Descriptive Metadata for Web Archiving

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

Jackie Dooley and Kate Bowers

OCLC

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

Jackie Dooley
OCLC Research

Kate Bowers
Harvard University

February 2018

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 978-1-55653-016-6
DOI: 10.25333/C3005C
OCLC Control Number: 1021307921

ORCID iDs
Jackie Dooley https://orcid.org/0000-0003-4815-0086
Kate Bowers https://orcid.org/0000-0002-2160-583X

Please direct correspondence to:
OCLC Research
oclcresearch@oclc.org

Suggested citation:
Dooley, Jackie, and Kate Bowers. 2018. *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*. Dublin, OH: OCLC Research. https://doi.org/10.25333/C3005C.

CONTENTS

EXECUTIVE SUMMARY

"The ways and means of conducting scholarly inquiry are experiencing fundamental change, with consequences for scholarly communication and ultimately, the scholarly record—the curated account of past scholarly endeavor." Thus begins the influential OCLC Research report *The Evolving Scholarly Record,* which articulates the rapid evolution of library and archives content from print to digital, and the profound implications for our collections and services. The content of many digital resources is inherently mutable and increasingly is being recorded and disseminated only on the Internet.

This renders it imperative that we preserve web content on a timely basis if we are to maintain the integrity and continuity of the historical, cultural and scholarly records. Libraries and archives have lengthy experience contending with the preservation challenges of short-lived media, but the longevity of a website is even more volatile: sites tend to be updated, expanded or reformulated repeatedly over time, and any site may disappear with no warning. If not preserved on a timely basis, a significant percentage of web content simply ceases to exist.

The work arose in part from two recent surveys—one of end users of archived web content and the other of web archiving practitioners—both of which showed that lack of a common approach to creating metadata was the most widely shared challenge across the web archiving community.

In response, OCLC Research established a Web Archiving Metadata Working Group (WAM) to develop recommendations for descriptive metadata. Our approach is tailored to the unique characteristics of archived websites, with an eye to helping institutions improve the consistency and efficiency of their metadata practices in this emerging area.

We began with a literature review to gain insight into the information of highest value to users. One highlight is that the context within which a website or collection was archived is of vital importance to end users for understanding potential uses of the content. In response to needs stated by metadata practitioners, we provide a bridge between bibliographic and archival approaches to description—an approach that grows in importance as new types of digital content permeate our collections.

Libraries and archives create metadata records for both live and archived sites, using bibliographic and archival approaches to describe both single websites and archived collections. Each community employs its own standards and practices for cataloging and description, though hybrid techniques are also in use. We defined a lean set of data elements with usage notes to guide the preparation of data content. These can be used in concert with existing standards that have far more granular data elements. It is a community-neutral, standards-neutral, scalable approach that does not mandate in-depth description or extensive changes to records over time.

Potential use cases include *scholars* building personal archives of websites for research purposes; *libraries* using Resource Description and Access (RDA)/MARC that seek specific guidance on the elements and content that are most pertinent to description of web content; *archives* that map their Describing Archives: A Content Standard (DACS)-based MARC 21 records and/or Encoded Archival Description (EAD)-encoded finding aids to the more simplified structure of a web tool such as Archive-It or a digital repository; *digital repositories* that encode metadata for web content in Metadata Object Description Schema (MODS) without reference to any content standard; and *Archive-It* users who seek guidance on creating content for Dublin Core elements.

INTRODUCTION

"Web Archiving operates at the frontier of capturing and preserving our cultural and historical record."
Sara Day Thomson, 2016.*

"It is far easier to find an example of a film from 1924 than a website from 1994."
M.S. Ankerson, 2011.†

"The ways and means of conducting scholarly inquiry are experiencing fundamental change, with consequences for scholarly communication and ultimately, the scholarly record—the curated account of past scholarly endeavor." Thus begins the influential OCLC Research report *The Evolving Scholarly Record*, which articulates the rapid evolution of content from print to digital, and the profound implications for library collections and services.[1,2] The content of many digital resources is inherently mutable and increasingly is recorded and disseminated only on the Internet.

This renders it imperative that we preserve web content on a timely basis if we are to maintain the integrity and continuity of the historical, cultural and scholarly records. Libraries and archives have lengthy experience contending with the preservation challenges of short-lived media, but the longevity of a website is even more volatile: sites tend to be updated, expanded or reformulated repeatedly over time, and any site may disappear with no warning. If not preserved on a timely basis, a significant percentage of web content simply ceases to exist.

The preservation challenges are daunting, as are those relating to descriptive metadata to render a web resource discoverable. *Descriptive Metadata for Web Archiving* was prepared in response to this need. The work arose in part from two recent surveys—one of end users of archived web content and the other of web archiving practitioners—both of which concluded that lack of a common approach to creating metadata is the most widely shared challenge across the web archiving community.[3,4]

In response, OCLC Research established a Web Archiving Metadata Working Group (WAM) to develop recommendations for descriptive metadata and facilitate discovery of archived web content.[5] Our approach is tailored to the unique characteristics of archived websites, with an eye to helping institutions improve the consistency and efficiency of their metadata practices in this emerging area.

---

* Thomson, Sara Day. 2016. "Surveying the Domain: Three Days with the Web Archiving Team." *The British Library Web Archive Blog.* Posted 14 September 2016. http://blogs.bl.uk/webarchive/2016/09/surveying-the-domain-three-days-with-the-web-archiving-team-.html.

† Ankerson, Megan Sapnar. 2011. "Writing Web Histories with an Eye on the Analog Past." *New Media & Society*. doi:10.1177/1461444811414834.

Two basic definitions define the context for this work: *Web archiving* is "the process of collecting, preserving, and providing enduring access to web content,"[6] while *web archives* has two definitions: "1. preserved copies of live web content collected for permanent retention and access 2. an organization devoted principally to the collection and preservation of web content."[7]

The working group recognized the importance of gaining a clear understanding of the needs of users of archived web content, and we have taken this into account throughout the project. We began with a literature review to gain insight into the information of highest value. One highlight is that the context within which a website or collection was archived is of vital importance to end users for determining potential uses. In response to needs stated by metadata practitioners, we have provided a bridge between bibliographic and archival approaches to description.

To ensure that we would not duplicate work planned by others, we consulted with the International Internet Preservation Consortium,[8] the Society of American Archivists Web Archiving Section,[9] and the Internet Archive's Archive-It[10] program before embarking on this project. All encouraged us to proceed. We also engaged in ongoing communications with the community, both to report on progress and to solicit feedback. During 2016 we published a work-in-progress report,[11] presented at numerous conferences, and posted updates to listservs, Twitter, and the OCLC Research blog, hangingtogether.org.[12]

Our work resulted in this publication, as well as two other publications: *Descriptive Metadata of Web Archiving: Literature Review of User Needs*[13] and *Descriptive Metadata of Archiving: Review of Harvesting Tools*.[14]

Objectives and Use Cases

WAM's overall objective was to develop practices for creating consistent metadata that address the unique characteristics of websites and collections. More specifically:

1. Develop community-neutral, standards-neutral practices for descriptive metadata for archived web content, taking into account the needs of end users and metadata practitioners.

2. Define a lean set of data elements with usage notes to guide the preparation of data content.

3. Ensure that the data elements can be used in concert with other standards that have far more granular data element sets.

4. Provide a bridge between bibliographic and archival approaches to description.

5. Use a scalable approach that requires neither in-depth description nor extensive changes to records over time.

6. Enable practitioners to have confidence that they are contributing to the application of consistent practice in this emerging area.

Descriptive metadata is only one type of metadata relevant for web archiving; preservation, technical and administrative metadata also are essential for enabling discovery and providing context. These were beyond the scope of WAM's work.

WAM's recommended practices can be used by any institution or person with a need to describe web content. Some potential use cases:

- *Scholars* building personal archives of websites for research purposes
- *Librarie*s and archives using RDA/MARC that seek specific guidance on the elements and content that are most pertinent to description of web content
- *Archives and libraries* having a need to map their DACS-based MARC records and/or EAD-encoded finding aids to the more simplified structure of a digital repository or a web tool such as Archive-It
- *Digital repositories* encoding metadata for web content in MODS[15] without reference to any content standard
- *Archive-It* users seeking guidance on creating content for Dublin Core elements

Institutions will be able to map WAM's lean element set to and from more granular content and structure standards. The crosswalks associated with each data element are meant to facilitate such conversion.

Bridging Descriptive Practices

Our literature review showed that metadata practitioners feel a need to bridge bibliographic and archival approaches to description—an approach that grows in importance as new types of digital content permeate our collections. This led WAM to incorporate such an approach into our objectives. We recognized at the outset that attempting to do so would be challenging, and that our recommendations might seem problematic to practitioners with deep experience in a single standards context. In studying differences across metadata communities, we sought to identify ways in which practices might be rendered compatible.

The descriptive practices prevalent in libraries and archives differ in many ways. At the most basic level, each community has its own set of standards for both description and data structure.[16] Metadata for web content is created in various organizational units depending on where responsibility for metadata and/or custody of materials is situated, including library cataloging, archives and digital repositories. Based on our sampling of extant records taken from WorldCat®, ArchiveGrid and Archive-It, data elements are being both selected and encoded in ways that do not follow the corresponding descriptive standards, nor do they necessarily include features that our literature review revealed are valuable to end users.

Further, many websites constitute complex research objects that may include a variety of formats and content types such as images, data and publications, as well as a complex array of links to external sites. This leads to the realization that the context of creation and use—a concept that has long been foundational in the archival landscape—is central to understanding such resources.[17] It is explored in some detail in *The Archival Advantage* as one area of archival expertise that applies to digital content, including websites.[18]

We encourage readers to consider the value of blending practices for web content, given that it may be described while either "live" or archived, has characteristics associated with both bibliographic and archival practice, and may be described at the item level, collection level, or both. In some settings, this work may offer opportunities for collaboration across multiple organizational units.

ARCHIVED AND LIVE WEBSITES

WAM's charge was to address metadata for web archiving; thus, our recommendations are based in that assumption. We readily acknowledge, however, that many libraries (and some archives) create metadata for live sites. The same data elements apply, but the specifics of metadata content and meaning often vary.

The essential difference between live and archived web content is that, in the latter case, an intervention has taken place that changes the very nature of the resource: each crawled version becomes a fixed object, preserved for the future in a particular digital location and associated with any other versions that have been captured. The *Code of Best Practices in Fair Use for Academic and Research Libraries* articulates this clearly:

> Selecting and collecting material from the Internet in this way is highly transformative. The collecting library takes a historical snapshot of a dynamic and ephemeral object and places the collected impression of the site into a new context: a curated historical archive. Material posted to the Internet typically serves a time-limited purpose and targets a distinct network of users, while its library-held counterpart will document the site for a wide variety of patrons over time. A scholar perusing a collection of archived web pages on the Free Tibet movement, or examining the evolution of educational information on a communicable disease, seeks and encounters that material for a very different purpose than the creators originally intended. Preserving such work can also be considered strongly transformative in itself, separate from any way that future patrons may access it. Authors of online materials often have a specific objective and a particular audience in mind; libraries that collect this material serve a different and broader purpose and a different and broader network of users. Libraries collect not only for a wide range of purposes today, but also for unanticipated uses by future researchers.[19]

These are some specific aspects of descriptive metadata that vary for live and archived content:

- Live sites inherently provide the most current information, whereas archived content gains research value over time. When a site has been repeatedly harvested, we can answer research questions such as “How have the Environmental Protection Agency’s policies differed across the Bush, Obama and Trump administrations?”

