Best Practices for Descriptive Metadata for Web Archiving

Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group

Alexis Antracoli, Kate Bowers, Jackie Dooley
THE PROBLEM
Absence of community best practices for descriptive metadata was the most widely-shared web archiving challenge identified in two surveys:

- Weber/Graham study of users of archived website (2016)
RBSC Subject-specific Databases

Filter by Division

Architectural Presentation Boards
Contains descriptive information on over 500 architectural presentation boards. Keyword searching retrieves a listing of boards that meet your criteria. Information returned will include a description of the board and the view depicted, and when available, the architect, architectural firm, and other contractor details. The boards typically include floor plans, artistic renderings, and campus footprints.

Archive-It
Provides access to archived versions of Princeton University websites starting from 2015. Archive-It provides browsing capabilities as well as full text search of all websites in the collection.

Catalog of Princeton University Senior Theses
List of theses starting in 1826 written by seniors at Princeton University. Not all departments are represented. Princeton University network connected patrons may view most 2004 theses. For Senior Thesis Searching and Ordering Tips, see the LibGuide: How to Search, Request to View, and Order Princeton University Senior Theses.

Faculty and Professional Staff Index, 1764-2006
Index for Faculty & Professional Research, Technical & Library Personnel files, 1764–2004. Contains the name, death date, departure date, and department for Princeton University personnel. Files for some trustees, administrators, and others may also be found. Explanation of Access to personnel

Princeton University Archives
Collected by: Princeton University Library
Archived since: Dec. 2014
Description: housed within the Seeley G. Mudd Manuscript Library, the Princeton University Archives consists of over 15,000 linear feet of materials including both paper and electronic records, as well as photographs and other visual materials that document the history of Princeton University. The University Archives is also the repository for Princeton senior theses and doctoral dissertations. To find more archiving holdings within the Princeton University Archives and the Princeton University Library, consult our finding aids at http://library.princeton.edu/research/finding-aids.
Subject: University & Libraries, Princeton University
Creator: Princeton University Archives

Narrow Your Results
Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here

Next Page >
College Republicans Records 2004-2016

Summary

This collection is stored at Mudd Manuscript Library.

Requests will be delivered to Princeton University Archives, MUDD Reading Room.

Collection Creator: Princeton University. College Republicans.


Extent: 1 website

Languages: English.

Access Restrictions

The collection is open for research use.

Description

This website is intended for prospective members of the group as well as the general public and includes select photographs of past events and a listing of the organization’s officers (incomplete) dating back to 1954.

Full text searching of this archived web site is available through the Archive-it interface.

Preferred Citation

Public Websites, 2016. College Republicans Records, Princeton University Archives. Department of Rare Books and Special Collections, Princeton University Library.
Web Archiving Metadata Working Group

CHARGE: The OCLC Research Library Partnership Web Archiving Metadata Working Group will evaluate existing and emerging approaches to descriptive metadata for archived websites and will recommend best practices to meet user needs and to ensure discoverability and consistency.

The Problem

Archived websites often are not easily discoverable via search engines or library and archives catalogs and finding aid systems, which inhibits use.
Objectives

• Recommend **best practices** for descriptive metadata for archived websites that are **community-neutral** and that address user needs (NOT a standard; displaces no standards)

• Identify the most relevant **data elements** and provide guidance on formulating **content**

• **Bridge** bibliographic and archival practices

• Address the differing characteristics of **item- and collection-level** descriptions
Potential user groups

- Practitioners who use standards but want **guidance on formulating content** for description of archived single sites and collections
  - **Dublin Core** (used by Archive-It) is a structure standard, not a content standard
  - Some archivists want web-specific guidance to supplement **DACS**
  - Those who want to export **MARC** records to a less granular data structure (such as Archive-It)
  - **RDA** (library content standard) focuses on transcription of live single sites, but websites lack standard elements and are not amenable to transcription

- Libraries and archives using a digital asset management system (**DAMS**), eschewing descriptive standards; brief records, scalable approach

- **Individuals** lacking metadata experience who build or contribute to web archives
Feedback received

• How much? A lot!

• Mostly positive; three negative comments

  – Positive: premise accepted; suggestions for improvement
  – Negative: premise rejected, including because …
    • Archival (two comments, including TS-DACS): recommendations are too bibliographic, don’t conform to DACS
    • Bibliographic (one comment): recommendations are too archival
LIT REVIEWS & OTHER RESEARCH

Who uses web archives?

