

SAA Description Section  
Portland, 26 July 2017

# Best Practices for Descriptive Metadata for Web Archiving

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

[oc.lc/wam](http://oc.lc/wam)

**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:

- OCLC Research Library Partnership (2015)
- Weber/Graham study of users of archived website (2016)

## RBSC Subject-specific Databases

Filter by Division  [Apply](#)



### 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 1926 written by seniors at Princeton University. Not all departments are represented. Princeton University network connected patrons may view most 2014 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](#)



[HOME](#) [EXPLORE](#) [LEARN MORE](#) [CONTACT US](#)

The leading web archiving service  
for collecting and accessing  
cultural heritage on the web  
*Built at the Internet Archive*



[Explore](#) >> [Princeton University Library](#) >> [Princeton University Archives](#)



## 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 audiovisual 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 archival holdings within the Princeton University Archives and the Princeton University Library, consult our [finding aids](http://findingaids.princeton.edu) at <http://findingaids.princeton.edu>.

**Subject:** [Universities & 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.

Group

Sort By: [Count](#) | [\(A-Z\)](#)

Academic Units (48)  
Administration (45)  
Alumni (2)  
Senior Theses (1)  
Student Life (20)

[Search](#)

[Clear](#)

[Sites](#)

[Search Page Text](#)

Page 1 of 2 (116 Total Results)

[Next Page](#)

Creator

Sort By: [Count](#) | [\(A-Z\)](#)

Princeton University. Office of the Dean of Undergraduate Students (7)  
Princeton University. School of Engineering

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

Title: Department of African American Studies Public Website


URL: <http://aas.princeton.edu/>

PRINCETON UNIVERSITY

Library Website | Create an Account | Site Feedback

Princeton University Library Finding Aids

TopicsNamesCollectionsLocations



Summary

Description

Collection History

Access and Use

Find More

Contents and Arrangement

College Republicans discussion dinner with former Congressman Ed Zschau '61 (3-26-2015), 2015

College Republicans Facebook page reaches 100 likes in less than six hours (2-16-2015), 2015

College Republicans host a 2016 GOP presidential debate party in Whig Hall (9-16-2015), 2015

College Republicans welcome back BBQ (9-24-2010), 2010

College Republicans welcome back BBQ (9-24-2010), 2010

CPAC selfie (3-9-2014), 2014

Members at an event with John Stossel '69 in McCosh Hall (3-30-2015), 2015

Members campaign for Barbara Comstock and Ed Gillespie in Virginia during fall break (10-31-2014), 2014

Members campaign for Mitt Romney in Virginia during fall break

AC441

College Republicans Records 2004-2016

Search This Collection

Search Tips | How to Browse this Collection

Public Websites

WEBSITE

View ContentAsk a Question

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..

Dates: 2016.

Extent: 1 website

Languages: English.

Access Restrictions

The collection is open for research use.

Description

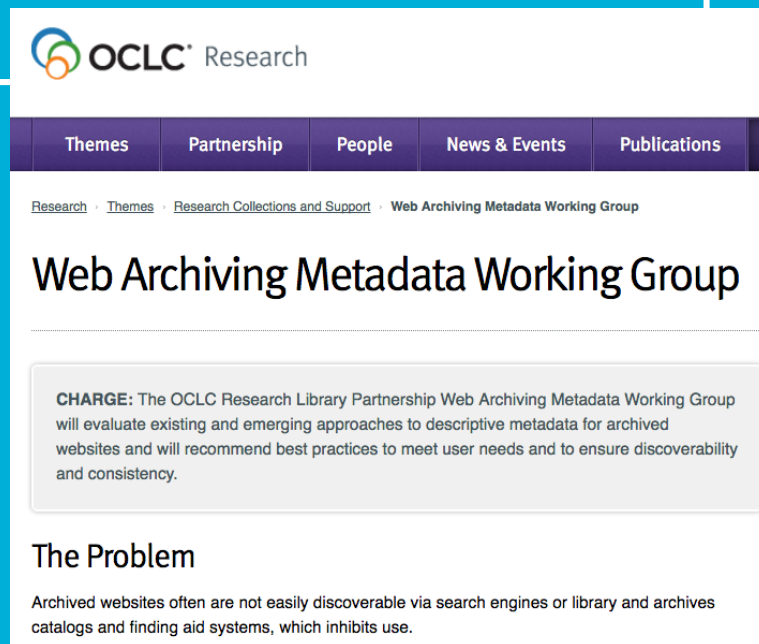
This website is intended for prospective members of the group as well as the general public and includes select photographs of past events and a listing of the organization's officers (incomplete) dating back to 1964.

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.