- A live website has a publisher but no inherent relationship to the institution that creates the metadata. In contrast, the institution that has collected and taken responsibility for archived content should be prominently identified in the metadata.

- By definition, metadata for a live site describes the current version; if the site is taken down, the seed URL will lead to a dead end.[20] When a site is harvested, whether once or repeatedly, the seed will lead to the archived version. Quality control procedures are used to identify scheduled crawls that must be re-evaluated. Metadata often includes an access URL (such as for a landing page or a collection in Archive-It) in addition to the seed(s).

- The date recorded in metadata for a live site may be that on which it was viewed and described, or, if known, the date it first went live. For archived content, the dates of capture are essential. Harvesting tools capture these dates, which enables viewing the chronological evolution of a site over time.

- Most live sites are publicly accessible and have no restrictions on access, while some sites are archived before an access mechanism has been set in place and can be viewed only onsite.

For these and other reasons, live and archived versions of the same single website warrant separate metadata records. On the other hand, when a host institution wants to make both versions accessible, it is far more practical from a staffing perspective, as well as more amenable within the limitations of our current metadata systems, to use a single metadata record that notes the existence and characteristics of an archived version. The single record should include, at minimum, the responsible institution and a stable access URL. See, for example, the approach taken by the New York Art Resources Consortium.[21]

BIBLIOGRAPHIC AND ARCHIVAL DESCRIPTION

Approaches to description vary substantially between the two traditions of bibliographic and archival description.

Libraries principally (but far from exclusively) catalog individual resources at the title level, and many resources are expected to, in effect, describe themselves via features such as a title page and verso (or CD label, sheet music cover or film credits) from which data elements can be transcribed. The nature of the content is revealed principally by the title (if it is descriptive) and subject terms, though notes describing content have become more common over time. Digital content that is openly accessible is considered published.

In contrast, archivists generally describe a collection of unpublished materials that are related by provenance in a single metadata record or multilevel finding aid. They routinely take descriptive information from sources both internal and external to the collection (such as biographical or historical secondary material). Titles are devised rather than transcribed. Extensive free-text notes are routinely used to describe both content and context of the material. Archival content that is born-digital or digitized is not considered to be published when made available online.

Many collections of archived web resources are thematically selected and are thus more similar to "artificial" archival collections than to those related organically by provenance.[22]

Commonality of practice exists for some types of information, particularly access points. These include vocabularies for subject or genre/form access; authority files for standardization of personal, organizational and geographic names; and international standards for expressing elements such as date and language in machine-actionable ways.[23]

In developing these recommendations for metadata for archived web content, we saw the blending of these practices in both local guidelines and extant records, reflecting the inherently hybrid nature of websites and collections. WAM's recommended data elements are equally relevant for description of materials in both libraries and archives, and at both the item and collection levels.

COLLECTION-LEVEL AND SITE-LEVEL DESCRIPTION

Either or both levels of description can be appropriate for describing archived web content, depending on the circumstances. The approach taken may depend on the type of institution creating the metadata (library, archive, museum, individual researcher), the human resources available to accomplish the work or other factors. To date, libraries have most often described *live* single sites in individual metadata

records to provide access to these high-value resources via their discovery systems. In contrast, archives almost exclusively describe web content that they have harvested for long-term preservation. Collection-level description is the norm.

The site-level approach can present an insurmountable workload, however, for institutions that harvest numerous sites or that are thinly resourced. This is particularly true if the descriptions are detailed or if a standard is employed that carries high overhead.

Many libraries “collect” thematically, such as to document a significant event or add to an existing collecting strength. In this context, collection-level description such as that routinely used in archives could be a cost-effective alternative to site-level description. Further, collective description can provide the context for an aggregation by describing its overall scope, significance and shared subjects in a way that single-site records cannot accomplish.

All WAM data elements are applicable to both site-level and collection-level description (whether the content is live or archived), though the nature of the content may differ. Three examples:

- The title of a single site usually is a transcription of prominent text from the site itself, whereas the title of a web collection is devised by the institution that collected the content.
- The creator of a single website, such as an institutional home page, blog or twitter feed, usually is easily identified unless purposely anonymous, while a collection of websites focused on a current event or topic rarely has an overall creator.
- The date recorded for a collection-level description of a thematic collection may include a range of dates between which harvesting was done. In contrast, the data recorded for a single live site may be that on which it was viewed and described.

A collection-level description can be a baseline for discovery, to be associated with single-site descriptions as feasible and justifiable. The nature of a collection may obviate the need for detailed site-level description. The records of an organization, for example, may warrant only collection-level description supplemented by a site-level title and URL, such as in a multilevel finding aid or Archive-It record. For an archived collection that encompasses a large number of sites, it would be impractical and potentially confusing to include a lengthy list of URLs in a single collection-level record.

Preparatory Research

We performed three studies prior to devising the data elements and usage guidelines so that our recommendations would be based in theory and practice rather than being speculative:

- A review of literature that covers the descriptive metadata needs of both end users of web archives and practitioners who create and manage such metadata
- An analysis of tools for capturing web content that appeared to have some level of metadata functionality
- An analysis of the applicability to web archiving of descriptive standards widely used by the library and archives communities, institutional guidelines contributed by seven institutions and sample metadata records gathered from a variety of sources

The conclusions we derived from each study are briefly summarized in the following sections. Separate reports for the literature review and analysis of tools were published simultaneously with this present one.

USER NEEDS LITERATURE REVIEW

Descriptive Metadata for Web Archiving: Literature Review of User Needs describes in detail the needs of end users and metadata practitioners as articulated in the scholarly literature and other current sources.[24] While there is clear overlap across the two communities, some needs vary. In devising the WAM data dictionary and recommendations for its implementation, the findings from this work were in the forefront of our thinking.

We were able to draw some conclusions relevant to expressed metadata needs of each type of user. Here are a few from each user group, with those that have direct *implications for descriptive practice in italics*:

End Users

- The literature on end-user needs focuses principally on academic researchers in a wide variety of disciplines.
- Users express a strong need for *provenance information* to add context beyond standard descriptive metadata elements. This reflects a widespread desire for transparency around the decision-making process in selecting sites for capture and building collections, as well as the *completeness of individual captures* and *changes that occur over time*.
- Given the ease and ubiquity of access to the open web, *restrictions on access*—such as being limited to onsite viewing in a library—are both mystifying and frustrating to users. A closely related need is clarity about *intellectual property rights* vis-à-vis re-use of content.
- Archived web data often is provided in a way that exceeds the limits of users' technical knowledge, constituting a widespread barrier to use. This could be alleviated by developing user-friendly tools and interfaces to enable content to be more readily accessed and repurposed.
- A need for user support services derives from the complexity of accessing and using web archives.
- Libraries and archives should actively engage in outreach to both current and potential web archives users: first, to provide an initial understanding of what web archives are, and how to find and use them; and second, to better understand users' research challenges in order to find ways to ameliorate them.

Metadata Practitioners

- *Scalable practices* are necessary because staff resources for this work are extremely limited at most institutions.

- An array of existing *library and archival standards* for data structure and content are being used for web archiving descriptive metadata.
- *Bibliographic, archival and hybrid approaches* are in use. The need to find *ways to bridge these* is widely perceived.
- The *application of standards* to archived web content is *highly inconsistent* in terms of both data elements and their content.
- Metadata describing archived web content is often delivered via multiple discovery systems, suggesting the need for smooth processes to re-use metadata generated across systems.
- Experimentation with nontraditional approaches is underway.

The literature review revealed pervasive needs that were beyond the scope of WAM's charge, such as the importance of better discovery tools, the challenge of archived web content being dispersed across many silos, the absence of software tools to facilitate sophisticated data analysis, and lack of user awareness that libraries and archives are preserving a great deal of web content. See the section on Future Research Needs for observations about these and other issues.

ANALYSIS OF HARVESTING TOOLS

The report *Descriptive Metadata for Web Archiving: Review of Harvesting Tools*,[25] is an analysis of 11 web harvesting tools with an eye to their functionality for extracting descriptive metadata. We pursued this work upon realizing that some practitioners wanted to know whether it would be possible to extract descriptive metadata from the crawled files rather than having to create and/or rekey it. The tools we reviewed are Archive-It, Heritrix, HTTrack, Memento, Netarchive Suite, SiteStory, Social Feed Manager, Wayback Machine, Web Archive Discovery, Web Curator Tool, and Webrecorder.

We came to several conclusions:

- Most tools built for web archives focus on capturing and storing technical metadata for accurate transmission and re-creation but capture minimal descriptive metadata, at least partially because so little exists in the captured files. *Descriptive metadata therefore must be created manually, either within the tool or externally*.
- The title of a site (as recorded in its metadata) and the date of capture are routinely captured, but it may not be possible to extract them automatically. Titles are sometimes unhelpful, such as "home page" or "title."
- Not all tools define descriptive metadata in the same way.
- The hope for auto-generation of descriptive metadata may be fruitless unless or until creators of textual web pages routinely embed more metadata that can be available for capture.
- Development of new tools and enhancement of existing ones are actively underway.

Ultimately, we came to wonder whether harvesting is the most appropriate stage of the web archiving process during which to add descriptive metadata for most types of content. Given the lack of metadata features in current web archiving tools, perhaps there is a clearer path forward for approaches that leverage external services and APIs.

ANALYSIS OF STANDARDS, GUIDELINES AND EXTANT RECORDS

We analyzed three standards, seven institutional sets of guidelines and a wide variety of extant metadata records. This analysis grounded WAM members in the data elements commonly used by libraries and archives that can apply to website description, approaches being taken by some individual institutions for describing archived websites, and developing a sense of the variations in overall practice.[26]

Three patterns emerged:

- Existing descriptive standards generally do not address the unique characteristics of either live or archived websites.
- Institutional metadata guidelines vary widely in both the elements included and in the choice of content within those elements.
- Some metadata practitioners follow bibliographic traditions, others take an archival approach (such as describing a collection of sites in a single metadata record), and hybrid approaches combining characteristics of both are common.

The standards we reviewed are Describing Archives: A Content Standard (DACS),[27] the Resource Description and Access (RDA)-based guidelines for describing integrating resources[28] and Dublin Core.[29] They differ greatly in purpose and approach, revealing a clear sense of the differences across standards-based descriptive practices.

