Apr 29th, 1:00 PM - 4:00 PM

Metadata for Digital Projects: An Overview of Practical Issues and Challenges

Murtha Baca PhD
Getty Research Institute (emerita) and UCLA (emerita)

Follow this and additional works at: https://digital.sandiego.edu/symposium

https://digital.sandiego.edu/symposium/2019/2019/7

This Workshop is brought to you for free and open access by Digital USD. It has been accepted for inclusion in Digital Initiatives Symposium by an authorized administrator of Digital USD. For more information, please contact digital@sandiego.edu.
Metadata for Digital Projects: An Overview of Practical Issues and Challenges

Presenter 1 Title
Head of Digital Art History, Getty Research Institute (emerita) ; Adjunct Professor, UCLA Department of Information Studies (emerita)

Session Type
Workshop

Abstract
This three-hour workshop will provide an overview of practical issues relating to metadata and controlled vocabularies for digital resources. There will be a review of metadata standards and vocabulary tools; project management and project planning considerations; and issues relating to publication formats, usability, and sustainability. Issues of metadata for the “Visible Web” and the “Deep Web” will also be addressed.

Workshop participants will do an in-classroom exercise in which they will create a "storyboard" for a proposed digital project, including a high-level metadata model and proposed vocabularies to be used, as well as their strategies for how users will find and interact with their proposed digital resources.

Location
KIPJ Room A

Comments
Murtha Baca has three decades of experience as an implementer and teacher of descriptive metadata and controlled vocabularies. She led the Getty Vocabulary Program, which developed and maintains multilingual controlled vocabularies for art, architecture, and material culture that are used all over the world. She developed and taught, for 18 years, a graduate seminar on metadata for the UCLA Department of Information Studies. She was a co-editor of Cataloging Cultural Objects, and the editor of Introduction to Metadata, in addition to writing and editing numerous articles and keynote speeches. As Head of Digital Art History at the Getty Research Institute (GRI) in Los Angeles, Murtha led a team of scholars and technical experts who developed the GRI's first "born-digital" scholarly publication. She has taught numerous workshops and seminars nationally and internationally, and written extensively on descriptive metadata, controlled vocabularies, and digital project management. She has twice received the De Laurier award from the Visual Resources Association (VRA); in 2017 she received the Distinguished Teaching award from the UCLA Department of Information Studies.

This workshop is available at Digital USD: https://digital.sandiego.edu/symposium/2019/2019/7
METADATA* FOR DIGITAL PROJECTS: THINKING ABOUT THE NUTS & BOLTS

Workshop presented by Martha Baca, PhD
University of San Diego Digital Initiatives Symposium, April 2019

“Making a Website” ≠ Doing a Digital Project

Images and other digital assets without accompanying metadata are mostly useless, and generally “unfindable,” unsharable, and not reusable.

Some digital projects in which I have participated

- Getty vocabularies: http://www.getty.edu/research/tools/vocabularies/ (+ Vocab LOD project)
- “Digital Mellini”—the Getty Research Institute’s first digital critical facsimile publication (with metadata record in the GRI’s OPAC, WorldCat, etc.): http://www.getty.edu/research/mellini/
- Getty Research Portal: http://portal.getty.edu/
The Role of Language

Weeping Woman
Crying Woman
Femme qui pleure
La larmoyante
La Mujer que llora
La Mujer llorando
Donna che piange
Donna piangente

Controlled vocabularies reflect the critical & linguistic history of an person, object, concept, etc., and provide important additional access points

Bulgarini, Bartolomeo
Bartolomeo Bulgarini
Bartolomeo Bulghini
Bartolomeo Bulgarini
Bartolomeo Bulgarini da Siena
Maestro d'Ovile
Master of the Ovile Madonna
Ovile Master
Lorenzetti, Ugolino
Ugolino Lorenzetti

Ayia Sophia
Ayasofya
Church of the Holy Wisdom
Hagia Sophia
Haghia Sophia
Saint Sophia
Santa Sophia
St. Sophia

Familie rose
British Museum, London

Ayia Sophia
Ayasofya
Church of the Holy Wisdom
Hagia Sophia
Haghia Sophia
Saint Sophia
Santa Sophia
St. Sophia