- Digital humanists
- Web scientists
- Computer scientists
- Data analysts
- Journalists
- Lawyers
- Website owners
- Website designers
- Government employees
- Genealogists
- Patent applicants
- Instructors
- Students
- Linguists
- Sociologists
- Political scientists
- Historians
- Anthropologists
Users need “provenance” metadata

- “The critical missing piece”
- Provide context
- Why was the content archived?
- Selection criteria
- Scope

Note: WAM focused solely on descriptive metadata, not technical or preservation metadata.
Metadata practitioner needs

- **Archival** and **bibliographic** approaches
  - DACS, EAD, MARC, Dublin Core, RDA, MODS

- **Data elements** vary widely
  - Same element name, multiple meanings

- **Level of description**
  - Single site, collection of sites, seed URLs

- **Scalability** and limited resources
Best practices methodology

• Analyze metadata standards & institutional guidelines
  – RDA (libraries), DACS (archives), Dublin Core (simplified)

• Evaluate existing metadata records “in the wild”
  – ArchiveGrid, Archive-It, WorldCat

• Identify dilemmas specific to web archiving

General finding: extreme inconsistencies in existing practice

• Incorporate findings from literature reviews

• Prepare data dictionary and report narrative
WEB-SPECIFIC DILEMMAS
• Is the **website creator/owner** the … publisher? author? subject?

• Should the **title** be … transcribed verbatim from the head of the site? Edited to clarify the nature/scope of the site? Append e.g. "web archive" for a collection?

• Which **dates** are important/feasible other than capture dates? Beginning/end of the site's existence? Date of the content? Copyright?

• How should **extent/size** be expressed? 6.25 Gb? 300 websites? 1 archived website? 1 online resource?

• Is the **host institution** that harvests and manages the archived content the repository? creator? publisher? selector?
• Is it important to clearly state that the resource is a website? If so, where? In the title? description? extent statement? all of these?

• Does provenance refer to …the site owner? the repository that harvests and hosts the site? ways in which the site evolved?

• Does appraisal mean …the reason the site warrants being archived? a collection of sites named by the repository? the parts of the site that were harvested?

• Which URLs should be included? Seed? access? landing page?
THE ARCHIVE-IT METADATA INTERFACE
Archive-It, the most widely used web archiving platform, provides for application of DC metadata for collections and seeds.
# Detail of the Collection-level metadata editing interface

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>A-Sites: Archived Harvard Websites</td>
</tr>
<tr>
<td>Creator</td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td>Harvard University</td>
</tr>
<tr>
<td>Description</td>
<td>The Harvard University Archives is charged with collecting and preserving the historical records of the University. Much of the information collected for centuries in paper form now resides on University websites. Beginning in</td>
</tr>
</tbody>
</table>
At the seed level, “grab title” can pull a title from HTML

Archive-It has support for non-DC fields
Bulk options include downloading all seed metadata

<table>
<thead>
<tr>
<th>Creator</th>
<th>Title</th>
<th>Relation (x 2)</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harvard University. Department of African</td>
<td>Department of</td>
<td>African and African</td>
<td></td>
</tr>
<tr>
<td>Harvard University. Faculty of</td>
<td>University.</td>
<td>University.</td>
<td></td>
</tr>
<tr>
<td>Harvard University. Information Technology</td>
<td>Harvard</td>
<td>University Student Info</td>
<td></td>
</tr>
<tr>
<td>Harvard University. Disability Services</td>
<td>Disability</td>
<td>Services</td>
<td></td>
</tr>
<tr>
<td>Harvard University. Adams House</td>
<td>Adams House</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harvard University. Administrative Board</td>
<td>Harvard</td>
<td>University.</td>
<td></td>
</tr>
<tr>
<td>Harvard College (1780-).</td>
<td>Administrative</td>
<td>Board</td>
<td></td>
</tr>
<tr>
<td>Harvard University. Advanced Leadership</td>
<td>Harvard</td>
<td>University Advanced</td>
<td></td>
</tr>
<tr>
<td>Harvard Medical School</td>
<td>Medical School</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harvard University. Accessible Education</td>
<td>Accessible</td>
<td>Educational</td>
<td></td>
</tr>
<tr>
<td>Harvard University. Center for African Stu</td>
<td>Center for</td>
<td>Af</td>
<td></td>
</tr>
</tbody>
</table>