- As the core *content standard* for library cataloging, RDA takes a fully bibliographic approach. The guidelines for "integrating resources" (i.e., those with content that is updated over time) address the characteristics of live websites. Some instructions—although defensible in theory—may be impractical for frequently-changing archived content; for example, Rule IR2.6.2.3 directs the cataloger to "Change the title proper to reflect the current iteration of an integrating resource if there is a change on a subsequent iteration."
- DACS is a *content standard* intended principally for describing groups of archival materials and takes a multilevel approach. For example, a description may include data elements describing the entire group of materials, a subset or individual items. The same data elements are used at all levels, though more elements usually are included at the top level of description than at lower levels.
- Dublin Core is a data *structure standard* that contains 15 elements for use in resource description to encourage a standard approach to simplified metadata for use by any community of practice. It does not include guidance for devising the content of elements. Dublin Core is widely used for describing digital objects, including by Archive-It users.[30]

Seven institutions responded to our call for local guidelines. All were developed for use with one or more data structure standards. In addition to the three standards we reviewed in depth, the latter include Encoded Archival Description (EAD),[31] MARC 21[32] and MODS.[33]

Our analysis revealed wide variation in the types of content included in—or omitted from—each set of guidelines, and to which data element each type of content was assigned. Only four data elements were present in all seven: title, creator, description and URL. Most guidelines contain only a simple list of data elements, leaving practitioners on their own to devise content that is both clear and consistent.

WAM also informally sampled extant descriptions of archived websites found in WorldCat® (MARC records),[34] ArchiveGrid (MARC records and finding aids),[35] Archive-It (Dublin Core)[36] and institution-specific discovery systems. By doing so, we were able to take into account the practices of some institutions that may not have formal guidelines. Our analysis of metadata created by a wide variety of institutions did not uncover any additional data elements, but it did reveal an even wider inconsistency of practice across the community than we had found in guidelines.

Questions about data elements emerged based on the inconsistencies we found. For example:

- *Website creator/owner:* Is this the publisher, creator, subject or all three?
- *Host institution:* Should the institution that selects, harvests and hosts the site be considered the repository, collector, publisher, selector or creator?
- *Title:* Should it be transcribed verbatim from the head of the site? Edited to clarify the nature/scope of the site? Should acronyms be spelled out? Should the title include a phrase such as “Website of the ⋯”?
- *Dates:* Which dates are both important and feasible to record? Beginning/end of the site's existence, date(s) of capture, dates of content or copyright date?
- *Extent:* How is this most usefully expressed? 1 archived website, 1 online resource, 6.25 GB or approximately 300 websites?
- *Provenance:* Does provenance refer to the creator of the website, the repository that harvests the site and hosts the web archive, ways in which the site evolved, frequency and dates of capture, or all of these?
- *Appraisal:* Does appraisal mean the reason why the site warrants being archived, a collection of sites named by the repository or the parts of the site that were and were not harvested?
- *Format:* Is it important that the description clearly state that the resource *is* a web archive? If so, how best to do this—in the title, extent or description?
- *URL:* Which URLs should be included? Seeds, access or landing page?
- *MARC 21 type of record[37]:* When coded in the MARC 21 format, should a website be considered a continuing resource, integrating resource, electronic resource, textual publication, mixed material, manuscript, or any of these, depending on the context?[38]

All of the above-listed possible choices were seen in institutional guidelines or extant records.

After establishing an understanding of the needs of users and current metadata practices for web archives, we developed the data elements and usage guidelines that follow.

Data Dictionary

CRITERIA FOR INCLUSION OF DATA ELEMENTS

In developing this data dictionary, we considered two key questions: Which types of content are most important to include in a metadata record describing an archived website or collection of sites, and what data element should be designated for each of these content types?

To guide our decisions about data elements to include and exclude, we followed these criteria:

- The data element set can be used either on its own or in concert with library and archival standards that are far more granular.
- The element set is lean to enable facilitate scalable metadata creation.
- Data element names and definitions were adopted or adapted from existing standards whenever feasible to enhance compatibility and to encourage consistency across discovery environments.
- Usage notes explain the recommended application of each element to assist practitioners in creating metadata whose meaning is completely clear to end users.
- Common elements that are key to identification and discovery of any resource are included (such as Creator, Date, Subject and Title).
- All other elements must have clear applicability for description of archived websites (such as Description, Rights and URL).
- Elements were *excluded* that rarely (if ever) appear in institutional guidelines and/or extant metadata records for archived web content and that have no special meaning in this context (such as audience, publisher and statement of responsibility).
- The same element is used to express a particular concept at all levels of description, in accordance with the multilevel description principles expressed in archival standards such as DACS and EAD.

Those familiar with Dublin Core will recognize substantial similarities with the WAM data element set. Rather than simply recommend use of Dublin Core for archived web content, however, we carefully considered which DC elements were applicable and which might be adaptable. In the end, the name and meaning of eight of the 15 DC elements match the WAM data dictionary: Contributor, Creator, Date, Description, Language, Relation, Subject and Title. Six of the 14 WAM elements are different in name and/or meaning: Collector, Extent, Genre/Form, Rights, Source of Description and URL.

In developing a lean set of data elements meant to be useable with a variety of standards, we did not extend the elements by using techniques such as modifiers or subfields. Users who require machine-

actionable data for elements that may have a variety of meanings, such as Date, Extent and URL, may want to investigate schemes that enable attributes and other qualifiers, such as Dublin Core, EAD, schema.org[39] or MODS.

DATA ELEMENTS AND USAGE GUIDELINES

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

These 14 data elements are listed below in alphabetical order. Each includes a definition, usage notes for preparing content of the element, examples taken from extant metadata records,[40] and brief crosswalks to Dublin Core, EAD, MARC 21, MODS, and schema.org. Each example was chosen for its usefulness in illustrating the respective element.

Each definition states that the element is used to describe "an archived website or collection; " this emphasizes the primary intended purpose of these recommendations. Thirteen of the elements are equally applicable in descriptions of single sites and collections, though both the nature of the content and number of elements may differ at the two levels.[41]

The usage notes interpret each data element in the specific context of web description. In that regard, they have value for those who will map these elements to a more granular data structure rather than using the WAM elements *per se*.

Unlike most other standards referenced in this report, no set of required or core elements is designated. Institutions that apply the data elements in concert with one or more standards should look to their stated minimum requirements. That said, Title and access URL are the absolute minimum necessary for identification and discovery of any archived site or collection. These are frequently the only elements included at the seed level in a multilevel description (such as in an archival finding aid or Archive-It). When describing a single site or collection in a scalable manner, other strongly recommended elements are Collector, Creator, Date of crawl(s) and Description.

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).

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

This is the only element for which selecting the most appropriate name was challenging, and so an explanation of the process by which we arrived at it may be useful. After reviewing institutional guidelines and numerous descriptive metadata records for web content, we compiled the following terms that are used to express the roles listed above.

- *Repository*: Required in all DACS-conforming records for archival collections to name the sole institution that collected, has custody of and provides access to the material. In contrast, "digital repository" is used by libraries for their storage environment.[42] Differing usage of this term by two allied communities would make its use problematic in an agnostic data dictionary.
- *Location*: Used by some institutions to record the name and address of the institution or repository that holds the resource.
- *Collected By*: Used by the Archive-It service for the name of the partner institution (i.e., subscriber) that selected the harvested content.
- *Collector*: Used by the Archive-It service for the name of the partner institution or its subunit that selected and manages the content.
- *Selector*: Used in multiple sets of institutional guidelines. Those using Archive-It convert it to "collector" or "collected by."
- *Owner*: Used in some institutional guidelines for the institution responsible for archiving the website or collection.
- *Publisher*: Used in one set of institutional guidelines for the person or entity responsible for publishing the archived website or content.[43]

Although the archival definition of "repository" encompasses many of these activities, we came to the conclusion that **Collector** is the most appropriate term for an element set intended to be agnostic across user communities.

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

Title: April 16 Web Archive
Contributor: Center for Digital Discourse and Culture (Virginia Tech)
Collector: Crisis, Tragedy and Discovery Network (Virginia Tech)

Title: Globalchange.gov *[single site]*
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 [single site]
Collector: New York Art Resources Consortium

The organization that selects the content for archiving is both **Creator** and **Collector** when describing its own harvested websites, though a subunit of the organization may be appropriate as the **Collector**.

Creator: Cornell University
Title: Cornell University Archives Web Archive
Collector: Cornell University Archives

In some extant metadata records, the web archiving software platform or tool used to manage the content (such as Archive-It) is designated as the collector, or even the creator. The tool or platform plays one or more related technical roles—and may warrant mention in a note—but it does not select or curate the content. Further, the tool or platform used may change: an institution may voluntarily stop using a particular tool and will then have to transform the access URLs in all metadata records—or the change may not be voluntary, such as when the California Digital Library decommissioned its former Web Archiving Service.[44] Therefore, it is rarely appropriate to designate the tool or platform as **Collector** nor **Creator**.

One situation exists in which the harvesting platform or tool is the **Collector**: when the organization that manages the tool takes responsibility for assembling and hosting a collection from disparate sources, as the Internet Archive sometimes has done.

Title: Internet Archive Global Events
Description: These collections of global events have been created by the Archive-It team in conjunction with curators and subject matter experts from institutions around the world.
Collector: Internet Archive

Title: Earthquake in Haiti Web Archive
Description: This collection is currently documenting the events of the January 2010 earthquake in Haiti and the aftermath, including the rescue efforts from around the world and the stories and circumstances of the Haitian people. Archive-It partners Library of Congress, Bibliothèque nationale de France, Virginia Tech and CTRnet, and University of Texas Libraries have all contributed websites for this collection.
Contributor: Library of Congress *[and others listed above]*
Collector: Internet Archive

Title: Political Campaign Web Archive
Description: The Political TV Ad Archive, a project of the Internet Archive, collects political TV ads and social media sites in key 2016 primary election states, unlocking the metadata underneath and highlighting quality journalism to provide journalists, civic organizations, academics, and the general public with reliable information on who is trying to influence them & how.
Contributor: Political TV Ad Archive
Collector: Internet Archive

Crosswalks	
Dublin Core	Contributor
EAD	<repository>
MARC 21	583 subfield h 710 852 subfield a 852 subfield b
MODS	<location><physicalLocation>

schema.org	schema:OwnershipInfo

CONTRIBUTOR

Definition: An organization or person secondarily responsible for the content of an archived website or collection.

Use **Contributor** for entities that have made meaningful but secondary contributions to the content of a website or collection and that are not specified in the **Creator** element.

Title: Read the Spirit Web Archives *[single site]*
Contributor: Baker, Wayne E. *[one segment of the site is his blog]*

[no Creator element]
Title: April 16 Web Archive
Contributor: Center for Digital Discourse and Culture (Virginia Tech)
Collector: Crisis, Tragedy and Discovery Network (Virginia Tech)

[no Creator element]
Title: Human Rights Web Archive
Contributor: Columbia University. Libraries. Web Resources Collection Program
Collector: Columbia University. Center for Human Rights Documentation and Research

Consider indicating the role played by a contributor. Examples include author, contributor, distributor or illustrator.[45]

Title: Globalchange.gov
Contributor: United States. Government Publishing Office, distributor

Crosswalks	
Dublin Core	contributor
EAD	<controlaccess><persname> <controlaccess><corpname> <controlaccess><famname>
MARC 21	700
MODS	<name>

schema.org	schema:contributor

CREATOR

Definition: An organization or person principally responsible for creating the intellectual content of an archived website or collection.

