Metadata, Controlled Vocabulary, and Thesauri

James Vorbach, Ph. D.

Outline

• Metadata
• Semantic Relationships
• Controlled Vocabulary
• Thesauri

Metadata

• Motivating Questions
  – How can a document be described?
  – Where does a document fit in the taxonomy?
  – What keywords should be assigned to the document?

Definition

1. Metadata literally “data about data”, is the sum total of what one can say about any information object. An information object is anything that can be addressed and manipulated by a human or a system as a discrete entity.

Characteristics

• Content relates to what an object contains or is about.
• Context indicates the who, what, why, where, how aspects associated with an object’s creation.
• Structure relates to the formal set of associations within or among information objects

Definition

2. A metadata record consists of a set of attributes, or elements, necessary to describe the resource in question. For example, a metadata system common in libraries – the library catalog – contains a set of metadata records with elements that describe a library item: author, title, date of creation or publication, subject, and the call number.
Applications

• Improving resource discovery by enabling field-based (e.g. author, title) searches
• Permitting indexing of non-textual objects
• Serving as a common exchange format such as the Open Archive Initiative

Dublin Core

• Metadata standard
• Consists of two levels
  – Simple: 15 elements
  – Qualified: includes an additional element, Audience, as well as a group of element refinements called qualifiers

Sample of the Dublin Core Element Set

<table>
<thead>
<tr>
<th>Element</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>A name given to the resource</td>
</tr>
<tr>
<td>Creator</td>
<td>An entity responsible for making the content of the resource.</td>
</tr>
<tr>
<td>Subject</td>
<td>A topic of the content of the resource.</td>
</tr>
<tr>
<td>Description</td>
<td>An account of the content of the resource.</td>
</tr>
<tr>
<td>Publisher</td>
<td>An entity responsible for making the resource available.</td>
</tr>
</tbody>
</table>

Relational Database Implementation

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
<th>Publisher</th>
<th>Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>UKOLN</td>
<td>UKOLN is a national focus of expertise in digital information management. It provides policy, research and awareness services to the UK library, information, and cultural heritage communities.</td>
<td>UKOLN, University of Bath</td>
<td><a href="http://www.ukoln.ac.uk">http://www.ukoln.ac.uk</a></td>
</tr>
</tbody>
</table>

XML Implementation

```xml
<?xml version="1.0"?>
<metadata
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  ...
  <dc:title>UKOLN</dc:title>
  <dc:description>UKOLN is a national focus of expertise in digital information management. It provides policy, research and awareness services to the UK library, information, and cultural heritage communities.</dc:description>
```
Semantic Relationships

- Equivalence
  - Terms that are equivalent in an information retrieval sense
  - Includes synonyms, misspellings, and closely related terms
- Hierarchical
  - Categories and subcategories
  - Types
    - Generic
      - A Subcategory member is a category member and inherits the properties of the category
    - Whole – Part
      - Subcategory is part of the category
    - Instance
      - Subcategory is an instance of category
- Associative
  - The other relationship; relationships not classified as equivalence or hierarchical
  - Types
    - Field of study and object of study
      - Example: Cardiology RT Heart (RT denotes related term)
    - Process and its agent
      - Example: Termite Control RT Pesticides
    - Action and the Result of Action
      - Example: Eating RT Indigestion

Measures of Effectiveness

- Precision Ratio
  \[ \frac{\text{# of relevant documents retrieved}}{\text{total number of documents retrieved}} \]
- Recall Ratio
  \[ \frac{\text{# of relevant documents retrieved}}{\text{total number of relevant documents in system}} \]
- In general, precision improves at the expense of recall

Controlled Vocabularies

- Synonym Rings
- Authority Files
- Classification Schemes
- Thesauri

Synonym Rings

- List of equivalent terms
- Example: cuisinart, food processor, cuizinart, kitchen aid, kitchenaid, blender
- Impact on the search: the terms in the list are logically 'Or'ed
- criteria: category=cuisinart Or category=food processor, Or category=cuizinart Or category=kitchen aid Or category=kitchenaid Or category= blender

Authority Files

- List of preferred terms
- Usually implemented in combination with synonym rings
- Examples:
  - 2 letter abbreviations for the 50 states
  - CT (preferred term), Connecticut, Conn, Connecticut, Constitution State
  - //authorities.loc.gov
Classification Schemes

- Hierarchical arrangement of preferred terms
  - Example 1: US -> Rhode Island -> Providence
  - Example 2: Dewey Decimal System (DDS)
    //www.oclc.org/dewey/about/default.htm
- Visible when using navigation systems; browsing before and after searching
- Serves as a backend tool for organizing and tagging documents

Thesaurus

- Semantic network of concepts where connections among words represent equivalence, hierarchical, and associative relationships
  - Equivalence relationships include synonyms, homonyms, antonyms
  - Hierarchical relationships include broader, narrower terms
  - Associative relationships include relationships such as 'Field of Study – Object of Study' (Cardiology RT Heart) and Process – Agent (Construction RT excavator)

Thesaurus

- Example: Library of Congress Thesauri
  //www.loc.gov/lexico/servlet/lexico
  Thesaurus for the Global Legal Information Network (GLIN)

Uses of Thesauri

- Point of indexing
  - Mapping variant terms to preferred terms when indexing at the document level
  - Integrating the controlled vocabulary into the organization, label, and navigation systems
- Point of searching
  - Retrieval; query terms are matched against the vocabulary represented in the thesaurus

Types of Thesauri

- Classic
  - Used at both point of indexing and searching
- Indexing
  - Used at the point of indexing only
  - Search engine may be ‘off limits’, either because of internal politics or vendor constraints
- Searching
  - Used at the point of searching only
  - Costs associated with document level indexing may be too high; third party content may preclude document level indexing