Developing a Data Format Standard Informed by CCO: The CDWA Lite XML Schema, and Thoughts on Cross-cultural Standards & Shareable Metadata

Murtha Baca
Getty Research Institute
ARLIS 2007
Atlanta

The state of art information standards pre-CCO

- **Data structure standards**: Categories for the Description of Works of Art (CDWA), VRA Core Categories, MARC (esp. MARC VIM)
- **Data value standards**: LCSH, LCNAF, Thesaurus for Graphic Materials (TGM), Art & Architecture Thesaurus (AAT), Union List of Artist Names (ULAN), Thesaurus of Geographic Names (TGN), ICONCLASS, others

Data Structure Standards for Art & Material Culture

- **CDWA**: http://www.getty.edu/research/conducting_research/standards/cdwa/
- **Object ID**: http://icom.museum/object-id/

The state of art information standards pre-CCO cont.

- **Data content standards**: There was no “AACR for art objects.” No tradition of standards specifically for art & cultural heritage museum & visual resources cataloging. (The response was CCO.)
- **Data format/technical interchange standards**: There was no “MARC for art objects.” (The CDWA Lite schema is one response to this need; the VRA Core 4.0 XML schema is another.)
Origin and Goals of CDWA Lite

- The Getty wanted to make authoritative, up-to-date information on the works in its collections widely available, in a variety of "venues."
- ARTstor ("service provider") asked the Getty to contribute to its Image Gallery.
- The Getty and ARTstor worked together to develop a replicable, standards-based way for institutions to contribute data and images relating to cultural heritage collections to union catalogs like ARTstor’s Image Gallery (and OCLC’s WorldCat, RLG’s Cultural Materials, etc.).

Goals cont.

- To develop a data dictionary (specification) & related XML schema suited to representing cultural objects, based on appropriate data structure (CDWA) & data content (CCO) standards
- To reduce overhead for contributing to union catalogs/service providers: do it once, do it right, share it with everybody
- To reduce labor and "delivery" costs
- To ensure a mechanism for updating data
- To include links from contributed metadata back to records in their "home" context

Why not just use Dublin Core?

- Dublin Core, currently the preferred schema for OAI harvesting, was designed for Web resource discovery
- Not adequate, or appropriate, for describing cultural materials
- Other schemas (MARC XML, MODS) are more appropriate than Dublin Core for other types of materials

Essential Elements

- An XML schema/data format standard appropriate for expressing information on cultural objects & their visual surrogates ⇒ CDWA Lite (not Dublin Core)
- A replicable, standard technical protocol for delivering, sharing, & disseminating the information expressed in that format ⇒ the OAI Protocol for Metadata Harvesting (OAI/PMH)
The CDWA Lite/OAI Harvesting Model

The CDWA Lite/OAI Harvesting Model involves a Data Provider that aggregates metadata records and "resources" from diverse data providers. The Metadata Repository "gets" records via OAI/PMH "gets" requests.

The Service Provider issues requests to the repository and provides an aggregated search service, browsing, and other enhancements for end-users. End-users access a "union catalog" of metadata records and resources, with links back to the original (possibly fuller) records and resources in their "home" environment.

Web References:
- http://www.getty.edu/research/conducting_research/standards/intrometadata/index.html
- http://www.getty.edu/research/conducting_research/standards/intrometadata/index.html

Origin of the CDWA Lite elements

Mapping data categories/elements

- CDWA/CCO required categories were mapped to an XML schema = CDWA Lite

 Metadata Standards Crosswalks

<table>
<thead>
<tr>
<th>CDWA</th>
<th>CCO</th>
<th>CDWA Lite</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJECT/WORK (core)</td>
<td>Record Type</td>
<td>cdwalite:objectWorkType</td>
</tr>
<tr>
<td>Objective/Work-Type (core)</td>
<td>Work Type</td>
<td>cdwalite:objectWorkType</td>
</tr>
</tbody>
</table>

- CDWA/CCO required categories were mapped to an XML schema = CDWA Lite
Data dictionary/spec and XML schema to describe core records for works of art, architecture, and material culture

Based on the required fields, structure, guidelines discussed in CDWA

Informed by the cataloging rules in CCO

The J. Paul Getty Museum at the Getty Center: Paintings

The J. Paul Getty Museum’s paintings collection includes European paintings from roughly 1300 to 1900. The collection also includes a number of pastels, predominately from the 18th century.

The Collections (OAI-harvestable “sets”)

Research Library, Getty Research Institute, Photo Study Collection: Tapestries

The Getty Research Institute’s Photo Study Collection contains more than 2 million study photographs, many of which are historic in nature. Featured here are 4,215 images from the Tapestries Collection, most of which are based on the French & Company dealer archive.

The Data “at home”

• J. Paul Getty Museum
  Paintings Collection (461 records and 461 digital still images)

  • Data lives in a collection management system (relational database)

  • Tapestries - Photo Study Collection of the Research Library, Getty Research Institute (4215 records and 9063 digital still images)

  • Data lives in a flat-file system with linking capabilities

The Data in ARTstor’s Image Gallery

Attributed to Hans Holbein the Younger
An Allegory of Passion
German, 1530
Oil on panel

The Data at the Getty Center: Paintings

The J. Paul Getty Museum’s paintings collection includes European paintings from roughly 1300 to 1900. The collection also includes a number of pastels, predominately from the 18th century.

The Collections (OAI-harvestable “sets”)

Research Library, Getty Research Institute, Photo Study Collection: Tapestries

The Getty Research Institute’s Photo Study Collection contains more than 2 million study photographs, many of which are historic in nature. Featured here are 4,215 images from the Tapestries Collection, most of which are based on the French & Company dealer archive.

