IIPC, London Web Archiving Week
16 June 2017

Best Practices for Descriptive Metadata

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

Alexis Antracoli, Karen Stoll Farrell, Jackie Dooley

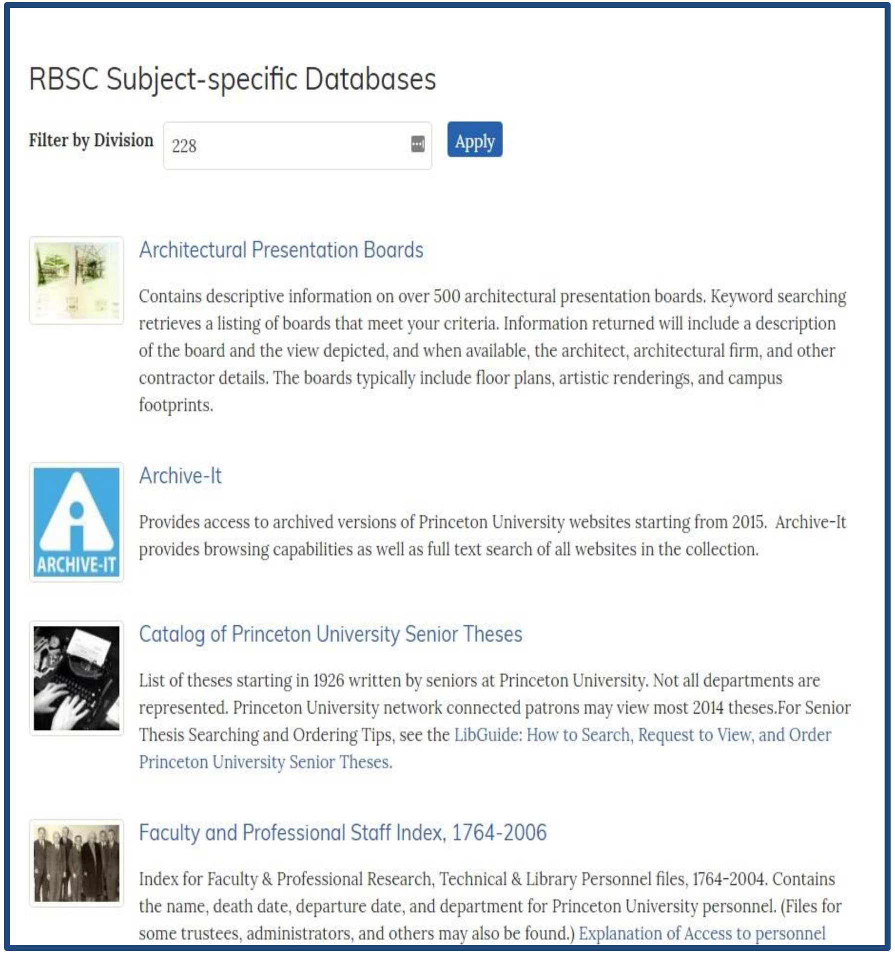
oc.lc/wam

OCLC

THE PROBLEM

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

- 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/Chapman study of users of archived website (2016)