Include this element for an organization only when it clearly has principal responsibility for having created the intellectual content. In case of doubt, use **Contributor**.

Creator: Occupy the Future *[organization name]*
Title: Occupy the Future Web Archive

Creator: Association for Research into Crimes against Art
Title: ARCAblog : promoting the study and research of art crime and cultural heritage protection

An individual person is the **Creator** only when he/she is clearly the creator of the intellectual content, such as an individual's personal blog or Twitter feed.

Creator: Sherman, Aliza
Title: Aliza Sherman rants and raves

Many sites about an individual are created by someone else (who may or may not be named), such as those for politicians, authors, musicians and other public figures. The individual's name is often the title of the site and should be repeated in the **Subject** element.

[no Creator element]
Title: Jacqui Lambie
Subject: Jacqui Lambie, 1971-

When describing a collection of sites related to each other only by subject, omit the **Creator** element.

[no Creator element]
Title: #blacklivesmatter Web Archive

[no Creator element]
Title: Human Rights Web Archive

If two entities share principal responsibility, place them both in **Creator** if the standards or software you use permits this. Otherwise, place one in the **Contributor** element. Use **Contributor** for all that have secondary responsibility.

Note: The software platform or tool used to manage the archived content (such as Archive-It, Facebook or Twitter) is not the **Creator**, because it did not create the intellectual content. (For more information, see **Collector**.)

Crosswalks	
Dublin Core	creator
EAD	<origination><persname> <origination><corpname> <origination><famname>
MARC 21	100, 110 700, 710
MODS	<creator>

schema.org	schema:creator

DATE

Definition: A single date or span of dates associated with an event in the lifecycle of an archived website or collection.

Use **Date** to record any known date or span of dates associated with an archived website or collection that will help users understand the content.

Always make clear the meaning of a **Date** element by adding appropriate wording to enable user understanding. Without this, any date for a website or collection is inherently ambiguous.

If known, include the date the site first went live and/or was taken down.

Site began in 2012.
Began in: 2012?
Site no longer active as of 2016

The date that a seed URL (or a collection of seeds) was crawled is essential. When a site will be crawled more than once, include the intended frequency and/or date range.

Archived May 2014
Archived since: May 2014
Date of first crawl: July 4, 2015
Captured 2013-2015
Captured 2010-ongoing
Captured monthly beginning January 2017
Captured 4 times between Jan 29, 2016 and Jan 5, 2017

All seeds are captured by harvesting software. In some contexts, seed URLs may appear on a calendar page in addition to (or instead of) the descriptive metadata record.

Be aware that copyright dates sometimes are not updated over time and thus may not match the currency of information on the site.

> Copyright 2012

The date a site was viewed for description should be placed in the **Source of Description** element.

> **Source of description**: Description based on archived web page captured Sept. 22, 2016; title from title screen (viewed Oct. 27, 2016)

It is advisable to express dates according to a standard such as ISO 8601 or in accordance with the conventions required by an encoding standard to enable them to be machine-actionable.[46]

Crosswalks	
Dublin Core	Date
EAD	<unitdate>
MARC 21	008 bytes 07-14 245 subfield f 260 subfield c 264 subfield c 583 subfield c
MODS	<dateIssued> <dateCreated> <dateCaptured> <copyrightDate> <dateOther>

schema.org	schema:dateCreated schema:dateModified schema:datePublished

DESCRIPTION

Definition: One or more notes explaining the content, context and other aspects of an archived website or collection.

The **Description** element is used for textual information on multiple aspects of the described content. Because it consists of unstructured text, this is where the content and context of the site or collection can be most clearly articulated.

To fulfill WAM's objective of a lean set of data elements, the **Description** element can contain any type of note. In this regard it matches Dublin Core, which is widely used to describe digital resource and also has a single description element. In contrast, content standards such as DACS and RDA enumerate many specific note fields that can be applied in conjunction with their companion data structure standards (EAD and MARC 21, respectively). When applying these recommendations with one or more detailed standards, use the more granular elements that they provide.

WAM's literature review describes the types of information of particular interest to users of web archives, many of which are appropriate in the **Description** element. A recurring theme across the

literature is the importance of what users refer to as “provenance” information, which includes the rationale for archiving a website, such as whether it is part of a broader thematic collection. Others match notes specified by DACS as described below.

Biographical or historical information about the organization or person responsible for creating the content, and a statement of scope and content (i.e., an abstract) are of preeminent importance in archival description and are essential for user understanding of website metadata. Other categories found in archival descriptions may be relevant, including custodial history, appraisal and accruals.

Some other desirable details specific to web content:

- A clear statement that the object of description is archived web content
- The reason for selecting the content for archiving, such as a notable event, or a legal or administrative mandate
- The nature of content on the live site that is absent from the archived version (such as file types that the crawler could not capture)

If your metadata system allows elements to be repeatable, consider splitting a lengthy **Description** into two or more occurrences.

The following examples are taken verbatim from extant metadata records created by web archiving institutions. Additional examples of **Description** elements for both collections and single sites are in appendix C.

Each of the following was chosen for its articulation of the context and scope of a *collection* of archived websites.

> Archive of Web sites of individuals, groups, the press, and institutions in the United States and from around the world in the aftermath of the attacks in the United States on September 11, 2001. The archive consists of over 30,000 selected Web sites archived from September 11, 2001 through December 1, 2001, and is intended to preserve the Internet reaction from U.S. and non-U.S. government sites; press, corporate/business, portal, charity/civic, advocacy/interest, religious, school/educational, individual/volunteer, and professional organization sites. Browse access to descriptions of approximately 2,300 Web sites is available, as well as a list of all sites archived.

> The UC Davis Web Archives preserves websites in the “ucdavis.edu” domain to document the history of the University's activities and accomplishments. Special Collections, as the repository for the University Archives, collects records of historical value for the University. Previously, many of these records have been in print form, now much of that information can be found on campus websites. This project, started in 2011, captures the websites of the University's administration, schools and colleges, academic departments, administrative units, organized research units, intercollegiate athletics, and student organizations …

Many extant metadata records that describe a *single site* include no such notes. As the following examples illustrate, however, even a brief **Description** can add value for the user.

> Homepage for Senator J.H. Schwarz. Includes internal links only, although URLs for external links remain as part of the structure of the document. Republican state senator from Battle Creek, Michigan, 1987-2002.
>
> A blog focusing on art thefts, cultural heritage protection and restitution all over the world, centered around the activities and conferences of ARCA.

Crosswalks	
Dublin Core	Description
EAD	<abstract> <accruals> <acqinfo> <bioghist> <scopecontent>
MARC 21	500, 505, 520, 540, 541, 545, 555, 561, 584
MODS	<abstract> <note>

schema.org	schema:description

EXTENT

Definition: An indication of the size of an archived website or collection.

Extent may be expressed as the number of websites, the quantity of data stored (in megabytes, gigabytes or another measure) or the number and/or type of files harvested.

Use an approach that will give users a useful indication of how much content has been archived. Consider whether the number of sites included in a web archive might be more useful than the aggregate size of the files in bytes. If the site or collection consists of data that has potential to be mined, the latter may be the more useful measure.

Libraries often state the extent of a single website as “1 online resource” because this is the default value suggested by RDA. Because it is used for a wide variety of resource types, however, the meaning is vague. A clear statement that the material is web content is more transparent to users, whether or not they are potentially interested in this type of content.

> 1 archived website

The **Extent** of a collection of sites often is expressed as an approximate number of sites. This approach is easy to maintain and so is recommended for sites that will be harvested repeatedly and for collections to which additional sites will be added.

> 1 collection of archived websites
> Approximately 150 archived websites
> 1 collection of approximately 75 archived websites

WAM does not recommend the use of Extent statements that include the exact number of sites, bytes or files require high maintenance if sites will be re-harvested periodically, because the information will become inaccurate unless revised following each crawl. Consider carefully whether approaches such as the following (taken from extant records) are either scalable or helpful to users.

173.0 online resources, (with 4,431,438 documents (375 gb))
1888876.8 megabytes
Approximately 1 million digital files (59 GB)
80000.0 megabyte(s) (740,752 files in 162,982 folders)
1.8 terabytes (approximately 450 websites)

Crosswalks	
Dublin Core	format
EAD	<physdesc><extent>
MARC 21	300, 347
MODS	<physicalDescription><extent>

schema.org	schema:description

GENRE/FORM

Definition: A term specifying the type of content in an archived website or collection.

Genre/Form terminology can provide useful access to various content types, including archived websites and collections.[47]

The use of a controlled vocabulary is strongly recommended to encourage consistency. Select a vocabulary that is used by the communities likely to benefit from the described content.

The generic term for “web sites” or “websites” is the most commonly used term for websites and collections.

Web archives *[for archived content]*
Websites or Web sites *[for live or archived content]*

Community practice varies as to whether **Genre/Form** terms should be used in the singular or plural form. WAM recommends the singular form in accord with both *Genre/Form Terms for Library and Archival Materials*[48] (lcgft) and the Art & Architecture Thesaurus[49] (aat). Whichever approach you use, be consistent across your metadata records.

Genre/form terms for social media also may be applicable.

Blog
Facebook page
Twitter page
Instagram page
Social media

Consider using compound terms for sites specific to particular domains.

Government website
Corporate website
News website
Personal website

A variety of more specific genre/form terms can be appropriate for web content, just as they are for analog materials.

Exhibition
News article
Periodical
Press release
Video

Crosswalks	
Dublin Core	Type
EAD	<controlaccess><genreform>
MARC 21	655
MODS	<genre>

schema.org	schema:genre

LANGUAGE

Definition: The language(s) of the archived content, including visual and audio resources with language components.

Include the **Language** element for any website or collection in which language is essential for understanding the content. If it is in more than one language, state all that seem significant.

Text in English and Italian.
Portuguese, Spanish, Russian, Thai, and Vietnamese; interface in English.
Users may select English, Spanish, or Italian versions of the site.
English, with some posts in Italian or Spanish.
Content is in a variety of languages, including English, Arabic, Chinese, French, Indonesian.
Websites are mostly in English, with contents of some websites in Spanish and other languages.

Use of a controlled source of language names such as the ISO 639 series is strongly recommended to encourage consistency of usage to enable them to be machine-actionable.[50]

Crosswalks	
Dublin Core	Language
EAD	<langmaterial><language>
MARC 21	008 bytes 35-37, 041, 546
MODS	<language>

schema.org	**schema:inLanguage**

RELATION

Definition: Used to express part/whole relationships between a single archived website and any collection to which it belongs.