The Data “at home”

• J. Paul Getty Museum
  Paintings Collection (461 records and 461 digital still images)

  • Data lives in a collection management system (relational database)

  • Tapestries - Photo Study Collection of the Research Library, Getty Research Institute (4215 records and 9063 digital still images)

  • Data lives in a flat-file system with linking capabilities

The Data in ARTstor’s Image Gallery

Attributed to Hans Holbein the Younger
An Allegory of Passion
German, 1530
Oil on panel

http://www.getty.edu/research/conducting_research/standards/cdwa/cdwalite.html
Harvested Getty Museum record in ARTstor

Harvested Getty Museum record includes link to Getty Web site

Paintings collection on Getty Web site

Getty museum object on the Getty Web site, with additional information & images
Harvested Photo Study Collection record on the Getty Web site.

Research Library, Getty Research Institute Photo Study Collection Study Images of Tapestries collection Web page.
Decisions

- The Getty Research Institute decided to contribute much fuller records to ARTstor than the Getty Museum did (Getty Museum records and images are "Visible Web," while the GRI Tapestries records & images are "Deep Web.")
- GRI contributed more than one "resource" (image) per metadata record, when available (Getty Museum did not).
- Data providers have considerable leeway in how basic or how full they want their harvestable records to be, and which resources they want to contribute.

Issues

- Like many VR collections, the GRI's Photo Study Collection uses a variety of non-standard (but mappable), locally-developed metadata schemas.
- The Tapestries metadata records are "hybrid"—they describe both the work and the image (also typical of many VR collections)
- Some cataloging decisions that work locally don’t translate well in a union environment.

What are the incentives for using CDWA Lite?
(or another standards-based schema + the OAI harvesting protocol)

Data that did not come from the collecting institution is out of date—these works were de-accessioned in the late 1980s!
Subject headings enhance access, but not all data is up to date or matches the repository’s data.

Work Type: sculpture
Title: Cult Statue of a Goddess, perhaps Aphrodite
Creator: Unknown Greek (South Italian)
Measurements: H: 220 x W: 67 cm (86 5/8 x 26 3/8 in.)
Materials: limestone and Parian marble with polychromy
Creation Date: 425–400 B.C.E.
Repository: The J. Paul Getty Museum at the Getty Villa (Malibu, California) 88.AA.76
Rights: © J. Paul Getty Museum; http://www.getty.edu/image_rights
Record Source: J. Paul Getty Museum. Object ID: 115100
Image: http://www.getty.edu/art/gettyguide/artObjectDetails?artobj=15050&handle=li

Another XML schema informed by CCO:
VRA Core 4.0

CDWA Lite and VRA Core 4.0 are included among the Potential Metadata formats for Use with the OAI/PMH on the OAI Best Practices pages of the National Science Digital Library (NSDL) and the Digital Library Federation (DLF)

http://oai-best.com/nsdl.org/cgi-bin/wiki.pl?MultipleMetadataFormats

CCO & CDWA Lite and Their Potential “Interaction” with Library and Archival Standards: “Cross-cultural” Metadata

EAD (data structure) & DACS (data content) used at the collection level for an archival collection with a common provenance

CDWA Lite (data structure/data format) & CCO (data content) at the item level for an individual work within the same collection
MARC (data structure/data format) and AACR (data content) used for a "parent" item (18th-century book with engravings) in OPAC.

There's no such thing as "one-standard-fits-all"!
Examples of AACR's inadequacies for describing art objects, and another argument for combining cross-cultural standards.

CCO (data content), and "Getty VRA*" (data format) used at the item level for an individual engraving from the "parent" work represented in the preceding MARC record.

"*an adaptation of VRA Core 3.0"
Title (CCO):
Arch Decorated with Double-headed Imperial Eagle and Gilt Lions Spouting White and Red Wine
Results list from same search, with CCO “display titles” co-existing with AACR titles.

Lessons Learned

- Metadata (descriptive, technical, rights, administrative, preservation) is one of our biggest investments.
- Do it right (consistent schemas, controlled vocabularies), and you can re-purpose metadata in a wide variety of ways.
- Good descriptive metadata records can be core—records don’t need to be “full” to be “good.” More data can be added later.
- Creation of consistent, standards-based descriptive metadata (a.k.a. cataloging) is time- and labor-intensive, but it’s worth it.

Some Emerging Trends in Metadata Creation

- “Schema-agnostic” metadata
- Metadata that is both shareable and re-purposeable
- Harvestable metadata (OAI/PMH)
- “Non-exclusive”/“cross-cultural” metadata—i.e., it’s okay to combine standards from different metadata communities—e.g. MARC and CCO, DACS and AACR, DACS and CCO, EAD and CDWA Lite, etc.
- Importance of authorities—and difficulties in “bringing along” the power of authorities with shared metadata records
- The need for practical, economically feasible approaches to metadata creation
Solutions to the Vocabulary Problem?

- Include variants, broader & narrower terms in metadata record for works (labor-intensive, redundant)
- Service providers/aggregators employ the appropriate controlled vocabularies & thesauri as "search assistants" (promising, but not now a reality)
- Can RLG Programs (including OCLC's Terminologies Service) help?

http://www.oclc.org/terminologies/

Standards are like toothbrushes...
Everyone agrees they are a good idea...
But nobody wants to use anyone else’s.

mbaca@getty.edu

http://www.getty.edu/research/conducting_research/vocabularies/

http://www.getty.edu/research/conducting_research/standards/