# OCLC RESEARCH LIBRARY PARTNERSHIP WEB ARCHIVING METADATA WORKING GROUP



The screenshot shows the OCLC Research website. At the top is the OCLC Research logo. Below it is a navigation bar with links: Themes, Partnership, People, News & Events, and Publications. The main content area has a breadcrumb trail: Research > Themes > Research Collections and Support > Web Archiving Metadata Working Group. The title 'Web Archiving Metadata Working Group' is prominently displayed. Below the title, a box labeled 'CHARGE:' describes the group's mission to evaluate metadata approaches and recommend best practices. The section 'The Problem' follows, explaining that archived websites are often hard to find via search engines or catalogs, which hinders their use.

OCLC<sup>®</sup> Research

Themes Partnership People News & Events Publications

[Research](#) · [Themes](#) · [Research Collections and Support](#) · [Web Archiving Metadata Working Group](#)

## 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

Bailey et al. Ben-David & Huurdeman Bernstein Bragg & Hanna Costa  
Costa & Gomes Costa & Silva Cruz & Gomes Dougherty & Meyer Galligan  
Gatenby Gibbons Goel Goethals Guenther Hartman et al. Hockx-Yu  
Jackson Jones & Shankar Lavoie & Gartner Leetaru Mannheimer Masanès  
Milligan Murray & Hsieh Neubert Niu O'Dell Peterson Phillips & Koerbin  
Pregill Prom & Swain Ras & van Bussel Reynolds Riley & Crookston  
Stirling et al. Sweetser Taylor Thomas et al. Thurman & O'Hanlon  
Tillinghast Truman Weber & Graham Webster Wuet et al. Zhang et al.

# 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

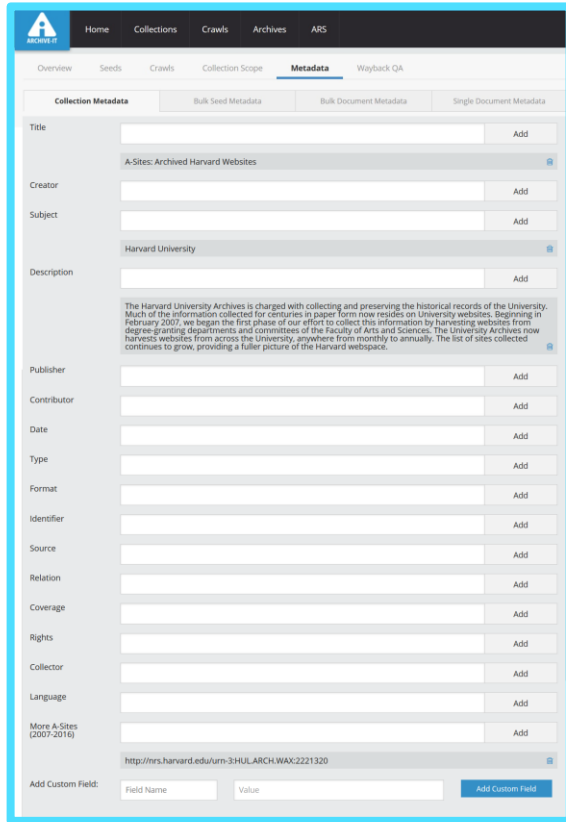
# WEB-SPECIFIC DILEMMAS

- 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



The screenshot shows the Archive-It Metadata form. At the top, there is a navigation bar with links: Home, Collections, Crawls, Archives, and ARS. Below this is a sub-navigation bar with links: Overview, Seeds, Crawls, Collection Scope, Metadata (selected), and Wayback QA. The main form is titled 'Collection Metadata' and has three tabs: Bulk Seed Metadata, Bulk Document Metadata, and Single Document Metadata. The form contains various fields for metadata entry, each with an 'Add' button. The fields are: Title (with a dropdown menu showing 'A-Sites: Archived Harvard Websites'), Creator, Subject, Description (with a text area containing a paragraph about Harvard University Archives), Publisher, Contributor, Date, Type, Format, Identifier, Source, Relation, Coverage, Rights, Collector, Language, and More A-Sites (2007-2016). At the bottom, there is a section for 'Add Custom Field' with input fields for 'Field Name' and 'Value', and an 'Add Custom Field' button.

Archive-It

Home Collections Crawls Archives ARS

Overview Seeds Crawls Collection Scope **Metadata** Wayback QA

**Collection Metadata** Bulk Seed Metadata Bulk Document Metadata Single Document Metadata

Title Add

A-Sites: Archived Harvard Websites

Creator Add

Subject Add

Harvard University

Description Add

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 February 2007, we began the first phase of our effort to collect this information by harvesting websites from degree-granting departments and committees of the Faculty of Arts and Sciences. The University Archives now harvests websites from across the University, anywhere from monthly to annually. The list of sites collected continues to grow, providing a fuller picture of the Harvard webpace.