In the web context, the most common relationship is between a collection and the subgroups or items within it. When describing a single site that is part of a collection of archived sites, include the collection title in the **Relation** element to provide the context within which the site was collected.

Title: ARCAblog : promoting the study and research of art crime and cultural heritage protection
Relation: Restitution of Lost or Looted Art Web Archive

Title: Snowden's father plans to visit son in Russia “very soon”
Relation: International Whistleblower Web Archive

Title: California Proposition 29 imposes additional taxes on cigarettes for cancer research
Relation: UCLA Online Campaign Literature Web Archives

Alternatively, formulate the information as a free-text note.

Title: Harvard University Committee on Degrees in Folklore and Mythology archived website
Relation: Part of the Archives’ collection of Faculty of Arts and Sciences departmental websites.

Title: GlobalChange.gov
Relation: Part of the Federal Depository Library Program Web Archive

The whole/part relationship is equally relevant when an archived website is part of an analog archival collection or is related to other digital material.

Title: Website for Yale's 300th anniversary
Relation: Part of Yale University's 300th Anniversary Commemoration Records

Title: April 16 Web Archive
Relation: This collection was specifically developed as a complement to the April 16 Archive (www.april16archive.org), which is dedicated to collecting and preserving individual stories, images, and files related to the events of April 16.

Include a URL for the related resource, if available.

Crosswalks	
Dublin Core	Relation
EAD	<unittitle> (of parent component) <relatedmaterial>
MARC 21	580, 730, 773, 787
MODS	<relatedItem>

RIGHTS

Definition: Statements of legal rights and permissions granted by intellectual property law or other legal agreements.

The **Rights** element is used for two distinct types of information: conditions that restrict user *access* to the archived content, and whether permission by holders of copyright for *re-use* after access has been gained. As appropriate, both types may be specified in a single occurrence of **Rights**.

Conditions of access might include the need to make an appointment for onsite use or a specified period of time during which the content is restricted. Such conditions may be imposed by an archival repository, donor, other agency, privacy law or other legal statute.

> Contact the Beinecke Library to make an appointment to view this content.
>
> Researchers seeking to examine archival materials are strongly encouraged to make an appointment. The Director, or an office of origin, may place restrictions on the use of some or all of its records. The extent and length of the restriction will be determined by the Director, office of origin, and the Archivist.
>
> 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. It can be accessed via an IP-restricted portal in the University of Miami Scholarly Repository. For specific questions, please contact chc@miami.edu.

When access is open, indicate this.

> This content is freely available online.
> No restrictions on access.
> Ford.com archived websites are searchable without restriction.

State whatever is known about holders of copyrights to either permit or restrict re-use of the archived content.[51] Even if the rights are unclear or unknown, state what you know.

> © Smith College. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
>
> The copyright and related rights status of the items in this collection have not been evaluated. Please refer to the individual items in this collection for more information about the rights holder. You are free to use this Item in any way that is permitted by the copyright and related rights legislation that applies to your use.

The copyright status of this material is not known. Users are responsible for determining whether or not the material may be re-used for publication.

Consider using or adapting the copyright statements available from RightsStatement.org.[52]

Crosswalks	
Dublin Core	Rights
EAD	<accessrestrict> <userestrict>
MARC 21	506, 540, 542
MODS	<accessCondition>

schema.org	schema:license schema:isAccessiblrForFree

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 Description 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 over time, 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.

Description based on contents viewed on November 4, 2003 (Community of Science website); title from main search screen

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)

If metadata is derived from a source external to the website, indicate this.

Title devised by archivist.

Crosswalks	
Dublin Core	Description
EAD	<processinfo>
MARC 21	588
MODS	<note>

schema.org	schema:description schema:disambiguatingDescription

SUBJECT

Definition: Primary topic(s) describing the content of an archived website or collection.

Topical subjects, geographic place names, and names of persons or organizations reflected in the content belong in the **Subject** element. Use as many terms as necessary to provide access to the primary subject content.

The use of a controlled vocabulary is strongly recommended to encourage consistency. Select a vocabulary that is used by the communities likely to benefit from the described content.

It may be appropriate to duplicate the **Creator** name in the **Subject** element, particularly for organizational sites in which the content is about the organization.

Creator: Duke University
Title: University Archives Web Archive Collection, 2010-[ongoing]
Subject: Duke University

Creator: Harvard University. Department of Anthropology
Title: Harvard University Department of Anthropology archived website
Subject: Harvard University. Department of Anthropology

Creator: 100,000 Poets for Change
Title: 100,000 Poets *[single site]*
Subject: 100,000 Poets for Change

Use the **Genre/Form** element, not **Subject**, for terms reflecting the type of content, such as "website," "blog" or "Facebook page."

Subject: Mormons
Genre/Form: Blogs

Subject: Political campaigns
Genre/Form: Campaign literature

Crosswalks	
Dublin Core	subject
EAD	<controlaccess>
MARC 21	600, 610, 611, 630, 647, 648, 650, 651, 653, 654, 656, 657, 658, 662, 69x
MODS	<subject>

schema.org	schema:about

TITLE

Definition: The name by which an archived website or collection is known.

Include at least one **Title** element.

September 11, 2001 Web Archive
Center for Working-Class Studies Website Archive
GlobalChange.gov *[a single site]*

Single sites

The title of a single site usually is transcribed from the head of the home page.

100,000 Poets

California Proposition 29 imposes additional taxes on cigarettes for cancer research
ARCAblog : promoting the study and research of art crime and cultural heritage protection
Snowden's father plans to visit son in Russia “very soon”
Aliza Sherman rants and raves

Optionally, add a phrase such as “archived website” to the transcribed title for a single site.

Welcome to the SAA 2000 Annual Conference Host Committee archived web site
Simplicissimus: The Harvard College Journal of Germanic Studies archived website

A title may instead be extracted from the metadata in the archived file if the wording is descriptive (i.e., more specific than “title” or “home page”).

Collections

Collection titles are usually devised. If the collection consists solely of sites harvested from a single institution’s domain, include the institution’s name.

Harvard University Faculty of Arts and Sciences departmental websites

If a collection is centered on a topic or event (or a place, person, etc.), devise a title that briefly articulates the collection’s scope.

Human Rights web archive
#blacklivesmatter web archive

Generally, append to the title a phrase such as “archived websites” or “web archives.”

Library of Congress web archives
Trinity University Student Organizations Website Archive
Grateful Dead web archive

If you choose not to add such a phrase, clearly state in one or more other elements that the collection consists of archived web content.

> **Title**: Wikileaks 2010 Document Release Collection
> **Description**: A collection of websites, news coverage, and commentary surrounding the Wikileaks releases …

The context within which the metadata record will be accessed can inform whether or not to add such a phrase. In certain website-specific interfaces it may be redundant, while in a library catalog it may provide a helpful distinction. Consider that your metadata may be repurposed in one or more other metadata environments in which web content would be integrated with descriptions of other resources.

If the content was harvested from a social media platform such as a blog, Facebook, Instagram, Twitter or any other platform, append wording to the title to indicate this.

> Darwin's Room Twitter feed
> Mormon blogs collection
> Occupy Harvard tumblr
> The Revenge of the Common Place Facebook Page
> Bob Brown (SenatorBobBrown) on Twitter

Variant Titles

Variant titles can enhance discovery. A common type is an acronym spelled out in whole words, or vice versa, depending on which version is transcribed as the primary title.

> **Title**: Federal Depository Library Program Web Archive
> **Title**: FDLP Web Archive

Multilingual versions are important title variants to include.

> **Title**: World Health Organization : WHO
> **Title**: Organisation mondiale de la Santé
> **Title**: Organización Mundial de la Salud

Include different spellings and ambiguity of page design that could lead to two or more feasible title choices.

> **Title**: Harvard University Department of Anthropology archived website
> **Title**: Harvard University Dept. of Anthropology archived website
> **Title**: Jacqui Lambie
> **Title**: Senator Jacqui Lambie

Crosswalks	
Dublin Core	Title
EAD	<unittitle>
MARC 21	245, 246, 730, 740
MODS	<titleInfo><title> <titleInfo type="alternative"><title>

schema.org	schema:name

URL

Definition: Internet address for an archived website or collection.

Use this element to record **URLs**, **URNs** or **URIs** that are useful to users, particularly seed and access **URLs**.[53] Include text to explain its function. Repeat the element as many times as necessary.

Internet Access: http://purl.fdlp.gov/GPO/gpo51421
Access: http://www.archive-it.org/collections/1068
Connect to this archived website online: https://webarchives.cdl.org/site/xxx/12345
Captured websites are accessible online at: https://archive-it.org/collections/3721
Web harvests, 2009-2016: http://nrs.harvard.edu/urn-3:HUL.ARCH.WAX:9361438
Web harvests, 2016 and later: https://archive-it.org/collections/5488
URL at time of capture: http://www.globalchange.gov/

When describing a single archived website, the seed **URL** may be a key piece of information to assist in discovery. It may no longer function when a site's address changes, and so may not be appropriate to include as access elements unless your system will resolve the change. An archived site must always have an access **URL** that will continue to be valid after the live site is taken down.

Crosswalks	
Dublin Core	Identifier
EAD	<dao href= >
MARC 21	856 subfield u
MODS	<url>

schema.org	schema:url

Future Research Needs

Archiving the web is a complex endeavor requiring a wide variety of skills and knowledge, and it would not be feasible to address all metadata and discovery issues in a single study. In developing the charge for the OCLC Research Library Partnership Web Metadata Working Group, we consciously scoped the work narrowly so that we could complete it in a reasonable time frame. We focused on descriptive metadata because it was the most widely shared need identified in surveys of end users and metadata practitioners.[54]

As we did our work, and as we solicited feedback on a draft of this report from the web archiving and metadata communities, many other issues surfaced for which investigation and solutions are needed.

Technical and preservation metadata: The lines are blurry between descriptive metadata and other types of metadata; crawl dates, for example are clearly both descriptive and technical. Metadata practitioners want to know more about all types of metadata that are harvested, including how they can be extracted, blended with descriptive metadata and made intelligible to end users.

Discovery systems: While not unique to the web archiving context, the complex relationships between structured metadata, free-text retrieval and other aspects of discovery systems are ripe for study. Further, much archived web content is siloed in systems separate from traditional library and archives discovery systems, and far from all of it is included in the latter contexts. End users are largely unaware that libraries and archives are doing this work.

Machine-actionable description: Until (or unless) it becomes possible to extract descriptive metadata from technical metadata in an automated way, what else might be possible to improve efficiencies in metadata creation? How much of the work can machines do for us? How can descriptive metadata elements be made machine-actionable so as to contribute to effective discovery and display? Would it be useful to map machine-readable metadata elements to library and archives descriptive concepts?

Multiple levels of description: WAM's recommendations touch on multilevel description issues, but not in any depth. These issues relate to the broader case of description of aggregates and parts and are relevant to both bibliographic and archival description. Sample questions: Would it be useful to examine the standards-based multilevel approach used in archival description through the particular lens of archived web content? How should standalone documents (such as PDFs of reports) contained within a site be described in relation to the broader metadata?

