LIBRARY OF CONGRESS COLLECTIONS POLICY STATEMENTS SUPPLEMENTARY GUIDELINES

Web Archiving

Contents

Preface

The Library's traditional functions to acquire, catalog, preserve, and serve collection materials of historical importance to Congress and the American people extend to digital materials, including websites. The Library acquires and makes permanently accessible born digital works that are playing an increasingly important role in the intellectual, commercial and creative life of the United States. Given the nature of web content and the rapidly changing technologies associated with its acquisition and use, the Library will review these guidelines on a regular basis to ensure that the Library's current and future research needs are met.

I. Scope

This document includes guidance for the recommendation of web-based content, primarily websites, which can be acquired using archival-quality harvesting software, known as crawlers. Web archiving preserves as much of the web-based user experience as technologically possible to provide future users snapshots of the content presented by organizations and individuals on the sites at particular moments in time, including how the intellectual content (such as text) was framed by the website implementation. The contents of a website may range from ephemeral unpublished blogs to digital versions of formal publications that are also available in print. For this document, web archiving efforts at the Library are defined as the intent to reflect as completely as possible how a website looked and functioned at the time it was harvested, acknowledging that web crawlers have technical limitations and may not be able to preserve all components of a website.

This *Supplementary Guidelines* document should be used in conjunction with subject-focused *Collections Policy Statements* and any Library-wide collecting priorities guidance when selecting content for the Library's web archives. Guidelines for the acquisition of other digital materials may be found in the following *Supplementary Guidelines*: *Datasets*, *Electronic Resources*, and *Open Digital Content*, which encompass materials acquired via acquisition methods other than web archiving, such as purchase/subscription, direct download, file transfer, or API-based accessioning. For guidance on the range of social media content that may be collected by the Library, the *Supplementary Guidelines for*

Social Media should be consulted.

II. Diverse and Inclusive Collecting Statement

As the nation's de facto national library, the Library of Congress strives to build an expansive, yet selective, collection that records the creativity of the United States and is reflective of the nation's diversity and complexity. The Library's mandate is to have collections that are inclusive and representative of a diversity of creators and ideas. A priority includes acquiring material of underrepresented perspectives and voices in the Library's collections to ensure diverse authorship, points of view, cultural identities, and other historical or cultural factors. The Library also seeks to build a research collection that comprises a globally representative sample of international materials that are diverse in voice and perspective, relative to their places of origin, further supporting the Library's mission to sustain and preserve a universal collection of knowledge and creativity for Congress and future generations.

Diverse collecting is mentioned within many of the Library's Collections Policy Statements. In addition, the Library has adopted several specific collection policies in an effort to ensure it is building an inclusive and representative collection. For more information, see the Library's Collections Policy Statements on *Ethnic Materials*, *LGBTQIA+ Studies*, *Women's and Gender Studies*, *Independently Published and Self-Published Textual Materials*, *and* *Countries and Regions with Acquisitions Challenges*.

III. Research Strengths

By amassing a collection of material, the Library of Congress strives to provide for future generations the keys to the interpretation of events that may not be extant elsewhere. While the Web Archiving program began in 2000, the bulk of archived content comprises websites crawled since 2013. Examples of older collections include the September 11th (2001), Iraq War (2003), and Hurricane Katrina (2015) web archives which house a wealth of unique and ephemeral information with historic value. Ongoing collections continue to provide coverage of the Library's comprehensive collecting areas with specific collection strengths in archives focused on U.S. national elections, Legislative Branch websites, and Congressional websites back to the 107th Congress (2003). Other diverse events and topics covered via curated representative collections include blogs about law, U.S. sites related to public policy topics, sites covering emerging cultural traditions and vernaculars on the web, sites covering performing arts, and websites documenting crises with global impacts such as the Coronavirus pandemic.

IV. Collecting Policies

The Library's subject-focused *Collections Policy Statements* and Library-wide collecting priorities guide decisions about whether archivable web content is in scope for collecting. Given the vast size and growing comprehensiveness of the Internet, the short lifespan of much of its content, and resource limitations, the Library narrowly defines the scope and priorities for its web collecting. In some instances, it develops partnerships and cooperative relationships to continue fulfilling its vital historic mission in order to supplement the Library's capacity.

The Library's permissions process impacts most sites archived by the Library of Congress. Under the Library's permissions process, some notice at a minimum must be provided to the site owner, with the exception of U.S. government sites or those that use Creative Commons or similar terms of service.

In general, the Library follows a collection-based approach to building its web archive. However, there also is a "Single Sites" capability that allows for the ad hoc collecting of a limited number of representative sites in a variety of subject areas. Recommending Officers may use this space when the content is ephemeral, high value, not in scope for an existing collection, or does not warrant the proposal of a full collection.

The selection of websites for the Library's subject, thematic, or topical collections depends on the scope and intent of the archive, as outlined in an approved Collection Proposal. Proposals must be approved either by Library management for Library-level efforts or by the Web Archiving Collection Development Group (WACDG), an interdepartmental group that reviews and approves routine new collection proposals and collection update requests quarterly, allocating available web archiving resources to approved requests.

