May 1st, 1:00 PM - 4:00 PM

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

Murtha Baca
Getty Research Institute (GRI)

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

https://digital.sandiego.edu/symposium/2017/2017/26

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 the Digital Art History program

Session Type
Workshop

Abstract
This 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. 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.

Location
KIPJ Room A

Comments
Murtha Baca is head of the Digital Art History program at the Getty Research Institute (GRI) in Los Angeles. She holds a PhD in art history and Italian language and literature from UCLA. Murtha is an adjunct faculty member in the Department of Information Studies at UCLA, where she teaches graduate seminars on metadata, indexing, and thesaurus construction. Her research and publications have focused on multilingual documentation of art, architecture, and material culture, and building tools and resources for conducting and publishing art-historical research in the digital age. Her edited volumes include Introduction to Art Image Access (Los Angeles: Getty Publications, 2002) and Introduction to Metadata (revised edition, 2016). She is the chair of the International Terminology Working Group (ITWG), which brings together cultural heritage documentation professionals who develop and utilize multilingual vocabulary tools. Murtha has published numerous articles on metadata and controlled vocabularies for art, architecture, and cultural heritage, and was a member of the editorial team that produced Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images (Chicago: American Library Association, 2006). With Patricia Harpring, with whom she co-authored Categories for the Description of Works of Art, she has twice received the Visual Resources Association’s DeLaurier Award for distinguished service to the field. Murtha was co-principal investigator and project leader for the GRI’s first born-digital critical facsimile publication, Pietro Mellini’s Inventory in Verse, 1681, released in 2015.
METADATA* FOR DIGITAL PROJECTS: THINKING ABOUT THE NUTS & BOLTS

Murtha Baca, PhD
Head, Digital Art History Program
Getty Research Institute (GRI)
University of San Diego DIGITAL SYMPOSIUM, May 2017

* and other essential elements

Agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:00</td>
<td>Introduction and goals for the workshop. Participants will introduce themselves and briefly describe the digital projects that they are planning or would like to do.</td>
</tr>
<tr>
<td>1:15</td>
<td>Instructor's presentation (with class discussion and questions): overview of metadata formats; &quot;Deep Web&quot; versus &quot;Visible Web,&quot; controlled vocabularies, &quot;project-specific vocabularies,&quot; social tagging; issues of access and metadata sharing; steps for developing a metadata strategy; entity-relationship diagrams and &quot;storyboards.&quot;</td>
</tr>
<tr>
<td>2:30</td>
<td>Break</td>
</tr>
<tr>
<td>2:45</td>
<td>Participants work on storyboards for their specific projects, possibly including a simple conceptual model, data model, or entity-relationship model.</td>
</tr>
<tr>
<td>3:15</td>
<td>Participants present their proposed projects and the instructor and class discuss them.</td>
</tr>
<tr>
<td>3:45</td>
<td>Lessons learned, final thoughts, wrap-up</td>
</tr>
<tr>
<td>4:00</td>
<td>Conclude</td>
</tr>
</tbody>
</table>
“Making a Website” ≠ Doing a Digital Project

Images and other digital assets without accompanying metadata are mostly useless, and generally “unfindable,” un sharable, and not re-usable.

Digital Projects—Why bother?

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
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
- for retention of context
- for expanding use & sharing
- for multi-versioning
- for legal issues
- for preservation of data

Information standards and controlled vocabularies can help extricate us from our metadata dilemmas...
What is a “record”? 

Allegory of Fortune

Michelangelo:
Allegory of Fortune: 1524

This image is available for download, without change, under the Getty’s Open Content Program.

Darian Demir
Bakov, about 1530
Oil on canvas
79 1/2 x 85 1/2 in.
89.PA.32
Currently on view at
The Getty Center Los Angeles
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)

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?
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 a person, object, concept, etc., and provide important additional access points.

