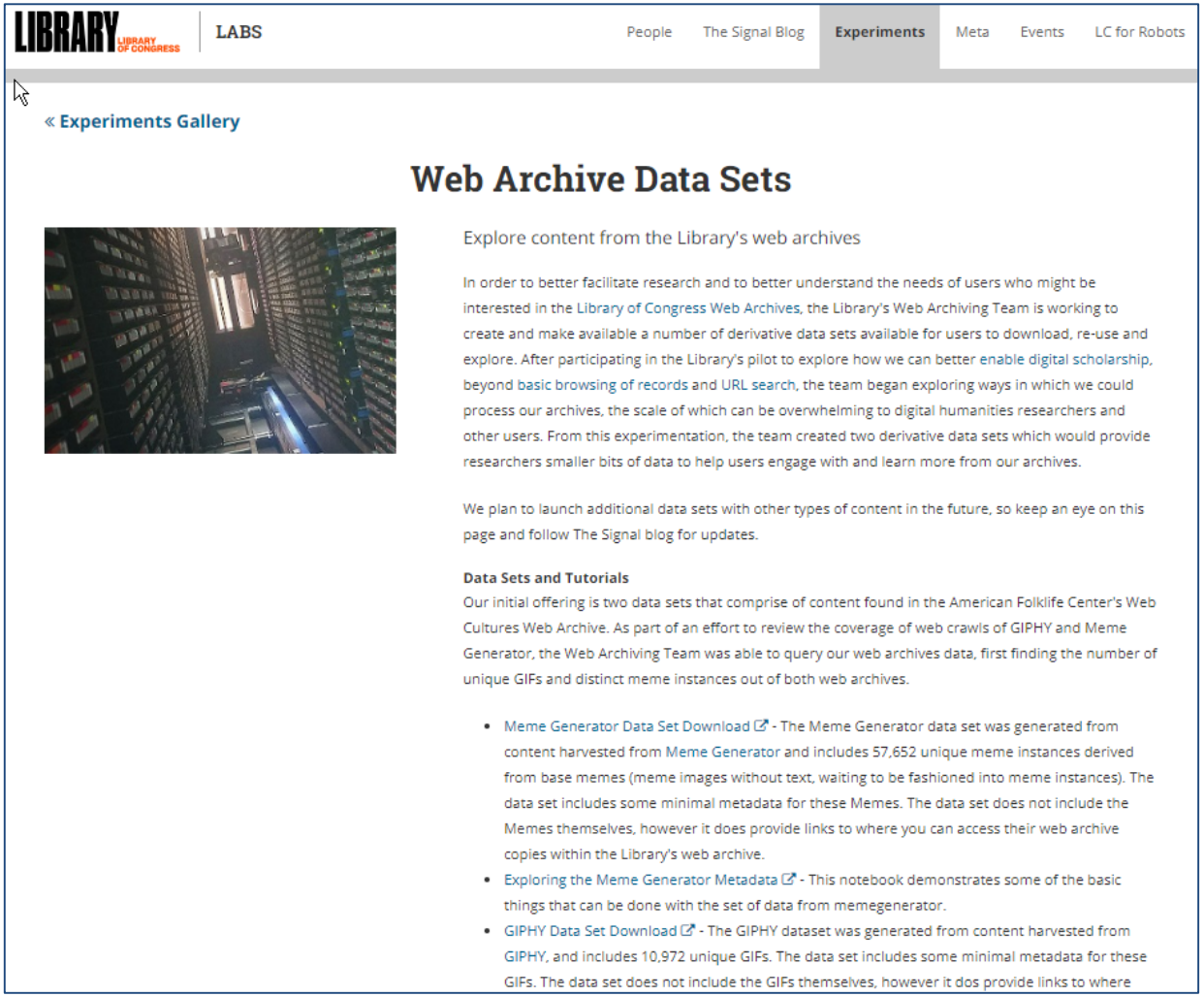
records describing each organization archived are searchable alongside other Library digital items

URL search at Webarchive.loc.gov →

A local instance of OpenWayback is used to display the archived content (including uncatalogued sites) currently up through **May 2018**



