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:

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 around 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=76689&utm%26f%3F&collections=American%3AMusic%3ACreator=Web%3AArchive

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/web/. If you have questions, comments or recommendations concerning the web archiving of your site, please email the Library’s Web Archiving Team at weboffice@loc.gov at your earliest convenience.

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

Records browse and search: [loc.gov/websites/collections](http://loc.gov/websites/collections)

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

Web Archive Data Sets

Explore content from the Library's web archives

In order to better facilitate research and to better understand the needs of users who might be interested in the Library of Congress Web Archives, the Library’s Web Archiving Team is working to create and make available a number of derivative data sets available for users to download, re-use and explore. After participating in the Library’s pilot to explore how we can better enable digital scholarship, beyond basic browsing of records and URL search, the team began exploring ways in which we could process our archives, the scale of which can be overwhelming to digital humanities researchers and other users. From this experimentation, the team created two derivative data sets which would provide researchers smaller bits of data to help users engage with and learn more from our archives.

We plan to launch additional data sets with other types of content in the future, so keep an eye on this page and follow The Signal blog for updates.

Data Sets and Tutorials

Our initial offering is two data sets that comprise of content found in the American Folk Life Center’s Web Cultures Web Archive. As part of an effort to review the coverage of web crawls of GIPHY and Meme Generator, the Web Archiving Team was able to query our web archives data, first finding the number of unique GIFs and distinct meme instances out of both web archives.

- **Meme Generator Data Set Download** - The Meme Generator data set was generated from content harvested from Meme Generator and includes 57,652 unique meme instances derived from base memes (meme images without text, waiting to be fashioned into meme instances). The data set includes some minimal metadata for these Memes. The data set does not include the Memes themselves, however it does provide links to where you can access their web archive copies within the Library’s web archive.

- **GIPHY Data Set Download** - The GIPHY data set was generated from content harvested from GIPHY and includes 10,972 unique GIFs. The data set includes some minimal metadata for these GIFs. The data set does not include the GIFs themselves, however it does provide links to where

[https://labs.loc.gov/experiments/webarchive-datasets/](https://labs.loc.gov/experiments/webarchive-datasets/)

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?

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

Thanks!
abgr@loc.gov
@agrotke