Web Archiving

Contents

I. Scope

The Library's traditional functions of acquiring, cataloging, preserving and serving collection materials of historical importance to Congress and the American people extend to digital materials, including web sites. 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 vast size and growing comprehensiveness of the Internet, as well as the short life-span of much of its content, the Library must: (1) define the scope and priorities for its web collecting, and (2) develop partnerships and cooperative relationships required to continue fulfilling its vital historic mission in order to supplement the Library's capacity. The contents of a web site may range from ephemeral social media content to digital versions of formal publications that are also available in print.

Web archiving preserves as much of the web-based user experience as technologically possible in order to provide future users accurate snapshots of what particular organizations and individuals presented on the archived sites at particular moments in time, including how the intellectual content (such as text) is framed by the web site implementation.

The guidelines in this document apply to the Library's effort to acquire web sites and related content via harvesting in-house, through contract services and purchase. It also covers collaborative web archiving efforts with external groups, such as the International Internet Preservation Consortium (IIPC).

II. Current Practice

Since its inception in 2000, web archiving at the Library has primarily been a *collection-based activity*. This means that the usual practice is not to acquire individual web sites one-by-one, but as part of a *named subject, event, or theme-based collection*. The sites harvested for the collection are curated by Recommending Officers (ROs), who set the frequency and scope of the harvesting of a site. The Library's goal is to create an archival copy – essentially a snapshot – of the site at a particular point in time or over a period of time.

A proposal for a collection is submitted to the Web Archiving Collection Development Group (WACDG) by a RO. The proposal may be comprehensive for a particular stated scope, or highly selective in representing a particular class or type of site. The attributes of the proposed collection are outlined in

the proposal: (1) name of the collection; (2) background information; (3) and a justification. In addition, the proposal includes the collection's expected scope in terms of types of sites; approximate number of seeds (the initial URLs specified by the RO); frequencies of harvest (i.e., weekly, monthly, biannually, annually); and whether the harvesting should be done for a limited specified time period or as an ongoing effort.

Once a proposal is approved by the WACDG, and the Office of the General Counsel has provided guidance on a permissions approach, the RO evaluates and then selects specific sites for the subject-based collection. The RO becomes the curator of the sites specified for the collection and is responsible for reviewing the selected sites on a periodic basis to ensure that they are still in scope for the collection.

Web archiving conducted by the Library of Congress is impacted by the Library's permissions process that applies to most types of sites. 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.

Archived sites are made available to the public via the Library's local installation of the Internet Archives *Wayback Machine*. More information on the program and collections can be found at the Library of Congress Web Archives homepage (http://www.loc.gov/webarchiving/)

In the case of web sites and web collections which the Library has collected or developed cooperatively with other research institutions, and which are stored in off-site repositories not under the jurisdiction of the Library, the Library will arrange with the repository to make the works available electronically to its patrons, ensuring permanent access or future transfer to the Library for archival storage.

III. Research Strengths

By amassing a collection of material, the Library of Congress will provide to future generations the keys to the interpretation of events that may not be extant anywhere else. While the Web Archiving program began in 2000, the bulk of content archived comprises web sites crawled since 2013. Examples of ongoing collections are U.S. national elections, the Legislative Branch and Congressional web sites back to the 107th Congress (2003), legal blogs, U.S. sites related to public policy topics, and sites covering international humanitarian crises and/or conflicts.

IV. Collecting Policies

The Library of Congress will acquire through web harvesting selected web sites and their multi-format contents for use by the U.S. Congress, researchers, and the general public. The Library selects web sites for its permanent collections which rank high on the following list of criteria: usefulness in serving the current or future informational needs of Congress and researchers, unique information provided, scholarly content, at risk of loss (due to ephemeral nature of some web sites), and currency of the information. The Library will define the attributes for selection, preserve the web content that reflects the original user experience and provide access to archived copies of the harvested material.

In general, the Library follows a collection-based approach to building its web archiving collection. However, there also is a "single sites" capability that allows for the collecting of representative sites in a variety of subject areas. As with any format, the cost of the work and the requirements of selecting, cataloging, serving, storing, and preserving must be considered in the decision to collect web sites.

Selection of works (i.e., web sites) for the LC Collections depends on the subject and extent of the web site harvest as described in a Collection Proposal and Specification that has been approved by the WACDG. 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. A Recommending Officer associated with the Collection with responsibility for the subject, language, or geographic area is responsible for recommending web sites, with the exception of collections and sites that are of interest to the institution as a whole.

Recommending Officers are not limited to a defined theme, subject or topic area for proposing a collection. Examples of collections recommended for web archiving range from foreign elections and public policy organizations to news sites and web comics. For each collection the scope and intent of the archive is outline in the collection proposal.

Examples of the Library's archiving efforts for sites of an institutional interest are the Library's own web presence and select Federal websites. The sites of Legislative Branch agencies, U.S. House and Senate offices and committees, and U.S. national election campaigns are acquired comprehensively. 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.

As with any format, the cost of the work and the requirements of selecting, cataloging, serving, storing, and preserving must be considered in the decision to collect web sites

Foreign web sites are collected on a highly selective basis. To avoid duplication of effort, recommenders of international sites should verify that the content is not already being archived and made publicly available by the host country. Exceptions to this policy can be made if there are concerns over the long-term accessibility of a foreign website. Proposals for new collections of foreign web sites should include a statement in the justification addressing how the foreign web sites are "of most immediate concern to the people of the United States."[1]

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 September 2017

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