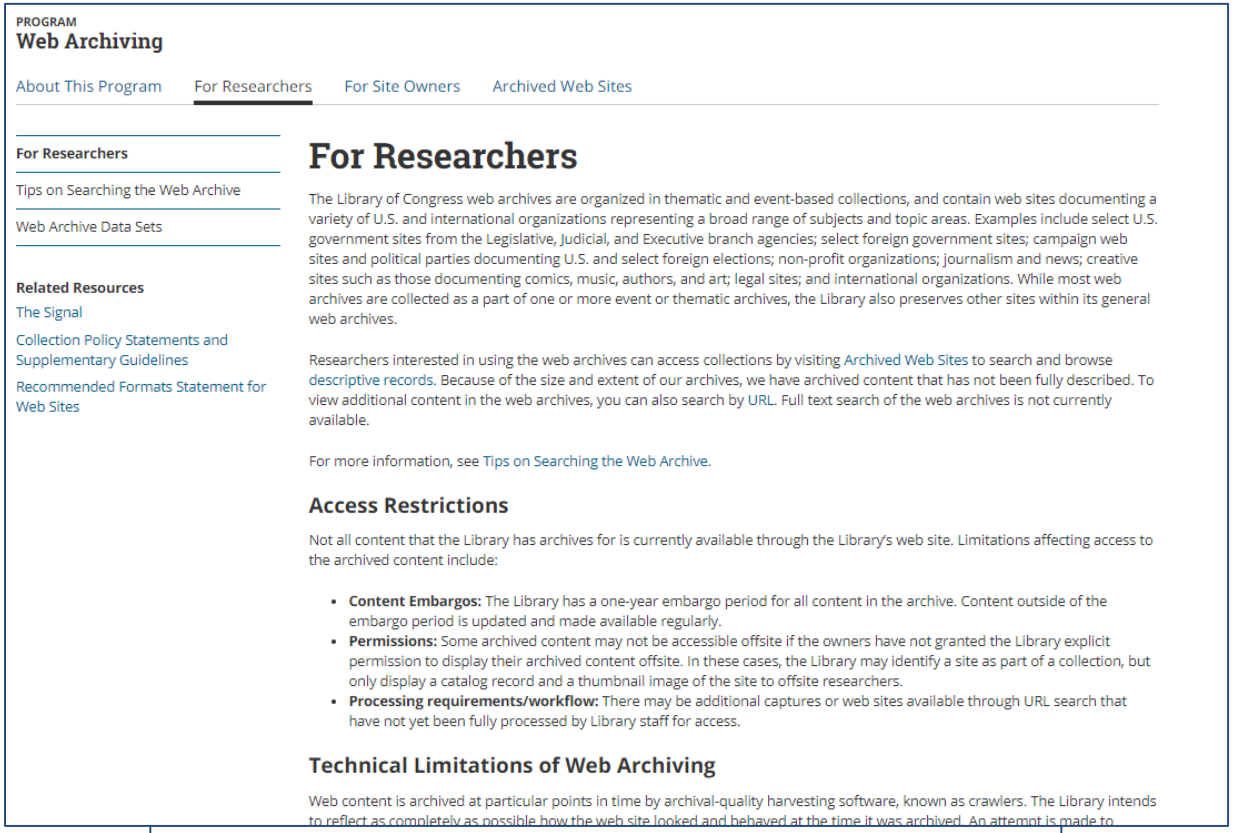
New! Web Archive Datasets



https://labs.loc.gov/experiments/webarchive-datasets/
http://blogs.loc.gov/thesignal/category/web-archiving-2/

Access Challenges:

- Keeping track of all of the various permissions approaches and helping our curators understand
- Attempts to explain embargos, permissions, etc. to researchers and site owners
- Trying to help staff and users make sense of what's been described vs. what we have available in Wayback
- Managing onsite/offsite access at the "site" level limits ability to upgrade/try out some new access tools
- How much can provide in terms of datasets with these mixed access situations?



PROGRAM
Web Archiving

About This Program | For Researchers | For Site Owners | Archived Web Sites

For Researchers

Tips on Searching the Web Archive

Web Archive Data Sets

Related Resources

The Signal

Collection Policy Statements and Supplementary Guidelines

Recommended Formats Statement for Web Sites

For Researchers

The Library of Congress web archives are organized in thematic and event-based collections, and contain web sites documenting a variety of U.S. and international organizations representing a broad range of subjects and topic areas. Examples include select U.S. government sites from the Legislative, Judicial, and Executive branch agencies; select foreign government sites; campaign web sites and political parties documenting U.S. and select foreign elections; non-profit organizations; journalism and news; creative sites such as those documenting comics, music, authors, and art; legal sites; and international organizations. While most web archives are collected as a part of one or more event or thematic archives, the Library also preserves other sites within its general web archives.

Researchers interested in using the web archives can access collections by visiting Archived Web Sites to search and browse descriptive records. Because of the size and extent of our archives, we have archived content that has not been fully described. To view additional content in the web archives, you can also search by URL. Full text search of the web archives is not currently available.

For more information, see Tips on Searching the Web Archive.

Access Restrictions

Not all content that the Library has archives for is currently available through the Library's web site. Limitations affecting access to the archived content include:

- **Content Embargos:** The Library has a one-year embargo period for all content in the archive. Content outside of the embargo period is updated and made available regularly.
- **Permissions:** Some archived content may not be accessible offsite if the owners have not granted the Library explicit permission to display their archived content offsite. In these cases, the Library may identify a site as part of a collection, but only display a catalog record and a thumbnail image of the site to offsite researchers.
- **Processing requirements/workflow:** There may be additional captures or web sites available through URL search that have not yet been fully processed by Library staff for access.

Technical Limitations of Web Archiving

Web content is archived at particular points in time by archival-quality harvesting software, known as crawlers. The Library intends to reflect as completely as possible how the web site looked and behaved at the time it was archived. An attempt is made to

https://www.loc.gov/programs/web-archiving/for-researchers/

Thanks!
abgr@loc.gov
@agrotke