MARC record types: The MARC format is widely used for encoding archived web content, and Type of Record is a one-character code (along with another closely related byte) that declares the basic nature of the material being described. The choice of code is a determining factor for the ways that search results can be faceted. Is it a serial, a manuscript, notated music, software, archival material, or a website? Given the many perspectives from which this content can be viewed, practitioners are coding this byte using a variety of codes, but many are uncomfortable that they are making the most appropriate choice.[55]

The web archiving community is robust, creative and growing by leaps and bounds. While the literature and the conversations are already substantial, we can look forward to seeing these and other pressing issues addressed in the coming years.

ACKNOWLEDGMENTS

We are grateful to the many colleagues who contributed to WAM's work. All members of the Working Group participated throughout the project (see Appendix B), a subset of which worked intensively on this report after all preparatory research was complete.

Karen Scott Farrell, Rick Fitzgerald, Tammi Kim, Mary Samouelian, Aislinn Sotelo and Jessica Venlet collaborated with the authors on the data dictionary, synthesizing the earlier work to analyze standards, guidelines and sample records. Selecting, naming and defining the data elements was an intensive process, and they rose to the occasion.

During this process, we received insightful questions and comments from WAM members Rebecca Guenther and Debbie Kempe, as well as from OCLC Research colleague Karen Smith-Yoshimura.

During the period for open comments from the community at large, a number individuals and groups offered particularly nuanced, insightful feedback: Ed Summers (University of Maryland), Erik Moore and colleagues (University of Minnesota), Megan O'Shea and colleagues (New York University), Maureen Callahan and Adrien Hilton on behalf of the Society of American Archivists' TS-DACS subcommittee, Alex Thurman (Columbia University) and Bertram Lyons (AV Preserve).

Finally, colleagues in OCLC Research were indispensable contributors. Program Officer Dennis Massie provided wise counsel and support to the Working Group throughout the project. Erin M. Schadt, Jeanette McNicol and JD Shipengrover efficiently shepherded our three publications through the production process.

APPENDICES

APPENDIX A: WORKING GROUP CHARGE

THE PROBLEM

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

A spring 2015 survey of members of the OCLC Research Library Partnership revealed the lack of descriptive metadata guidelines as the biggest challenge related to website archiving among this cohort. The second most-cited challenge is understanding the needs of users who seek to use website content in their work.

Review of existing guidelines, as well as sampling of descriptions in WorldCat and ArchiveGrid, reveals widely variable practice. This can be traced, at least in part, to the fact that some characteristics of websites are not addressed by existing descriptive rules such as RDA (Resource Description and Access) and DACS (Describing Archives: A Content Standard). Some record creators follow bibliographic traditions, while others use an archival approach, such as describing multiple sites in one record. Sometimes the two approaches are blended.

ADDRESSING THE PROBLEM

A call for volunteers to join a working group to address these challenges led to an enthusiastic response, and work began in January 2016.

The Working Group will study archival and bibliographic description practices for archived websites, consider when each approach might be most appropriately used, and determine how the two might be made compatible. We will keep in mind that metadata is sometimes repurposed for reuse in a variety of different tools and contexts. We will also consider issues related to description of archived websites in relation to live/active sites.

In the coming months, members of the Working Group will:

1. Finalize the issues to be addressed.
2. Perform desk research to learn about user needs and behavior relative to websites to inform our approach to defining best practices for descriptive metadata.
3. Develop best practices for metadata, informed by the study of existing guidelines for describing archived websites—such as those developed by the Program on Cooperative Cataloging, the New York Art Resources Consortium and a variety of individual institutions.
4. Study the published literature and online sources to identify metadata issues identified by researchers in the field.
5. Informally sample and evaluate existing descriptions of archived websites in WorldCat (MARC records), ArchiveGrid (MARC records and finding aids), Archive-It and other sources.

6. Investigate available tools for web archiving and the ways in which they enable production of descriptive metadata.

The full working group will meet monthly via WebEx. Subgroups will undertake specific tasks and report their findings to the group. We have tentatively targeted January 2017 to complete the work.

We will liaise with other groups that are active and influential in the web archiving sphere of practice. These include the Web Archiving Roundtable of the Society of American Archivists, the International Internet Preservation Consortium (IIPC) and the Internet Archive.

PROJECTED OUTCOMES

We will develop best practices that bridge existing approaches to description of archived websites. These will enable practitioners to create appropriate metadata consistently and with confidence that they are following community practice. Best practices will also benefit users by increasing discoverability through consistency of metadata and by providing contextual information that meets identified needs.

A report on user needs will inform community-wide understanding of documented user needs and behaviors as evidence to underlie the best practices for metadata for archived websites.

APPENDIX B: MEMBERS OF THE WEB ARCHIVING METADATA WORKING GROUP

Alexis Antracoli, Princeton University

Penny Baker, Clark Art Museum

Kate Bowers, Harvard University

Lori Dedeyan, University of California at Berkeley

Karen Stoll Farrell, Indiana University

Rick Fitzgerald, Library of Congress

Ben Goldman, Pennsylvania State University

Rebecca Guenther, metadata standards consultant

Claudia Horning, UCLA

Deborah Kempe, Frick Museum

Tammi Kim, University of Nevada Las Vegas

Jason Kovari, Cornell University

Matthew McKinley, California Digital Library

Rosalie Lack, UCLA

Allison O'Dell, University of Florida

Joy Panigabutra-Roberts, University of Tennessee at Knoxville

Dallas Pillen, University of Michigan

Lily Pregill, Frick Museum

Mary Samouelian, Harvard Business School

Aislinn Sotello, University of California at San Diego

Jessica Venlet, Massachusetts Institute of Technolocy

Olga Virakhovskaya, University of Michigan

Jackie Dooley, OCLC Research

Dennis Massie, OCLC Research

APPENDIX C: DESCRIPTION ELEMENT EXAMPLES

These examples are taken verbatim from extant metadata records. They were chosen for their clear statements of selection criteria, provenance, scope, content and other characteristics.

COLLECTIONS OF ARCHIVED WEB SITES

A growing collection of websites selected by the Avery Architectural and Fine Arts Library staff for web archiving preservation by the Columbia University Libraries' Web Resources Collection Program. Website captures began in 2010 and are ongoing. The collection's principal thematic focus is documenting the evolution of the built environment and public spaces through the interaction of historic preservation efforts and new development projects within urban planning debates. Selected websites are mostly published by nonprofit groups or individuals based in the New York City area, including historic preservation groups, neighborhood associations, public policy organizations, parks conservancies …

Archive of Web sites of individuals, groups, the press, and institutions in the United States and from around the world in the aftermath of the attacks in the United States on September 11, 2001. The archive consists of over 30,000 selected Web sites archived from September 11, 2001, through December 1, 2001, and is intended to preserve the Internet reaction from U.S. and non-U.S. government sites; press, corporate/business, portal, charity/civic, advocacy/interest, religious, school/educational, individual/volunteer, and professional organization sites. Browse access to descriptions of approximately 2,300 Web sites is available, as well as a list of all sites archived.

Collection of websites published by David Crumm Media LLC, a multimedia publishing company located in Canton, Mich., that focuses on religion and spirituality. Headed by partners David Crumm and John Hile. Read the Spirit Editor in Chief is Wayne E., Baker, University of Michigan Ross School of Business faculty. Web site includes Wayne Baker's blog.

In October of 2007, a series of wildfires broke out throughout southern California. This web archive documents that event as it appeared on California State Agency sites, federal government sites, news, blogs, social networking sites. In addition to providing lasting access to the web's coverage of the fires, this archive may serve as a point of comparison to other web archives of dramatic and quickly unfolding historical events. In particular it may show how people communicated about the fire, how they got critical information such as evacuations and how they documented the impact of the fire as it unfolded.

Legal blawgs is a selective collection of authoritative sites (associated with the American Bar Association-approved law schools, research institutes, think tanks, and other expertise-based organizations) that contain unique, born-digital content. These blogs contain journal-style entries, articles and essays, discussions, and comments on emerging legal issues, national and international. Sites are domestic (US) and in English, although foreign sites may be included later in the duration of this project.

On July 1st, 2005, Associate Justice Sandra Day O'Connor announced her resignation from the United States Supreme Court … nomination of John G. Roberts, Jr. heightened the interest in the nominations process, the constitutional advice and consent role of the United States Senate, and the importance of the Judicial Branch. Although over 150 names have been sent to the Senate by the president since 1789, the nomination of John G. Roberts, Jr. is the first during an era of widespread use of the Internet as a communication medium. To preserve the nomination context and discussion for future scholars, the Library of Congress identified and archived numerous Web sites relating to the Supreme Court nomination and appointment process.

The 2014 US-Cuba Policy Change Twitter archive collection contains a data set of tweets collected from the Twitter microblogging platform when President Barack Obama announced on December 17, 2014, that the United States would begin normalizing full diplomatic relations between the US and Cuba after more than half a century of minimal relations. President Obama's announcement included plans to re-establish the US embassy in Havana, allowing official visits of Cuban diplomats and officials to the United States, and increased official dialogue on public policy issues affecting both countries. During the announcement, the Cuban Heritage Collection collected tweets relating to the hashtags #cuba, #cubapolicy, #cubalibre, #cubausa, #uscuba, and #cubanmiami between December 9, 2014, and January 28, 2015 … This data archive is available for download to the University of Miami community via the University of Miami scholarly repository. The data is presented in JSON structured text files. For information on accessing the archive, see the Access and Restrictions section of this finding aid.

The April 16 Web Archive captures a wide variety of content related to the April 16, 2007, tragedy at Virginia Tech. It includes memorial and tribute sites, commercial and noncommercial media, and other relevant web-based materials. This collection was specifically developed as a complement to the April 16 Archive (www.april16archive.org), which is dedicated to collecting and preserving individual stories, images, and files related to the events of April 16.

> “The Library and Archives of Canada Act received Royal Assent on April 22, 2004. For the purposes of preservation it allows Library and Archives Canada (LAC) to collect a representative sample of Canadian websites. To meet its new mandate, LAC began to harvest the web domain of the Federal Government of Canada starting in December 2005. As resources permit, this harvesting activity will be undertaken on a semi-annual basis. The website data which is harvested is stored in the Government of Canada Web Archive (GC WA). Client access to the content of the GC WA is provided through searching by keyword, by department name, and by URL. It is also possible to search by specific format type, e.g. .pdf. At the time of its launch in Fall 2007, approximately 100 million digital objects (over 4 terabytes) of archived Federal Government website data was made accessible via the LAC website”. --Introduction, home page.

The Manuscript Division Archive of Organizational Web Sites includes principally Web sites of organizations with whom the division has an existing relationship … they fall into several broad categories. First, there are sites for non-governmental, voluntary organizations, including civil rights and political advocacy groups whose records the division holds. Examples include the National Urban League, Leadership Conference on Civil Rights … professional and honorary organizations such as the American Historical Association … [and much more about organizations whose websites have been archived.]

