Metadata, Ontologies, Taxonomies, Oh My!

Presentation for the ALCTS Symposium
"Preparing 21st Century Cataloging and Metadata Professionals,"
San Diego, CA, January 9, 2004
by Arlene G. Taylor

Tradition! Tradition!

- Description
- Access points for names and titles
- Subject headings
- Classification
- Encoding

Encoding - Purposes

- To provide for access to each part of a record
- To provide for display
- To allow integration of many languages and scripts
- To allow for data transmission

Encoding – How accomplished

- Assign tags, numbers, letters, or words (i.e., codes) to discrete pieces of information
- Use standard codes
- Use frameworks and wrapper technologies
  - Warwick Framework
  - RDF (Resource Description Framework)
  - METS (Metadata Encoding & Transmission Standard)
  - Semantic Web

Standard Encoding Schemes

- Current standards for encoding records
  - MARC
    - MARC 21 (formerly USMARC and CAN/MARC)
    - UNIMARC
  - SGML / XML
    - DTDs and XML Schemas
      - TEI – for encoding literary texts
      - HTML/XHTML – for encoding Web pages
      - EAD – for encoding archival finding aids
      - MARC DTDs and XML schemas – for encoding MARC 21 records
      - ONIX – for encoding publishers' records

5

Description - Purposes

- To present the characteristics of an information package
- To give enough information about an information package to identify it uniquely and to distinguish it from every other information package
- To aid in evaluating or selecting (e.g., Is the original manuscript of a book needed, or will a printed copy do? Is the 8th ed. as good as the 9th ed. for my purposes? Is a vinyl LP o.k., or can I only play a CD?)
- To provide a filter that serves as a surrogate for a full information package so that users do not have to examine a multitude of complete (e.g., full text) packages in order to find what is needed

6

Description – How Accomplished

- Determine the unit to be described
  - "catalogable" unit
  - finite vs. continuing resources
  - work – expression – manifestation – item
- Create surrogate records by selecting important pieces of information from or about the information package
- Use rules or conventions created by different communities to determine which pieces of information will be included

7

Standard Descriptive Schemas

- Bibliographic and General Metadata Schemas
  - ISBD (International Standard Bibliographic Description)
  - AACR2R (Anglo-American Cataloguing Rules, Second Edition)
  - Dublin Core
  - MODS (Metadata Object Description Schema)

8

Standard Descriptive Schemas (cont.)

- Domain-Specific Metadata Schemas
  - ISAD(G) (General International Standard Archival Description)
  - APPM (Archives, Personal Papers, and Manuscripts)
  - EAD (Encoded Archival Description)
  - TEI (Text Encoding Initiative) Headers
  - GILS (Government Information Locator Service)
  - FGDC Content Standard for Digital Geospatiol Metadata (CSDGM)
  - VRA Core Categories for Visual Resources
  - ONIX (Online Information eXchange)

9

Access Points - Purposes

- To identify (e.g., an entity known to the user)
- To collocate (i.e., bring together related information packages)
- To aid in evaluating or selecting (e.g., Has this author written something newer on the subject? Which of several works with the same title do I want? What level of subject treatment is needed – a whole work on the subject? a chapter? a paragraph?)
- To locate a copy of the information package represented

10

Access Points for Names and Titles - Purposes

- To facilitate the retrieval of names and titles that are imperfectly remembered
- To facilitate the retrieval of names and titles that are expressed differently in different information packages
- To facilitate the retrieval of names and titles that have changed over time
- To collocate expressions and manifestations of works
- To collocate works that are related to other works

11

Access Points for Names and Titles – How Accomplished

- Name and Title Authority Control
  - All access points (whether main or added entries) need to be under authority control so that
    - persons or entities with the same name can be distinguished from each other
    - all names used by a person or body, or all manifestations of a name of a person or body will be brought together
    - all differing titles of the same work can be brought together
  - Therefore, current practice dictates either the establishment of a "heading" for each name or title as an access point or the provision of pointers to draw different representations of names or titles together
  - Headings are kept track of in authority files; RDF provides a model for linking entities

12

Name Authority Standards

- LCNAF (Library of Congress Name Authority File) – constructed according to principles set out in AACR2R
- Getty Vocabulary tools (artist names; geographic names) – VRA Core Categories calls for use of the Getty vocabulary
- ISAAR(CPF) – International Standard Archival Authority Record for Corporate Bodies, Persons and Families
- EAC – Encoded Archival Context (for describing creators of archival collections)
- DCMI Agents – creators, contributors, and publishers – to be used in Dublin Core records

Controlled Subject Terminology - Purposes

- To provide subject access to information packages in a catalog or index
- To collocate surrogate records for information packages of a like nature
- To provide suggested synonyms and syndetic structure to aid a user in subject searching
- To save the users' time

Controlled Subject Terminology – How Accomplished

- Conceptual analysis – describe aboutness in natural language
- Translate that analysis into the framework of the controlled vocabulary system (e.g., use of single concept terms vs. use of phrases, compound concepts, and precoordinated subdivisions)
- Use controlled vocabulary system rules to create controlled subject access points to be added to metadata records

Controlled Vocabularies

- Subject heading lists
  - LCSH (Library of Congress Subject Headings)
  - FAST (Faceted Access to Subject Terminology)
  - Sears List of Subject Headings
  - MeSH (Medical Subject Headings)
- Thesauri
  - AAT (Art & Architecture Thesaurus)
  - Thesaurus of ERIC Descriptors
  - INSPEC Thesaurus
  - Many more...

Controlled Vocabularies (cont.)

- Ontologies
  - OWL Web Ontology Language
  - WordNet®
  - UMLS® (Unified Medical Language System)
- Natural Language Processing
  - Semantic, syntactic, and morphological analysis that provides "control" of a user's natural language queries

Classification - Purposes

- To categorize information packages into knowledge organization schemes
- To collocate information packages by subject or form/genre
- To provide a logical location for similar information packages
- To arrange and retrieve information packages and/or their surrogates

Classification – How Accomplished

- Conceptual analysis – describe aboutness in natural language
- Translate that analysis into the framework of the categorization system (e.g., hierarchical or faceted)
- Use classification system rules to create notations to be added to metadata records or to create categories into which to anchor metadata records

Categorization systems

- Bibliographic classifications
  - DDC® (Dewey Decimal Classification)
  - UDC (Universal Decimal Classification)
  - LCC (Library of Congress Classification)
- Taxonomies
  - usually subject-specific categorized lists of terms
  - lists of "taxonomies" include classification schemes, subject heading lists, Internet directories and gateways, as well as subject-specific tools
  - often proprietary to the organizations that created them

Categorization systems (cont.)

- Internet categorized "drill down" approaches (e.g., Yahoo!, Google, etc.)
- Artificial neural networks
  - automatic categorization of documents (often Web pages)
  - categories displayed in a visual representation of a collection of information, with similar documents clustered together and with similar subjects displayed near each other (e.g., WEBSOM, Smartmoney.com)

Finale

- Principles for creation of surrogate records that have been developing over hundreds of years can be used to catalog (to metadate?) anything!
- The challenge is to get students to concentrate on applying principles rather than obsessing on rules.