RBSC Subject-specific Databases

Filter by Division 228

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

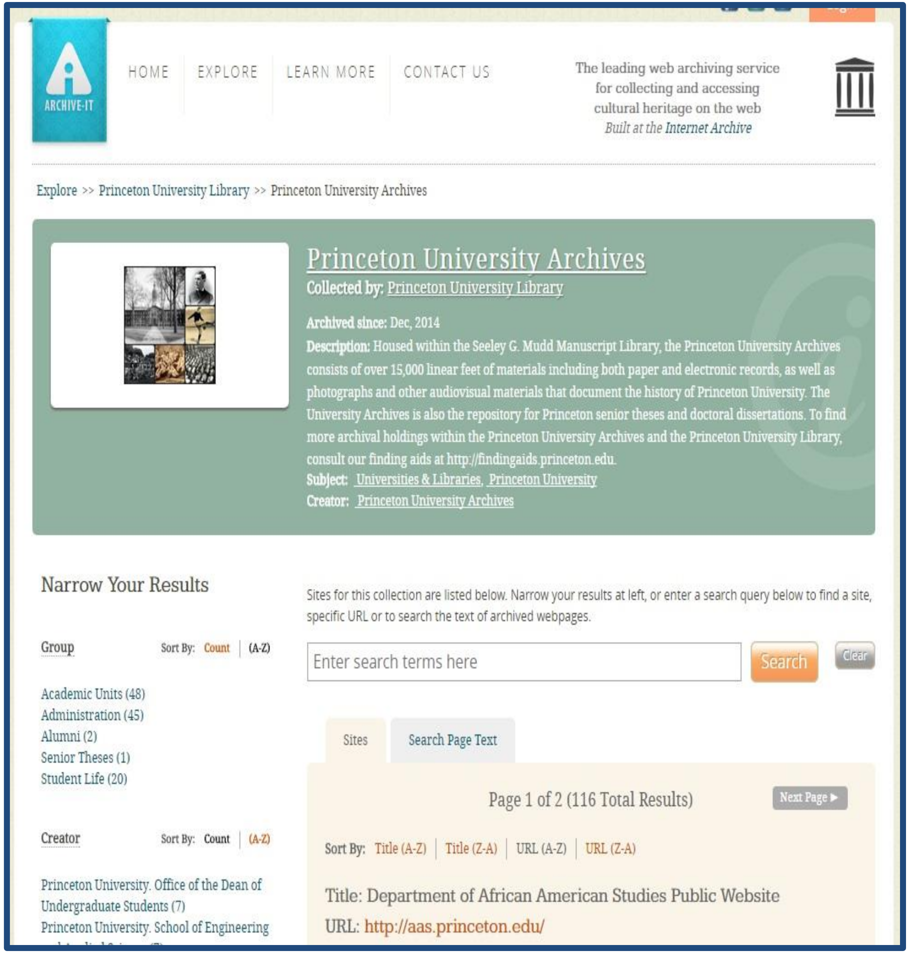
Provides access to archived versions of Princeton University websites starting from 2015. Archive-It provides browsing capabilities as well as full text search of all websites in the collection.

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



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 at http://findingaids.princeton.edu.

Subject: Universities & Libraries, Princeton University

Creator: Princeton University Archives

Narrow Your Results

Group Sort By: Count | (A-Z)

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

Creator Sort By: Count | (A-Z)

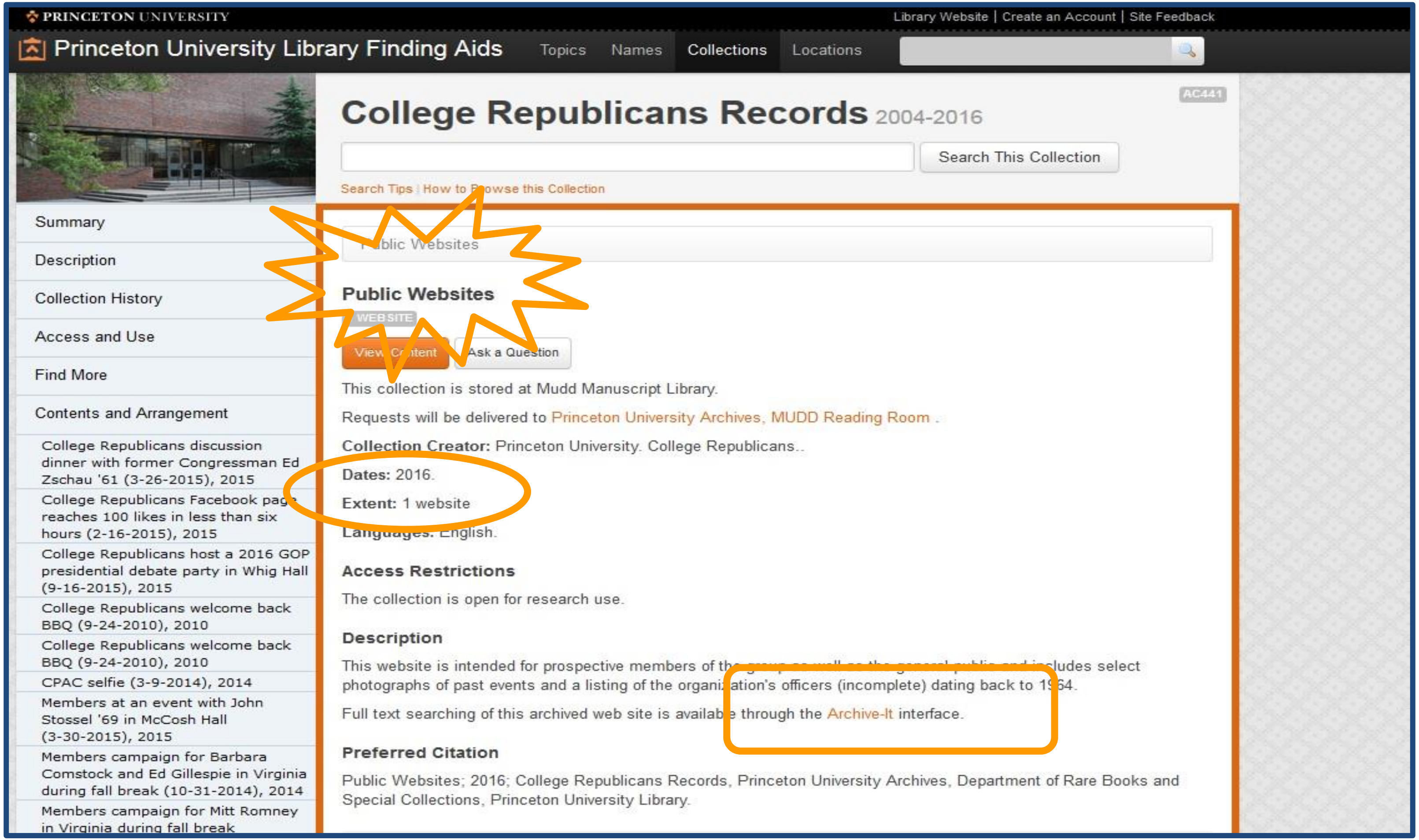
Princeton University. Office of the Dean of Undergraduate Students (7)
Princeton University. School of Engineering