The UC Davis Web Archives preserves websites in the “ucdavis.edu” domain to document the history of the University's activities and accomplishments. Special Collections, as the repository for the University Archives, collects records of historical value for the University. Previously, many of these records have been in print form, now much of that information can be found on campus websites. This project, started in 2011, captures the websites of the University's administration, schools and colleges, academic departments, administrative units, organized research units, intercollegiate athletics, and student organizations …

This archive preserves and makes permanently accessible exclusively digital government information presented in official US Federal agency Web sites. Iterations of sites are captured periodically to reflect Federal government sites' evolution, and to enable users to access publications and information that are no longer accessible from the sites' current versions.

This archive provides ongoing access to the websites of Olympic committees for nations that competed in the 2010 Vancouver Olympic Games. Snapshots were taken of Olympic committee websites before and

throughout the games, with more frequent captures of nations with many athletes competing. Note that we were not able to archive sites for every participating nation; in several cases Olympic committee sites were unavailable for capture. The CTV and NBC websites devoted to the games are also included here. This archive was created as part of an effort by the International Internet Preservation Consortium (IIPC) to begin experimenting with cross-archival search strategies, using the Olympics as a common theme. Along with several national libraries, the California Digital Library is providing content related to the Olympics to serve as a basis for international cross-archival searching.

This web archive preserves internet sites that contain information, resources, or online materials documenting the Grateful Dead phenomenon, a term that includes both band and fans. Many of these websites are devoted to the band's music and history; some document other aspects of the remarkable relationship the Dead created with their fans. In addition to these resources, the sites contain stories, opinions, and the digital objects embedded in these websites, such as podcasts, tweet feeds, and digital images.

Web archive of the site for 100,000 Poets for Change, an international educational grassroots organization focusing on the arts, especially poetry, music, and the literary arts. 100TPC events take place simultaneously around the world in September of each year. The first event took place on September 24th, 2011, in a demonstration/celebration of poetry to promote social and political change and over 700 events in 550 cities representing 95 countries signed up to make this global initiative a success through poetry readings, political demonstrations, community picnics, awareness events, parades, and more.

SINGLE ARCHIVED WEBSITES

A blog focusing on art thefts, cultural heritage protection and restitution all over the world, centered around the activities and conferences of ARCA.

Online archive of the print issues of The Record (2004-2009). The web site is no longer live.

Twitter page for the leader of the Australian Greens party Bob Brown. Contains short posts from Senator Brown about his activities and policies.

Website of a grassroots advocacy, low-income rights and social justice organization based in Springfield, Massachusetts.

Established with a vision to “empower the nation with global change science,” GlobalChange.gov comprises thirteen Federal agencies and is a web-based portal serving structured global change data and information with an emphasis on the National Climate Assessment. This official website presents its mission and strategic plan and provides access to relevant resources generated or sponsored by the US Government and other authoritative scientific bodies.

Latvia National Participation at the Venice Biennale 2015. Instagram. Exhibition title: Armpit. Artists: Katrina Neiburga, Andris Eglitis. Commissioner: Solvita Krese (Latvian Centre for Contemporary Art). Deputy Commissioner: Kitija Vasiljeva. Curator: Kaspars Vanags. Venue: Pavilion at Arsenale.

This is an archived web page collected at the request of Sterling and Francine Clark Art Institute Library using Archive-It. This page was captured on 15:37:34 Sep 16, 2015, and is part of the Venice Biennale 2015 on the Web collection https://archive-it.org/collections/5748.

The site contains archived versions of Ford.com that were captured beginning in December 2007. Plans are for crawls to be ongoing with captures on a quarterly basis. The content reflects the material on the site as of the given dates. Offers, policies, pricing and other content may have changed since that date and may no longer be valid.

This site is a collaborative effort by federal agencies formed as a group in 2007 to define common guidelines, methods, and practices to digitize historical content in a sustainable manner. Recognizing that the effort would require specialized expertise, two separate working groups were formed with the possibility that more tightly focused groups might be necessary as the work progressed. The Federal Agencies Still Image Digitization Working Group will concentrate its efforts on image content such as books, manuscripts, maps, and photographic prints and negatives. The Federal Agencies Audio-Visual Working Group is focusing its work on sound, video, and motion picture film.

APPENDIX D: ENCODED EXAMPLES

The following are adapted from extant examples of descriptions for archived web content to reflect the recommendations of WAM's best practice report.

SINGLE ARCHIVED WEBSITE ENCODED IN MARC

WAM element	MARC tag	Content
Creator	110	Harvard Ventures
Title	245	Harvard College Ventures archived website
Language	008 546	eng In English.
Date	008 245 $$f	2014 9999 2014 and later accruals
Extent	300	1 archived website
Access conditions	506	Unrestricted online access.
Source of description	500	Name of organization "Harvard Ventures" based on web harvest of July 2014. Title devised by archivist based on June 2016 web harvest.
Description	520 545	Periodic captures of the Harvard College Venture Partners Website, harvardventures.org, beginning in July 2014. The Harvard Ventures website provides information on leadership, members, committees, events, sponsors, and mailing lists. Harvard Ventures is an undergraduate entrepreneurship club founded in 2011 at Harvard University.
Subject	650 610	Entrepreneurship -- Societies and clubs -- Massachusetts -- Cambridge Harvard Ventures
Genre/Form :	655 655	Web sites Web archives
Relation	773 830 830	Forms part of the Records of the Harvard Ventures. Collections of the Harvard University Archives. Web archives. Collections of the Harvard University Archives. Records of associated organizations.
URL	856	http://nrs.harvard.edu/urn-3:HUL.ARCH.WAX:13957645
Collector	852	HUA *[transformed to "Harvard University" by online catalog interface]*

ARCHIVED WEB COLLECTION ENCODED IN MARC

WAM element	MARC tag	Content
Title	245 246	Iraq War, 2003 web archive Iraq War web archive
Date	008 245 $$f	20039999 2003-
Language	008 546	eng Archived websites are in English.
Extent	300	1 web archive collection (231 websites)
Description	500	Title from overview page, as viewed on December 23, 2008.
Relation	500	Part of the Library of Congress Web Archives, Minerva
Access conditions	506	Descriptions of the archived Web sites are searchable without restriction; access to some archived sites is restricted to on-site users of the Library of Congress.
Description	520	Selective collection of 231 Web sites, archived beginning on March 13, 2003 related to the Iraq War. Included in the archive are websites from the U.S. and foreign governments, public policy and advocacy groups, educational organizations, religious organizations, support groups for military personnel, anti-war groups, sites that target children, and news sources. The Iraq War Web archive consists of three phases of collection: the first phase, a weekly capture, began on March 13, 2003 with the commencement of the war and ended June 30, 2003. Phase 1 has been processed and is available from this site. Phase 2 is a weekly capture and covers December 2003 to December 2004. Phase 3, also a weekly capture, was begun in January 2005 and is ongoing (archives from these later phases are not yet available).
Form/genre	655	Web archives.
Subject	650 651	Iraq War, 2003-2011 United States -- Military policy
Collector	710 852	Library of Congress, |e collector lcwa *(transformed to human-readable text by online interface)*
URL	856	http://hdl.loc.gov/loc.natlib/collnatlib.00000003

SINGLE ARCHIVED WEBSITE ENCODED IN DUBLIN CORE

WAM element	Dublin Core element	Content
Title	Title	Women's Worlds in Qajar Iran archived website Women's Worlds in Qajar Iran دنیای زنان در عصر قاجار
Date	Date	2015 and later accruals
Description	Description	Web archive of http://www.qajarwomen.org, a discovery and research platform on the topic of women during the Qajar era (1796-1925) in Iran. Website provides bilingual access to thousands of personal papers, manuscripts, photographs, and other materials in both private and institutional collections. The Harvard University Library (HUL) central infrastructure accommodates all image, text, and audio materials on this website. Web archive of this site collected by the Harvard University Archives as part of an ongoing effort to document the University's role in support of scholarship.
Extent	Format	1 archived website
Genre/form	Format	Web sites Web archives
Access conditions	Rights	Live website is publicly available. The archived version is embargoed for duration of the live site's accessibility.
Language	Language	In English and Farsi (Persian).
Relation	Relation	Collections of the Harvard University Archives. Web archives.
Source of description	Description	Title devised by archivist based on of the English title on the 2015 website harvest.
Subject	Subject	Iran -- History -- Qajar dynasty, 1794-1925 Women -- Iran -- History -- Sources
Contributor	Contributor	Harvard University Library, host institution
Collector	Contributor	Harvard University Archives, collector
URL	Identifier	[Access URL to archived website embargoed]

Multilevel Example Encoded in EAD

Language, creator, history note and selected subject terms are applicable to the whole collection and expressed within the collection description. Many other elements of the collection description have been left out to make the example shorter.

```
<?xml version="1.0" encoding="UTF-8"?>
        <ead xmlns="urn:isbn:1-931666-22-9" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                xsi:schemaLocation="urn:isbn:1-931666-22-9 http://www.loc.gov/ead/ead.xsd">
        <eadheader countryencoding="iso3166-1" dateencoding="iso8601" langencoding="iso639-2b"
                repositoryencoding="iso15511" scriptencoding="iso15924">
        <eadid countrycode="US" mainagencycode="MH-Ar">HUANNNNN</eadid>
        <filedesc>
        <titlestmt>
        <titleproper>Records of Harvard Ventures: an inventory</titleproper>
        </titlestmt>
        </filedesc>
        <profiledesc>
        <langusage><language langcode="eng">English</language></langusage>
        </profiledesc>
        </eadheader>
        <archdesc level="collection">
        <did>
        <repository>Harvard University Archives</repository>
        <origination><corpname>Harvard Ventures</corpname></origination>
        <unittitle>Records of Harvard Ventures</unittitle>
        <unitdate calendar="gregorian" era="ce" normal="2011/2016">2011-2016 </unitdate>
        <langmaterial><language langcode="eng">English</language></langmaterial>
        </did>
        <bioghist>
        <p>Harvard Ventures is an undergraduate entrepreneurship club founded in 2011 at Harvard University.</p>
        </bioghist>
        <dsc>
                        <c level="series">
                        <did>
                                <unittitle>Harvard College Ventures archived website</unittitle>
                                <unitdate normal="2014/2016" calendar="gregorian"era="ce">2014-2016 and later
                                accruals</unitdate>
                                <physdesc><extent>1 archived website</extent></physdesc>
                                <dao xmlns:xlink="http://www.w3.org/1999/xlink" xlink:href="http://nrs.harvard.edu/urn-
                                3:HUL.ARCH.WAX:13957645" xlink:actuate="onRequest" xlink:show="new"/>
                        </did>
                        <accessrestrict>
                                <p>Unrestricted online access.</p>
                        </accessrestrict>
                        <scopecontent>
                                  <p>Periodic captures of the Harvard College Venture Partners website beginning in
                                  July 2014.</p>
                        </scopecontent>
                        <processinfo>
                                 <p>Name of organization based on “Harvard Ventures” based on web harvest of July
                                 2014. Title devised by archivist based on June 2016 web harvest.</p>
```

```
                    </processinfo>
                    <controlaccess>
                    <genreform>Web sites.</genreform>
                    </controlaccess>
                    <controlaccess>
                    <genreform>Web archives.</genreform>
                    </controlaccess>
                    </c>
</dsc>
<controlaccess>
<corpname role="subject">Harvard Ventures.</corpname>
</controlaccess>
<controlaccess>
<subject>Entrepreneurship -- Societies and clubs -- Massachusetts -- Cambridge.</subject>
</controlaccess>
</archdesc>
</ead>
```