Familie rose
British Museum, London

Example from Getty Research Institute
Cultural Objects Name Authority (CONA)

Ch'ien-lung Reign, Ch'ing dynasty (1644-1911)

Fencai National Palace Museum, Taipei

Famille rose British Museum, London
Using terminology to reach broader audiences: this is where “local” “collection-specific” or “resource-specific” controlled vocabularies, which can include non-expert and even “wrong” terms (as in “folksonomies”), can help.

cabinet? desk?
cartonnier?

cabinet? desk?
cartonnier?
The publication also has an ISBN number, like all of the Getty’s print publications.
GETTY RESEARCH PORTAL: AGGREGATED METADATA RECORDS FROM LIBRARIES AROUND THE WORLD

Search Results Display

GETTY RESEARCH PORTAL SEARCH RESULTS DISPLAY

Record Detail

COMPLETE BROWSABLE, DOWNLOADABLE DIGITAL FACSIMILE IS AVAILABLE AT THE INTERNET ARCHIVE

M. Baca, Metadata for Digital Projects
Information technology makes it possible to frame research questions in a computational way, to use electronic tools and new research methods to work (and collaborate!) more efficiently, and to ask new kinds of questions. It also facilitates sharing of both raw data and research findings—*if data and metadata are carefully and thoughtfully formatted.*
The “Visible Web” versus the “Deep Web”

- The Visible Web is what you see in the results pages from commercial search engines like Google.
- The Invisible or Deep Web consists of data from dynamically searchable databases that are not automatically indexed by search engines, because they are not static HTML pages that “live” somewhere—they are created on the fly when a user does a search.

METADATA FOR THE WEB

- The Web is not a “library”!
- Web searching is very hit-and-miss
- Some “places” for Web metadata exist, but not all institutions implement them consistently:
  - TITLE HTML tag
  - DESCRIPTION META tag
  - KEYWORDS META tag
  - “No index, no follow” META tag

METADATA FOR THE WEB CONTINUED

The most important elements for search engine optimization (SEO) are:

- The HTML “TITLE” TAG (appears at the top of a web page, and is used to bookmark the page)
- The actual indexable text on the page
- Referring links (the Google “popularity contest”)

Speaking of the Web...

- Will your digital resource be “reachable” by commercial search engines?
- If yes, how will you “contextualize” individual objects?
- If not, what is your strategy to lead Web users to your main page/search page?
Order from Chaos: The Pieces of the Puzzle

- Data (aka "metadata")
- Assets (e.g., images, media files, texts, bibliography, etc.)
- People (with clearly defined roles)
- Skill sets (e.g. cataloging, TEI markup, software administration, database management, copy editing, Web writing/editing, interface/UX design)
- Standards

The Pieces of the Puzzle, continued

- Appropriate software AND software support
- Institutional support
- A project manager
- Physical & virtual space to work, and an institutional "venue" to publish research and supporting data, and to maintain (or, eventually, "retire") resources

WHAT IS METADATA?

- "Metadata" is often used interchangeably (and confusingly) with "data."
- "Metadata" is often used to refer to meta tags on HTML pages on the Web.
- "Metadata" (like "data") is a plural word, but usually used as if it were singular.

WHAT IS METADATA?

A structured description of the essential attributes of an information object. (Tony Gill, Chapter 2, Introduction to Metadata 3.0)

Metadata is normally structured to model the most important attributes of the class of information objects being described (e.g., the MARC format).
WHAT IS METADATA?

Metadata is structured information associated with an object for purposes of discovery, description, use, management, and preservation.

from the NISO Framework of Guidance for Building Good Digital Collections, 3.0.

TYPES OF METADATA

- **Administrative**: for managing and administering information resources (e.g. location information, version control)
- **Descriptive**: for the description or identification of information resources (e.g. specialized indexes, finding aids, individual object records)

TYPES OF METADATA (CONT.)

- **Preservation**: for the preservation management of information resources (e.g. documentation of data “refreshing” and migration)
- **Technical**: related to how a system functions or how metadata behaves (e.g. hardware and software documentation, tracking of system response times)
- **Use**: (e.g. use and user tracking, usability studies)

WHY IS METADATA IMPORTANT?

