LIBRARY OF CONGRESS COLLECTIONS POLICY STATEMENTS SUPPLEMENTARY GUIDELINES

Web Archiving

Contents

I. Scope
II. Research Strengths
III. Collecting Policy
IV. Acquisition Source: Current and Future
V. Collecting Levels

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.

In 2000, the Library of Congress established the MINERVA Web Preservation Project in order to "initiate a broad program to collect and preserve primary source materials." A multi-disciplinary team of Library staff developed thematic web archives on such topics as the U.S. national elections, the Iraq War, and the events of September 11th, among others.

In 2003, the Library and other national libraries and the Internet Archive formed the International Internet Preservation Consortium, an acknowledgement of the importance of international collaboration for preserving web content. Also in 2003, a Web Archiving Collections Policy Statement was created and approved by the Collections Policy Committee.

The Office of Strategic Initiatives created a team in 2004 to: collect web content, test and model a variety of digital content and associated metadata harvesting mechanisms, and build on an enterprise-wide understanding of technical decisions and tools relevant to harvesting content and developing a strong web archiving infrastructure.

In 2005 the "Selecting and Managing Content Captured from the Web" (SMCCW) project looked at the criteria for selecting Web content for harvest, the extent to which technical capabilities enable, affect or prevent the building of Web site collections according to those criteria, and the custodial activities required to ensure the continued viability and value of the content, as well as other aspects of the full digital life cycle processes related to web archiving.

In 2008, all of the Library's Collections Policy Statements and related documents were reviewed and updated. At that time, the Web Archiving Collections Policy Statement was revised and became a *Supplementary Guidelines* document.

In 2010, Library Services (LS) appointed a Web Archiving Coordinator to coordinate all web archiving activities in LS, with an emphasis on liaison functions to stakeholders inside and outside the service unit and the Library.

The web is growing steadily, and at the same time is continually disappearing. Web sites disappear, and site content tends to change rapidly. Given the vast size and growing comprehensiveness of the digital universe, as well as the short life-span of much of its content, it is clear that 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.

Current Practices

Since its inception, 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 Management Oversight Committee (MOC) by a RO. The proposal may be comprehensive for a particular stated scope, or very 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 MOC, 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 steward 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, with the exception of government sites or those that use Creative Commons or similar terms of service. Under the Library's permissions process, some notice at a minimum must be provided to the site owner.

This *Supplementary Guidelines* document should be used in conjunction with the *Electronic Resources Supplementary Guidelines* document and other subject *Collections Policy Statements*.

II. Research Strengths

The contents of a web site may range from formal publications that differ from print publications in the classified collections only by format to ephemeral 140 character "tweets." Web archiving preserves as much of the web-based user experience as technologically possible in order to provide future users of these archived web sites 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.

By amassing a collection of this material, the Library of Congress will provide to future generations the keys to the interpretation of events that may not be extant anywhere else. As of 2013, the Library had collected over 475 terabytes of archived web site content, with collections including: Elections 2000-2012; September 11th; legislative branch and Congressional Web Sites back to the 107th Congress; Legal Blawgs; Iraq War; Crisis in Darfur, visual materials, and selected single sites not related to a theme.

III. Collecting Policies

The web archiving mission statement adopted by the Web Archiving Management Oversight Committee and the Collections Policy Committee in 2013 is:

> *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 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. The sites of Legislative Branch agencies, U.S. House and Senate offices and committees, and U.S. national election campaigns will be acquired comprehensively. For other categories of web sites, only representative sites will be chosen, primarily as part of collections on a defined topic, theme or event.*

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.

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.

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 MOC. 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.

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. Storage is costly in time and in money; hence the selection must be considered carefully. Evolving web technologies require new tool sets to accurately harvest web content.

The Library is committed to preserving its web sites and web collections just as it is to ensuring enduring access to its analog collections in print and other formats.

IV. Acquisition Sources: Current and Future

Current: Recommending Officers select web sites, or the seed URLs, based on the guidance from the approved Collection Proposal and Specification and any relevant Library of Congress selection criteria. Recommending Officers and Library management are responsible for identifying potential “events” or “themes” for collections. A content scope is outlined and submitted to Library management to determine if resources are available. Proposals or ideas for new web archiving projects must be approved within the division and directorate prior to their start. Recommending Officers also select single sites in their subject fields.

Future: 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.

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.

V. Collecting Levels

Collecting Levels are determined by the scope of the Library’s Collections Policy Statements, the selections by the Recommending Officers, and considerations of the cost of the work and the requirements of selecting, cataloging, serving, storing, and preserving the web sites.

Revised by the Collection Development Office, December 2013.