Recommending Officers select websites, based on the guidance from the approved Collection Proposal and any relevant Library of Congress selection criteria. Recommending Officers and Library management are responsible for identifying potential "events" or "themes" for collections. Recommending Officers recommend content in their assigned areas but are not limited to a defined theme, subject or topic for proposing a collection within their area. Examples of collections recommended for web archiving range from science blogs and public policy organizations to news sites and web comics. Examples of the Library's archiving efforts for sites of an institutional interest are the Library's own web presence and select Federal websites.

As with any format, the resources necessary to sustainably steward web archives, from selection, cataloging, and access, to storage and preservation, must be considered in the decision to collect websites. Given that the universe of web archivable web content will always be greater than the Library's capacity to acquire and provide enduring access to the content, the Library further promotes selectivity by relying on annual collecting priorities, which are collaboratively developed in consultation with subject experts. While collecting may be initiated on any topic covered by the Library and for which content is in scope as per subject-focused Collections Policy Statements, recommendations that respond to the annual areas of focus are prioritized in terms of resource allocation.

Comprehensive collecting areas include the sites of Legislative Branch agencies, U.S. House and Senate offices and committees, and U.S. national election campaigns. In addition to legislative websites, the Library seeks to broadly archive websites from all branches of government. The Library comprehensively harvests all Judicial Branch websites. The Library collects selectively for the Executive Branch due to the large number and size of the Executive Branch websites and the commitments by other agencies (GPO, NARA, etc.) to archive. As a result, the Library focuses its archiving effort on cabinet-level agencies and the affiliated programs that complement the Library's Judicial and Legislative collections. The Library does not archive the sites of national laboratories, the majority of dot.MIL sites, and only selected smaller agencies on a case-by-case basis. State-level websites are not collected on any systematic basis. Websites for all other areas except U.S. elections and the federal materials mentioned here are

collected on a representative basis, namely materials that serve to introduce and define a particular theme, subject, or topic area and to indicate the varieties of information available.

Non-U.S. websites are collected on a highly selective basis at the national level. The Library does not collect subnational materials such as state, provincial, or city-level websites except in rare cases. Typically, non-U.S. content that is archived by host country institutions is not prioritized; however, factors such as usefulness to the current or future needs of Congress and Library researchers or concerns over enduring public access to the archived content are taken into consideration when making collecting determinations. Collecting focused primarily on non-U.S. websites should be "of most immediate concern to the people of the United States."[1]

When a large collection of web-based born digital publications of interest can easily be determined, the recommending division, in consultation with the acquiring divisions, may authorize the selection of in-scope websites by acquisitions staff as part of collections initiated by senior Library management.

V. Collecting Guidelines

The Library of Congress acquires, through web harvesting, selected websites and their multi-format contents for use by the U.S. Congress, researchers, and the general public. Consideration should be given to the factors below when recommending web content for acquisition via web archiving.

- High priority for acquisition should be given to the following:
    - Web content that falls under categories in which the Library collects comprehensively as articulated in the Collecting Policies section of this document
    - Web content that responds to the Library's annual collecting priorities for web archiving

- In addition, consideration should be given to the following factors when recommending web archivable content for acquisition:
    - usefulness in serving the current or future informational needs of Congress and researchers
    - uniqueness and quality of information provided
    - scholarly content
    - at risk of content loss (due to ephemeral nature of some websites)
    - currency of the information
    - relationship to other resources in the Library's collections
    - whether web archiving is the appropriate method to document the topic or event if other resources exist

- Formats which are included in a site may be: audio-visual materials, prints, photographs, maps, or related items required to support research in the subject covered. The Library makes an attempt to gather multi-format objects associated with a website, however it's important to note that web platforms and presentations evolve continuously and existing web archiving tools

[1] Canon Three, Canons of Selection (https://www.loc.gov/acq/devpol/cps.html)

have technical limitations.

- Due to technical challenges, the following categories of web content are better suited to other modes of acquisition and are typically not acquired via web archiving:
    - Form and database-driven content
    - Dynamic content based on programming scripts and content requiring user input or plug-ins for rendering
    - Streaming media and audio/visual content such as podcasts and YouTube
    - Materials residing in the deep web (parts of the Internet not fully accessible through standard search engines)
    - Password-protected sites and those requiring user registration
    - Content behind a paywall

- As web content can be highly interdisciplinary and can exhibit significant format diversity within a site, staff are encouraged to consult across recommending divisions and with other relevant Library departments when judging the appropriateness and technical specifications of web content for the Library's collections.

- The Library's web archiving program adheres to Copyright law and operates in a permissions based environment requiring the Library to provide notice of crawls at minimum or explicitly secure permissions from content owners depending on the country and category of the content. As such, contact information (preferably an email address for the site owner) should be identified and provided.

With the nature of the web and related technology constantly changing, the Library will need to periodically re-evaluate the best methods for selecting works to archive.

Revised July 2022