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.

Page 1 of 2 (116 Total Results)

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

AC441

College Republicans Records 2004-2016

Search Tips | How to Browse this Collection

- 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

Public Websites

Public Websites

WEBSITE

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.

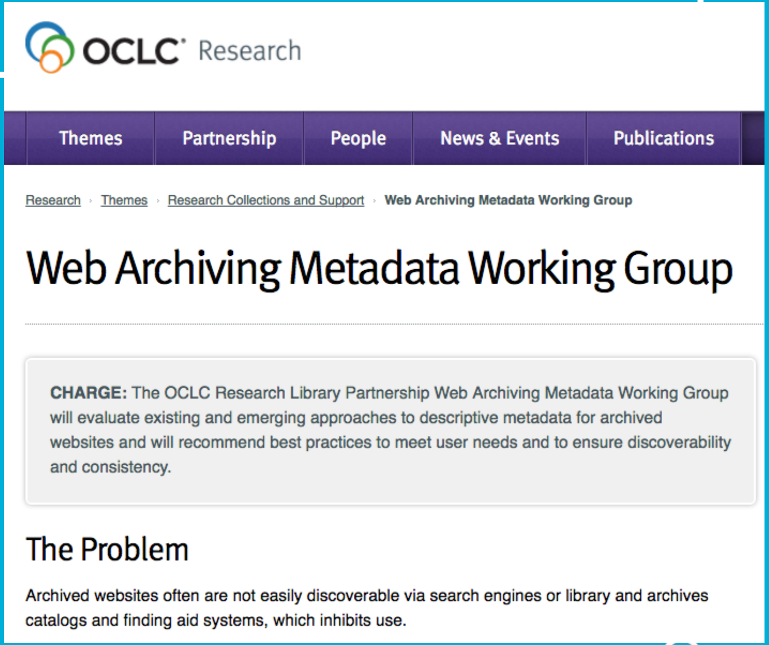
Full text searching of this archived web site is available through the Archive-It interface.

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


OCLC 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.

Objective

- Recommend best practices for web archiving descriptive metadata that are **community-neutral** and **standards-neutral**

- A set of defined data elements (i.e., a **data dictionary**)

Outputs (July 2017)

- Literature review to inform our understanding of documented **user needs** and behaviors

- Best practices for **descriptive metadata** address both single-site and collection approaches

- Analysis of descriptive metadata functionalities of eleven harvesting **tools** [not covered in today's session]

LITERATURE REVIEWS

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 al. Zhang et al.

Who are the end users of 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

How are they using web archives?

- Read specific web pages/sites
- Data and text mining
- Technology development

What behaviors do they use?

Costa and Silva (2010) classify needs into three behavioral groups; much cited by others.

- Navigational
- Informational
- Transactional

Takeaways for end-user needs

- Flexible Formats
- Engagement
- Access and re-use/rights statements
- Archived vs. live
- Subject access

“Provenance” metadata

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

Takeaways for metadata practitioners

- **Archival** and **bibliographic** approaches
  - RDA, MARC, Dublin Core, MODS, finding aids, DACS
- **Data elements** vary widely
  - Same element name, different meanings
- **Level of description**
  - Single site, collection of sites, seed URLs
- **Scalability** and limited resources

DEVELOPING DESCRIPTIVE METADATA BEST PRACTICES

Methodology

- Analyze metadata **standards & institutional guidelines**
  - RDA (libraries), DACS (archives), Dublin Core (simplified)
- Evaluate **existing metadata records** “in the wild”
  - WorldCat, ArchiveGrid, Archive-It
- Identify **dilemmas** specific to web archiving
- 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”?
- 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? 1 archived website? 1 online resource? 6.25 Gb? approximately 300 websites?
- 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?