NOTES

1. Lavoie, Brian, Eric Childress, Ricky Erway, Ixchel Faniel, Constance Malpas, Jennifer Schaffner, and Titia van der Werf. 2014. *The Evolving Scholarly Record*. Dublin, OH: OCLC Research. http://www.oclc.org/research/publications/library/2014/oclcresearch-evolvingscholarly-record-2014.pdf.
2. The ideas formulated in the 2014 report are elaborated in: Lavoie, Brian, and Constance Malpas. 2015. *Stewardship of the Evolving Scholarly Record: From the Invisible Hand to Conscious Coordination*. Dublin, OH: OCLC Research. http://www.oclc.org/content/dam/research/publications/2015/oclcresearch-esrstewardship-2015-a4.pdf.
3. Ricky Erway. 2015. "Thoughts from Partner Staff about Web Archiving" *hangingtogether.org* (blog). Posted 29 October 2015. http://hangingtogether.org/?p=5450.
4. A research team led by Matthew Weber at Rutgers University surveyed users of web archives in the winter of 2016. They expect to publish their data soon.
5. OCLC Research Library Partnership "Web Archiving Metadata Working Group." Last updated 10 March 2016. http://oc.lc/wam.
6. "Web Archiving." 2017. Word of the Week, Society of American Archivists. http://us3.campaign-archive2.com/?u=56c4cfbec1ee5b2a284e7e9d6&id=40edf162c4.
7. See note 6.
8. http://netpreserve.org/.
9. https://www2.archivists.org/groups/web-archiving-section.
10. https://archive-it.org/.
11. Dooley, Jackie M., Karen Stoll Farrell, Tammi Kim, and Jessica Venlet. 2017. "Developing Web Archiving Metadata Best Practices to Meet User Needs." *Journal of Western Archives* 8:2. http://digitalcommons.usu.edu/westernarchives/vol8/iss2/5/.
12. A sample blog post: Dooley, Jackie. 2017. "Best Practices for Web Archiving Metadata: Watch this Space!" *hangingtogether* (blog). Posted 5 April 2017. http://hangingtogether.org/?p=5918.
13. Samouelian, Mary, and Jackie Dooley. 2018. *Descriptive Metadata for Web Archiving: Review of Harvesting Tools.* Dublin, OH: OCLC Research. doi:10.25333/C37H0T.
14. Venlet, Jessica, Karen Stoll Farrell, Tammy Kim, Allison Jai O'Dell, and Jackie Dooley. 2018. *Descriptive Metadata for Web Archiving: Literature Review of User Needs.* Dublin, OH: OCLC Research. doi:10.25333/C33P7Z.
15. http://www.loc.gov/standards/mods/.
16. See appendix E for a list of the standards cited in this report.

17. The 2005 Society of American Archivists, *A Glossary of Archival and Records Terminology* by Richard Pearce Moses includes two definitions of "context:" "n. ~ 1. The organizational, functional, and operational circumstances surrounding materials' creation, receipt, storage, or use, and its relationship to other materials. - 2. The circumstances that a user may bring to a document that influences that user's understanding of the document," 90. http://files.archivists.org/pubs/free/SAA-Glossary-2005.pdf.

18. Dooley, Jackie. 2015. *The Archival Advantage: Integrating Archival Expertise into Management of Born-digital Library Materials*. Dublin, OH: OCLC Research, 16-17. http://www.oclc.org/content/dam/research/publications/2015/oclcresearch-archivaladvantage-2015.pdf.

19. *Code of Best Practices in Fair Use for Academic and Research Libraries*. 2012. Washington, DC: Association of Research Libraries, 276. http://www.arl.org/storage/documents/publications/code-of-best-practices-fair-use.pdf.

20. A seed is the starting point URL for a crawler and an important access point for an archived site or collection. See Praetzellis, Maria. 2017. "Glossary of Archive-It and Web Archiving Terms." Updated June 2017. https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms.

21. New York Art Resources Consortium. 2015. *Metadata Application Profile and Data Dictionary for Description of Websites with Archived Versions, Version 1, (June 2015).* http://www.nyarc.org/sites/default/files/web-archiving-profile.pdf.

22. DACS Principle 1 states: "… institutions may also acquire and assemble records that do not share a common provenance or origin but that reflect some common characteristic, such as a particular subject, theme, or form." "Describing Archives: A Content Standard (DACS), Second Edition." 2013. Society of American Archivists. http://www2.archivists.org/groups/technical-subcommittee-on-describing-archives-a-content-standard-dacs/dacs#.V45ZcCMrJmA.

23. International standards are cited for date and language in the descriptions of these data elements.

24. See note 14.

25. See note 13.

26. Most of the institutional guidelines are not publicly available, so the analysis without is presented without attributions.

27. *Describing Archives: A Content Standard (DACS)*. 2013. 2nd Edition. Chicago, IL: Society of American Archivists. Last modified 22 April 2016. http://www2.archivists.org/groups/technical-subcommittee-on-describing-archives-a-content-standard-dacs/dacs#.V45ZcCMrJmA.

28. "Integrating resources are continuations in which new material is incorporated with older material, such as loose-leaf services (common. for law material), and websites." See "Integrating Resources: A Cataloging Manual." 2015 (Draft revision). Module 35. In *CONSER Cataloging Manual.* Washington, D.C.: Program for Cooperative Cataloging. www.loc.gov/aba/pcc/conser/word/Module35.doc. Also available as "Appendix A: Integrating Resources Manual" In *BIBCO Participants' Manual. 2014. 3rd Edition.* Washington, D.C.: Program for Cooperative Cataloging. http://www.loc.gov/aba/pcc/bibco/documents/bpm.pdf.

29. Dublin Core Metadata Initiative. “User Guide.” Last modified 6 September 2011. https://github.com/dcmi/repository/blob/master/mediawiki_wiki/User_Guide.md.

30. Praetzellis, Maria. 2017. “Add, Edit, and Manage Your Metadata.” Posted May 2017. https://support.archive-it.org/hc/en-us/articles/208332603-Add-edit-and-manage-your-metadata.

31. https://www.loc.gov/ead/.

32. https://www.loc.gov/marc/.

33. http://www.loc.gov/standards/mods/.

34. http://www.worldcat.org.

35. https://beta.worldcat.org/archivegrid/.

36. https://archive-it.org/.

37. MARC 21 format for Bibliographic Data. Leader 06. Washington, DC: Library of Congress. https://www.loc.gov/marc/bibliographic/bdleader.html.

38. We did not further explore issues related to MARC record types, which are relevant to a single standard. See the section on Future Research Needs.

39. http://schema.org/.

40. The content of some examples was slightly edited to conform to these guidelines.

41. The only element not relevant for description of live sites is Collector.

42. “A digital repository is a mechanism for managing and storing digital content.” In “What is a Repository?” Repositories Support Project.” Accessed 30 August 2017. http://www.rsp.ac.uk/start/before-you-start/what-is-a-repository/.

43. No publisher element is included in this data dictionary because the object of description is an archived website or collection, not a live website.

44. “The Web Archiving Service (WAS) collections and all core infrastructure activities, i.e., crawling, indexing, search, display, and storage, have been transferred to Internet Archive's Archive-It.” n.d. California Digital Library. http://webarchives.cdlib.org/. See also “Announcing a New Partnership: California Digital Library, UC Libraries, and Internet Archive’s Archive-It Service.” 2015. *CDLINFO News*. Posted 14 January 2015. http://www.cdlib.org/cdlinfo/2015/01/14/announcing-a-new-partnership-california-digital-library-uc-libraries-and-internet-archives-archive-it-service/.

45. In the MARC21 format, role terms are called Relator Terms. See “Term Sequence” MARC Code List for Relators. Updated 21 October 2014. https://www.loc.gov/marc/relators/relaterm.html.

46. “Date and Time format - ISO 8601”. International Organization for Standardization (IS). https://www.iso.org/iso-8601-date-and-time-format.html.

47. Wide variation exists across community practices and style guides as to whether this is spelled as one word or two.

48. “Library of Congress Genre/Form Terms.” Library of Congress. http://id.loc.gov/authorities/genreForms.html.

49. “Art & Architecture Thesaurus® Online.” The Getty Research Institute. http://www.getty.edu/research/tools/vocabularies/aat/.

50. Language codes - ISO 639. “Using a code rather than the name of a language has many benefits as some languages are referred to by different groups in different ways, and two unrelated languages may share the same or similar name.” International Organization for Standardization. Accessed 30 August 2017. https://www.iso.org/iso-639-language-codes.html.

51. “It is fair use to create topically based collections of websites and other material from the Internet and to make them available for scholarly use. ... To the extent reasonably possible, the legal proprietors of the sites in question should be identified according to the prevailing conventions of attribution.” Code of Best Practices in Fair Use for Academic and Research Libraries. 2012. Washington, DC: Association of Research Libraries, 27. http://www.arl.org/storage/documents/publications/code-of-best-practices-fair-use.pdf.

52. “These Rights Statements are necessarily limited to copyright and related rights, which may vary by national law. For example, assertions about the Public Domain status of a Work may be defined differently depending on the jurisdiction in which that determination is made.” International Rights Statements Working Group. 2016. *Recommendations for Standardized International Rights Statements*, *October 2015 (updated November 2017)*, 15. http://rightsstatements.org/files/171116recommendations_for_standardized_international_rights_statements_v1.2.pdf.

53. A digital object identifier (DOI) is not necessarily a URI, unless it is managed as one via representation in a resolver service. In such cases, a DOI is a URL as defined in this data dictionary.

54. See note 3. Erway, *Web Archiving,* and note 4. Weber, “Web Archives Users Survey.”

55. An OCLC Research query to WorldCat in April 2017 showed that roughly equal numbers of records that include the genre/form term “web site” or “website” are coded as texts (type of record a), integrating resources (bibliographic level I) and or mixed materials (type of record p).

For more information about our work related to digitizing library collections, please visit: **oc.lc/digitizing**

6565 Kilgour Place
Dublin, Ohio 43017-3395

T: 1-800-848-5878
T: +1-614-764-6000
F: +1-614-764-6096
www.oclc.org/research

ISBN: 978-1-55653-016-6
DOI:10.25333/C3005C
RM-PR-215938-WWAE 1709