- for enhanced accessibility, in multiple “places”
- for retention of context
- for expanding use & sharing
- for multi-versioning
- for legal issues
- for preservation of data
"RE-CONTEXTUALIZED" METADATA: EXAMPLES OF METADATA IN A "FEDERATED" CONTEXT

In addition to the presence of metadata on an institution’s own website, and in the local OPAC, metadata can (and should) be shared in a variety of “union” environments, e.g.:

- WORLDCAT
- OCLC ART RESOURCE DISCOVERY
- GROUP CATALOGUE
- ARCHIVES GRID
- GOOGLE!

Information standards and controlled vocabularies can help extricate us from our metadata dilemmas...

DON’T GO INTO THIS BLINDFOLDED!

- What is the focus of your project, and what research questions do you want to ask?
- Where will your data come from?
- What is your source of labor?
- What are the intended users and uses?
- What is your data model?
- What standards will you follow?
- What will be the end-product?
- Where will your end-product “live”?
- How will users find it?

What is a “record”?
What is a “record”? Descriptive metadata records for an image in 19th-century album, the object depicted, and link to object on current repository’s website (from the INHA “Digital Montagny” project).

A Typology of Data Standards

<table>
<thead>
<tr>
<th>Type of Data Standard</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data structure standards (metadata element sets, schemas)</td>
<td>MARC (Machine-Readable Cataloging format), Dublin Core Metadata Element Set (DCMES), Categories for the Description of Works of Art (CDWA), VRA Core Categories</td>
</tr>
<tr>
<td>Data value standards (controlled vocabularies, thesauri, controlled lists)</td>
<td>Library of Congress Subject Headings (LCSH), Library of Congress Name Authority File (LCNAF), LC Thesaurus for Graphic Materials (TGM), Medical Subject Headings (MeSH), Art &amp; Architecture Thesaurus (AAT), Union List of Artist Names (ULAN), Getty Thesaurus of Geographic Names (TGN), ICONCLASS</td>
</tr>
<tr>
<td>Data content standards (cataloging rules and codes)</td>
<td>Anglo-American Cataloguing Rules (AACR), Resource Description and Access (RDA), International Standard Bibliographic Description (ISBD), Cataloging Cultural Objects (CCO), Describing Archives: A Content Standard (DACS)</td>
</tr>
<tr>
<td>Data format/technical interchange standards (metadata standards expressed in machine-readable form)</td>
<td>MARC21, MARCXML, BIBFRAME, EAD XML DTD, METS, MODS, CDWA Lite XML schema, Simple Dublin Core XML schema, Qualified Dublin Core XML schema, VRA Core 4.0 XML schema</td>
</tr>
</tbody>
</table>

RELATIONSHIP BETWEEN “RECORDS” AND CONTROLLED Vocabularies: Data “structures” populated with data “values”
DETERMINING WHAT METADATA IS NEEDED
➢ Who are your users? (current as well as potential) (e.g., library or registrarial staff, curators, professors, advanced researchers, students, general public)
➢ What information do you already have (even if it’s only on index cards)?
➢ What information is already in automated form?
➢ What metadata categories & vocabulary tools are you currently using? Are they adequate for all potential uses and users? Do they map to any standard?

WHAT DATA DO YOU NEED?
❑ What common or core data is needed?
❑ What data do your various user groups need?
❑ What established metadata standards (e.g., MARC, METS, EAD, Dublin Core, VRA Core, LIDO) might fit the information needs of your collections and/or institution and your USERS?

