⊖ Collections Policy Statement Index

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 now acquires and makes permanently accessible "born
digital" works (i.e. digital materials which do not have an analog equivalent) 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 these primary source materials. A multi disciplinary team of
Library staff representing cataloging, legal, public services, and technology services studied methods to
evaluate, select, collect, catalog, provide access to, and preserve these materials for future
generations of researchers. As part of this program, the Library developed thematic Web archives on
such topics as the United States national elections, the Iraq War, and the events of September 11th and
others. The Library of Congress Web Archives (LCWA) is available through the Library's Web.

In July 2003, the Library and other national libraries and the Internet Archive (USA) acknowledged the
importance of international collaboration for preserving Internet content for future generations and
formed the International Internet Preservation Consortium (IIPC). At present, 38 institutions are
members of the IIPC.

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

In 2005, the "Selecting and Managing Content Captured from the Web" (SMCCW) project was initiated
to address some of the Life Cycle processes related to Web capture.

Recommending Officers have been engaged throughout this development process in identifying and
selecting appropriate Web sites to capture. With Web capture, Recommending Officers must identify

URLs to add to the collections, rather than evaluate what is received through traditional acquisition mechanisms (such as copyright deposit.)

The web is growing steadily, and at the same time is continually disappearing. Web sites disappear and the 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 of its collecting responsibilities in this new world, and (2) develop partnerships and cooperative relationships required to continue fulfilling its vital historic mission.

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

II. Research Strengths

Web sites offer up-to-date information on given topics and/or document the history of a government agency or commission. As such, they may be ephemeral, disappearing after a short period of time, but their impact may be immense and provide historical sociological data that may not be found elsewhere. By amassing a collection of this material at the Library of Congress, we provide to future generations the keys to the interpretation of events that may not be extant anywhere else. As of 2008, the Library of Congress had collected over 80 terabytes sites, with collections relating to : Elections 2000-2006; September 11th ; Olympics 2002; the 107th Congress to current Congressional Web Pages; Iraq War; Papal Transition; Darfur, visual materials, organizational sites and other collections as well as single sites not related to a theme. The collections continue to grow.

III. Collecting Policies

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 web sites), and currency of the information.

Selection of works for the collection depends on the subject and extent of the collection capture as defined by the Collection Proposal and Specification that has been approved. Formats which are included may be: audio-visual materials, prints, photographs, maps, or related items required to support research in the subject covered. The Recommending Officer responsible for the subject, language, or geographic area is responsible for recommending web sites. At present, the Library of Congress has elected to perform selective collecting, by which the sites captured are determined by Recommending Officers in consultation with management.

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 capture 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 Capture projects must be approved within the division and directorate prior to the start of any new Web Capture project. Further approvals are needed before work begins. 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 reconsider the best methods for selecting works to capture.

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 legally contract 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 Humanities and Social Sciences Division, November 2008.