Publisher Add

Contributor Add

Date Add

Type Add

Format Add

Identifier Add

Source Add

Relation Add

Coverage Add

Rights Add

Collector Add

Language Add



More A-Sites (2007-2016) Add

<http://nrs.harvard.edu/urn-3:HULARCH.WAX:2221320>

Add Custom Field: Field Name Value Add Custom Field

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

Title	<input type="text"/>	Add
	A-Sites: Archived Harvard Websites 	
Creator	<input type="text"/>	Add
Subject	<input type="text"/>	Add
	Harvard University 	
Description	<input type="text"/>	Add
	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	

At the seed level, “grab title” can pull a title from HTML

Title	<input type="text"/>	Add	Grab Title
-------	----------------------	-----	------------

Archive-It has support for non-DC fields


Add Custom Field:	<input type="text" value="Field Name"/>	<input type="text" value="Value"/>	Add Custom Field
-------------------	---	------------------------------------	------------------

# Bulk options include downloading all seed metadata

Overview Seeds Crawls Collection Scope **Metadata** Wayback QA

Collection Metadata **Bulk Seed Metadata** Bulk Document Metadata Single Document Metadata

**Add ODS Seed Metadata File**

 [Download Existing Seed Metadata](#)

	A	B	C	D	
1	Creator	Title	Relation (x 12)	URL	Description (x2)
2	Harvard Law School	HLS 200	Harvard Law School	http://200.hls.harvard.e	Harvard Law School pr
3	Harvard University. Department of African and African Studies	Department of African and African Studies	Harvard University. Faculty of Arts and Sciences	http://aaas.fas.harvard.e	The Department of Afri
4	Harvard University. Information Technology Services	Harvard University Student Information System	Central administration	http://about.my.harvard.e	The Student Informati
5	Harvard University. Disability Services	Harvard University Disability Services	Central administration	http://accessibility.harv	Harvard University Dis
6	Harvard University. Adams House	Adams House	Harvard University. Faculty of Arts and Sciences	http://adamshouse.harv	The Adams website pr
7	Harvard College (1780- ). Administrative Board	Harvard College Administrative Board	Harvard University. Faculty of Arts and Sciences	http://adboard.fas.harv	The Harvard College A
8	Harvard University. Advanced Leadership Program	Harvard University Advanced Leadership Program	Harvard Medical School	http://advancedleadersh	The Harvard Universit
9	Harvard University. Accessible Education Office	Harvard University Accessible Education Office	Harvard University. Faculty of Arts and Sciences	http://aeo.fas.harvard.e	The Accessible Educat
10	Harvard University. Center for African Studies	Harvard University Center for African Studies	Harvard University. Faculty of Arts and Sciences	http://africa.harvard.ed	The Center for African

# RECOMMENDED BEST PRACTICES

# WAM data elements

<b>Collector</b>	<b>Extent</b>	<b>Source of Description</b>
<b>Contributor *</b>	<b>Genre/Form</b>	<b>Subject *</b>
<b>Creator *</b>	<b>Language *</b>	<b>Title *</b>
<b>Date *</b>	<b>Relation *</b>	<b>URL</b>
<b>Description *</b>	<b>Rights *</b>	

\* = 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 will access 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

Crosswalks	
Dublin Core	Contributor
EAD	<repository>
MARC	524 852 subfield a 852 subfield b
MODS	<location>
schema.org	schema:OwnershipInfo

# 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

Crosswalks	
Dublin Core	Description
EAD	<processinfo>
MARC	588
MODS	<note>

schema.org	schema:description schema:disambiguatingDescription
------------	--

# FORTHCOMING PUBLICATIONS

# Three simultaneous reports (autumn 2017)

- Best practices for **descriptive metadata**
  - With data dictionary

**Send comments by August 4<sup>th</sup>!**

- **User needs**
  - With annotated bibliography
- **Tools**
  - With evaluation grids

# Q&A

# SAA Description Section

## 26 July 2017

For more information, please contact:

**Jackie Dooley**

Program Officer, OCLC Research

[dooleyj@oclc.org](mailto:dooleyj@oclc.org)

@minniedw

[oclc.org/wam](http://oclc.org/wam)

**Because what is known must be shared.<sup>SM</sup>**



©2016 OCLC. This work is licensed under a Creative Commons Attribution 4.0 International License. Suggested attribution:  
"This work uses content from **Developing Best Practices for Web Archiving Metadata to Meet User Needs**  
© OCLC, used under a Creative Commons Attribution 4.0 International License:  
<http://creativecommons.org/licenses/by/4.0/>."