DATA STANDARDS: ESSENTIAL STEPS

First Step: Select and Use Appropriate Metadata Element Sets
Data Structure Standards
(a.k.a. metadata standards)
✓ Guidelines for the structure of information systems: What elements should a database include?
✓ Meant to be customized according to institutional and/or project needs.
✓ MARC, EAD, MODS, Dublin Core, LIDO, VRA Core are examples of data structure standards.
Second Step: Select and Use Vocabularies, Thesauri, Classifications, and “Folksonomies”

- Data values are used to “populate” or fill metadata elements
- Examples are LCSH, AAT, TGM, MeSH, etc., as well as “local” vocabularies

Third Step: Follow Guidelines for Documentation

**Data Content Standards**
- Best practices for documentation (i.e., implementing data structure and data value standards)
- Rules for the selection, organization, and formatting of content.
- AACR, RDA (the successor to AACR), DA:CS (Standard), CCO (Cataloging Cultural Objects)

Fourth Step:
Select the Appropriate Format for Expressing Data

**DATA FORMAT STANDARDS**
- How will you “publish” and share your data in electronic form?
- How will service providers obtain, add value, and disseminate your data?
- Candidates are Dublin Core XML; MARC21; MARC XML; VRA XML schema; LIDO XML schema; MODS, etc. And more recently—Linked Open Data (LOD).
Looking at a tried-and-true metadata standard for libraries:

MARC

MARC (MACHINE-READABLE CATALOGING) FORMAT

- MARC is the technical “container” for the data in a bibliographic record (both a data structure and a data format standard)
- MARC records are formulated according to the Anglo-American Cataloguing Rules, 2nd edition, 1988 revision (AACR2), and now according to Resource Description and Access (RDA)
- MARC can be used to catalog books, audiovisual materials, sound recordings, computer files, and archival materials

http://lcweb.loc.gov/marc/

Library of Congress MARC record

Brief display of previous LC MARC record, with human-readable field labels instead of alphanumeric indicators.
MARC records can also be expressed in XML format:

See http://www.loc.gov/standards/marcxml/
METS:
A METADATA “WRAPPER” FOR DIGITAL INFORMATION OBJECTS

METS

(Metadata Encoding & Transmission Standard)

METS is an XML schema designed for creating XML document instances that express the complex structure of digital objects, the names and locations of the files that comprise those objects, and the associated metadata.

DUBLIN CORE:
“METADATA WITHOUT PAIN”?

WHY IS DUBLIN CORE SO PREVALENT?

• Dublin Core is the basic required metadata schema for OAI metadata harvesting
• DC is widely used in “aggregated” resources and for metadata mapping/crosswalks (e.g. Getty Research Portal: http://portal.getty.edu/)
• “Lowest common denominator”
• The format is incorporated into systems such as CONTENTdm (http://www.oclc.org/en-US/contentdm.html) and Omeka (https://omeka.org/)
LINKED OPEN DATA (LOD): THE HOLY GRAIL FOR CONNECTING & SHARING METADATA FOR RE-USE & RE-PURPOSING?

THINKING ABOUT AND VISUALIZING DATA AND RELATIONSHIPS: ENTITY-RELATIONSHIP MODELS

ENTITY-RELATIONSHIP MODEL
— first posited by Peter Chen of M.I.T. in 1976

http://portal.acm.org/citation.cfm?id=320440#abstract

FRBR (Functional Requirements for Bibliographic Records) Entity-Relationship Diagram
Cataloging Cultural Objects/CDWA
Entity-Relationship Diagram

CDWA/CCO Entity-Relationship Diagram

Entity-relationship diagram of information and resources relating to item from Getty Research Institute Special Collections

Storyboard

A sequence of drawings, typically with some directions and dialogue, representing the shots planned for a movie or television production.
### Main Points to Address in the Class Exercise & Presentation

- database, interactive website, data repository, digital publication, collection of digital objects, something else
- Who are your intended users, and what do you expect they will want to do with your resource?
- What metadata standard(s) will you use, and why?
- What controlled vocabularies or thesauri will you use, and why?
- Will your resource be “open content,” and if so, what issues will you need to address? How will you make your metadata “shareable?” and “re-usable?”

### Main Points to Address Continued

- reformatted/scrubbed, created from scratch, or a combination?
- What is your strategy for the discoverability of your resource? (e.g. from search engines like Google and/or online catalogs like Worldcat). Will your resource be discoverable in multiple “places”?
- What resources (human, technical, monetary) will you need to build your resource?
- How will your resource be maintained and, if appropriate, updated or “retired”?
- How will you measure success?

---

**Over to you!**