WEB ARCHIVING

- **Web Archiving Home**
- **About the Program**
- **Cataloging**
- **Collections in Progress**
- **FAQs**
- **Contact**

Resources

- Digital Preservation
- Library of Congress Web Archives

Cataloging

Archived Web sites were cataloged using the Metadata Object Description Schema (MODS). Preliminary keyword, title, and subject metadata were extracted from the archived Web sites to create preliminary MODS records that were subsequently reviewed and/or enhanced by catalogers who assigned controlled subjects from Library of Congress Subject Headings (LCSH) or Thesaurus of Graphic Materials (TGM). A Lucene search interface was developed to search the MODS records both within and across the archived collections.

Collection-level:

In addition, a MARC record for each **collection** is available in the Library of Congress Online Catalog so that the collection can be found along with other Library materials in the catalog.

Metadata included in **collection**-level records in Library of Congress Online Catalog:

245 $a Collection title $h [electronic resource].
520 $a General description of the collection content and number of Web sites and date range when Web sites were captured
6XX $a Collection-level subject heading (usually several 6XX fields)
856 $a http://hdl.loc.gov/loc.natlib/collnatlib.12345678

Web site level:

MODS data included in record for each archived Web site:

TITLE INFO
<titleInfo><title> - Title extracted by system from HTML title tag (when available) and reviewed by cataloger, otherwise supplied by cataloger
<titleInfo type="alternative"><title> - Alternative Title supplied by cataloger if different and useful.

NAME
<name type="personal"><namePart> - Name of Web site creator in inverted order; supplied by cataloger
<name type="corporate"><namePart> - Corporate Name of Web site creator; supplied by cataloger

TYPE OF RESOURCE
<typeOfResource> -"text"; supplied by system

GENRE

<genre> -"Web site"; supplied by system

ORIGIN INFO (A single site may have multiple captures--the first and last dates of capture are recorded)
<originInfo>
<dateCaptured encoding="iso8601" point="start">
- Date of first capture of site; extracted by system from site
<dateCaptured encoding="iso8601" point="end">
- Date of last capture of site; extracted by system from site

LANGUAGE (languageTerm repeated for languages as needed)
<language>
<languageTerm authority="iso639-2b" type="code"> - 3 letter code supplied by cataloger

PHYSICAL DESCRIPTION (internetMediaType repeated for types as needed)
<physicalDescription>
<internetMediaType> - MIME type; supplied by system

ABSTRACT
**** - Extracted by the system from the META name="description" tag in archived Web site (when available); reviewed and/or edited by cataloger

NOTE
<note type="system details"> - A note that records the URL of the Web site at the time of capture; supplied by system

SUBJECT (Subject repeated for subject headings and key words as needed)
<subject authority="lcsh"> - Collection-level and Web site specific (item-level) LCSH headings; supplied by cataloger (Collection-level headings are the same as are in collection-level record in LC Online Catalog)
<subject authority="lctgm"> - Collection-level and Web site specific (item-level) TGM headings; supplied by cataloger (Collection-level headings are the same as are in collection-level record in LC Online Catalog)
<subject authority="local"> - Subjects assigned by cataloger
<subject authority="keyword"> - Subject keywords extracted from META name=keywords tag in archived Web site (when available); reviewed, augmented, and/or edited by cataloger

RELATED ITEM (Contains the collection title and the persistent ID for the collection)
<relatedItem type="host">
<titleInfo><title> - Collection Title; supplied by system
<location><url> - Persistent ID for the collection, e.g., http://hdl.loc.gov/loc.natlib/collnatlib.12345678
that resolves to the collection Overview page; supplied by system

IDENTIFIER (Contains the Resource ID for the Web site for single sites and for the resource page for a site with multiple captures)
<identifier> - Resolvable persistent identifier for the archived web site at the Library of Congress; supplied by the system

LOCATION
<location><url usage="primary display"> - Resolvable persistent identifier for archived Web site; supplied by system

ACCESS CONDITION

<accessCondition> - Rights/permissions information; supplied by system

RECORD INFO
<recordInfo>
<recordCreationDate encoding="iso8601"> - Record creation date; supplied by system
<recordIdentifier source="dlc"> - Identifier for the MODS record; supplied by system

Connect with the Library

All ways to connect »

Find Us On

Subscribe & Comment

RSS & E-Mail

Blogs

Download & Play

Apps

Podcasts

Webcasts

iTunesU

Questions

Ask a Librarian

Contact Us

About | Press | Jobs | Donate
Inspector General | Legal | Accessibility | External Link Disclaimer | USA.gov

Speech Enabled