RECOMMENDED BEST PRACTICES

Setting the context

- Use cases: library, archives, researcher
- Comparisons between …
  - **Bibliographic and archival** approaches to description
  - Description of **archived and live** sites
  - **Collection, site, and document-level** descriptions

Data dictionary characteristics

- **Lean** (14 elements); use on its own or with granular library and archives standards
- Element **names and definitions** adopted or adapted from standards
- **Usage notes** explain how to formulate the content of each element
- The **same element** is used for a concept **at all levels of description** as per multilevel principles expressed in archival standards (DACS and EAD).

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. Access Conditions, 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)

WAM data elements

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

* = 9 of 14 element names/meanings match Dublin Core

Access Conditions [to be renamed Rights]

Definition: Circumstances that affect the availability [and/or re-use] of an archived website or collection.

Use **Access Conditions** to record *whether or not conditions exist that restrict user access* to the archived content. These might include the need to make an appointment for onsite use or a specified period of time during which the content is embargoed. Such conditions may be imposed by an archival repository, donor, other agency, or legal statute.

> This content is embargoed from public access until 2025.
>
> Due to Twitter's Terms of Service, this data archive is accessible only to the University of Miami community …

Maps to “Rights” in Dublin Core.

Access Conditions: Crosswalks

Crosswalks	
Dublin Core	Rights
EAD	<accessrestrict> <userestrict>
MARC	506
MODS	<accessCondition>
schema.org	schema:license schema:isAccessiblrForFree

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

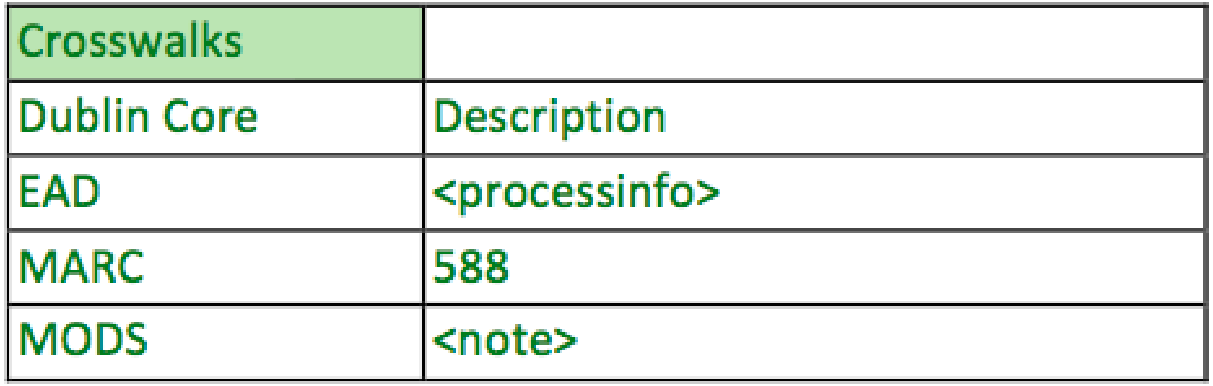
Description based on archived web page captured Sept. 22, 2016; title from title screen (viewed Oct. 27, 2016)

Title from home page last updated June 21, 2012 (viewed June 22, 2012)

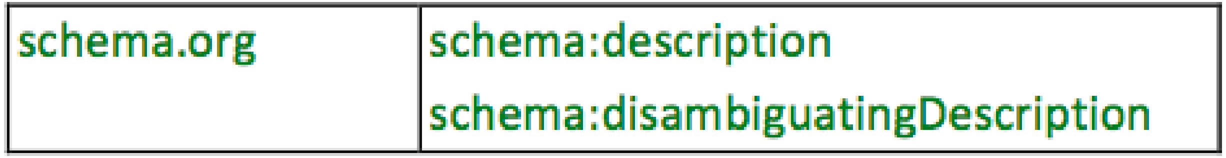
Title from home page (viewed on Oct. 11, 2007)

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

WAM data elements (14)

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

* = 9 of 14 element names/meanings match Dublin Core

PUBLICATION IN LATE JULY

Three simultaneous reports

- Best practices for **descriptive metadata**
  - With data dictionary
- **User needs**
  - With annotated bibliography
- **Tools**
  - With evaluation grids

Q&A

IIPC, Web Archiving Week
16 June 2017

For more information, please contact:

Jackie Dooley
Program Officer, OCLC Research

dooleyj@oclc.org
@minniedw

oc.lc/wam

Because what is known must be shared.SM