Bulgarini, Bartolomeo
Bartolomeo Bolgarini
Bartolomeo Bolghini
Bartolomeo Bulgarini
Bartolommeo Bulgarini da Siena
Maestro d’Ovile
Master of the Ovile Madonna
Ovile Master
Lorenzetti, Ugolino
Ugolino Lorenzetti

names from Getty Union List of Artist Names (ULAN)
Ayia Sophia
Ayasofya
Church of the Holy Wisdom
Hagia Sophia
Haghia Sophia
Saint Sophia
Sancta Sophia
St. Sophia

Constantinople
Constantinopolis
Costantinopoli
Estambul
Istanbul
Konstantinopol
New Rome
Mikligard
Tsargrad
Tsarigrad

names from Getty Thesaurus of Geographic Names (TGN)
Using language (and metadata) to reach broader audiences: this is where “collection-specific” or “resource-specific” controlled vocabularies can help.

desk?
desk?
cabinet?
cabinet?
chest?
cartonnier?
cartonnier?

### A Typology of Data Standards
(from Introduction to Metadata)

<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>the set of MARC (Machine-Readable Cataloging format) fields, Encoded Archival Description (EAD), 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)</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), ICONICLASS</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>

This type of standard is often a manifestation of a particular data structure standard (type 1 above), encoded or marked up for machine processing.
RELATIONSHIP BETWEEN “RECORDS” AND CONTROLLED VOCABULARIES: DATA “STRUCTURES” POPULATED WITH DATA “VALUES”

LINKED OPEN DATA (LOD): THE HOLY GRAIL?
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, and Classifications

Data Value Standards

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

Data Value Standards continued

- Used as controlled vocabularies or authorities to assist with documentation and cataloguing.
- Used as research tools—vocabularies contain rich information and contextual knowledge.
- Used as search assistants in database retrieval systems and online collections.
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 (Anglo American Cataloguing Rules), RDA (Resource Description and Access, the successor to AACR), DA:CS (Describing Archives: A Content 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/](http://lcweb.loc.gov/marc/)
MARC records can also be expressed in XML format:

See

http://www.loc.gov/standards/marcxml/

“Classic” MARC record
### MARC XML record

```xml
xmlrecord
- <datafield ind1="1" ind2=""
  tag="008"
/>  
- <subfield code="t">Geographies, etc.</subfield>
- <subfield code="h">1976-1997</subfield>
- <subfield code="q"></subfield>
- <datafield ind1="1" ind2=""
  tag="245"
/>  
- <subfield code="a">Authentic</subfield>
- <subfield code="d">
  Our Sandburg; illustrated as an accomplice adventure by Ted Rand
</subfield>
- <subfield code="h">1995</subfield>
- <subfield code="n">1</subfield>
- <subfield code="i">1st ed.</subfield>
- <subfield code="c">
  San Diego, Calif., Harcourt Brace Jovanovich
</subfield>
- <subfield code="p">1993</subfield>
- <subfield code="a">26 cm.</subfield>
- <datafield ind1="1" ind2=""
  tag="500"
/>  
- <subfield code="a">Our Mylar slab included in pocket.</subfield>
- <subfield code="a">
  A poem about numbers and their characteristics. Features an example, a distorted drawing which can be restored to normal by viewing from a particular angle or by viewing the image's reflection in the provided Mylar case.
</subfield>
```

### MODS:

**RICHER THAN DUBLIN CORE, SIMPLER THAN MARC**

![MODS Image]
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/](http://portal.getty.edu/))
- “Lowest common denominator”
- The format is incorporated into systems such as CONTENTdm ([http://www.oclc.org/en-US/contentdm.html](http://www.oclc.org/en-US/contentdm.html)) and Omeka ([https://omeka.org/](https://omeka.org/))
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

CDWA/CCO Entity-Relationship Diagram

Cataloging Cultural Objects/CDWA
Entity-Relationship Diagram
Entity-relationship diagram of information and resources relating to item from Getty Research Institute Special Collections

Another simple visualization of the CCO/CDWA Entity-Relationship Model, stressing the use of authorities (aka controlled vocabularies)
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

- What type of resource will you create? (e.g. searchable 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?
- Will your resource be “open content,” and if so, what issues will you need to address? Will your data be “shareable?”
- What metadata standard(s) will you use, and why?
- What controlled vocabularies or thesauri will you use, and why?
MAIN POINTS TO ADDRESS CONTINUED

- Will the data for your digital resource be re-purposed from an existing source, created from scratch, or a combination of both?
- 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 you measure success?

Over to you!