Add ODS Seed Metadata File

Download Existing Seed Metadata
RECOMMENDED BEST PRACTICES
WAM data elements

<table>
<thead>
<tr>
<th>Collector</th>
<th>Extent</th>
<th>Source of Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contributor *</td>
<td>Genre/Form</td>
<td>Subject *</td>
</tr>
<tr>
<td>Creator *</td>
<td>Language *</td>
<td>Title *</td>
</tr>
<tr>
<td>Date *</td>
<td>Relation *</td>
<td>URL</td>
</tr>
<tr>
<td>Description *</td>
<td>Rights *</td>
<td></td>
</tr>
</tbody>
</table>

* = 9 of 14 element names/meanings match Dublin Core
Data dictionary inclusion criteria

- Includes **common elements** used for identification and discovery of all types of resource (e.g., Creator, Date, Subject, Title)
- Other elements must have **clear applicability** to archived websites (e.g., Rights, Description, URL)
- Elements **excluded** that rarely (if ever) appear in guidelines and/or extant metadata records and have no web-specific meaning (e.g. audience, publisher, statement of responsibility)
Data element features

- Element name
- Definition
- Usage note
- Examples
- Crosswalk
Collector

Definition: The organization responsible for curation and stewardship of an archived website or collection.

Use **Collector** for the organization that selects the web content for archiving, creates metadata and performs other activities associated with “ownership” of a resource. Stated another way, this is the organization that has taken responsibility for the archived content, although the digital files are not necessarily stored and maintained by this organization (collections harvested using Archive-It are a prominent example).

*No equivalent in Dublin Core.*
Collector: Lifecycle activities

Institutions involved in web archiving engage in a variety of activities during the lifecycle of archiving web content. We identified four activities performed by the institution that assumes responsibility for archiving web content:

- **Selecting** websites for archiving
- **Harvesting** the content of the designated seed URLs
- Creating and maintaining **metadata** to describe the content
- Making **decisions** about other aspects of **collections management**, including how the harvested files will be preserved and how access will be provided.
Collector: Examples

Creator: Seattle (Wash.)
Title: City of Seattle Harvested Websites
Collector: Seattle Municipal Archives

Title: Globalchange.gov
Contributor: U.S. Global Change Research Program
Collector: Federal Depository Library Program

Creator: Association for Research into Crimes Against Art
Title: ARCAblog: promoting the study and research of art crime and cultural heritage protection
Collector: New York Art Resources Consortium
## Collector: Crosswalks

<table>
<thead>
<tr>
<th>Crosswalks</th>
<th>Dublin Core</th>
<th>EAD</th>
<th>MARC</th>
<th>MODS</th>
<th>schema.org</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Contributor</td>
<td></td>
<td>524</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;repository&gt;</td>
<td>852 subfield a</td>
<td>&lt;location&gt;</td>
<td>schema:OwnershipInfo</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>852 subfield b</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Source of description

Definition: Information about the gathering or creation of the metadata itself, such as sources of data or the date on which source data was obtained.

**Source of Information** is used to identify the source of all or some of the metadata, particularly for descriptions of single sites. Basic aspects of a website (creator name, title, etc.) may change significantly, but the responsible institution is unlikely to have the resources to become aware of changes, let alone update the metadata. Include the date on which the site was examined and the location from which the information was taken.

No equivalent in Dublin Core.
Source of description: Examples

Description based on archived web page captured Sept. 22, 2016; title from title screen (viewed Oct. 27, 2016)

Title from home page last updated June 21, 2012 (viewed June 22, 2012)
Title from home page (viewed on Oct. 11, 2007)
Title from HTML header (viewed Feb. 16, 2006)
### Source of description: Crosswalks

<table>
<thead>
<tr>
<th>Crosswalks</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dublin Core</td>
<td>Description</td>
</tr>
<tr>
<td>EAD</td>
<td>&lt;processinfo&gt;</td>
</tr>
<tr>
<td>MARC</td>
<td>588</td>
</tr>
<tr>
<td>MODS</td>
<td>&lt;note&gt;</td>
</tr>
<tr>
<td>schema.org</td>
<td>schema:description</td>
</tr>
<tr>
<td></td>
<td>schema:disambiguatingDescription</td>
</tr>
</tbody>
</table>
FORTHCOMING PUBLICATIONS
Three simultaneous reports (autumn 2017)

- Best practices for descriptive metadata
  - With data dictionary

Send comments by August 4th!

- User needs
  - With annotated bibliography

- Tools
  - With evaluation grids
Q&A
For more information, please contact:

**Jackie Dooley**  
Program Officer, OCLC Research  

dooleyj@oclc.org  
@minniedw