5. Metadata and Metadata Standards

INFO 202 - 12 September 2007

Bob Glushko
Plan for INFO 202 Lecture #5

- What is metadata?
- Types of metadata
- How much metadata?
Plan for lectures 5-12 - Defining What Something Means

- Metadata [Lecture 5 - today]
- Controlled names and controlled vocabularies [Lecture 6]
- Classification [Lecture 7]
- Metadata for multimedia [Lecture 8]
- Models of documents (document types) [Lecture 9]
- Models of data collections (database schemas) [Lecture 10]
- Ontology and the Semantic Web [Lectures 11 and 12]

Literally "data about data"

A fancy name for an inferior form of cataloguing (disgruntled librarian)

The properties of a thing or its relations to other things serving to
describe it; the physical and production attributes of (information
entities) or documents (Svenonius)

"A description of the attributes and contents of an information
package... that may include descriptive information about the context,
quality and condition, or characteristics of the data" (Taylor, p. 139)

"Metadata consists of data structures used to discuss other data structures. Metadata augments the values of information (or data) with additional properties that explain its meaning, organization, and other characteristics of interest in our models" (Glushko & McGrath, p. 88)

"Information on the organization of the data, data domains, and the relationship between them" (Baeza-Yates, p. 142)
Why Metadata?

Objectives from International Federation of Library Associations
"Functional Requirements for Bibliographic Records" (Svenonius p. 17)

FIND / LOCATE
IDENTIFY
SELECT
OBTAIN
NAVIGATE
But "Meaning is Use"

The IFLA framework takes a narrow view of information resources and information uses.

For a librarian, the Library of Congress classification number is a critical metadata element for a book.

For a bookseller, the LOC number is useful but its current sales price is a more important metadata element.

For a webmaster or IT person providing access to information resources via a portal, metadata like URLs, protocols, and passwords are the most critical metadata.

These latter cases satisfy the traditional information science definition of metadata but only retrospectively.
Types of Metadata (Taylor, "The Organization of Information")

DESCRIPTIVE metadata - what the information object is about; inherently intrinsic properties

ADMINISTRATIVE metadata - who, what, why, where of the object's creation and management; inherently extrinsic properties

STRUCTURAL metadata - information about the structure, format, and composition of the thing being described; can be intrinsic or extrinsic
Descriptive Metadata

Data derived from an information object that describes it

A piece of descriptive data is the content of one of these metadata elements:

- Title
- Name(s) associated with it
- Edition or version
- Publication date

CONTENT STANDARDS govern the datatypes and values that these metadata elements can have
What is Being Described?

Two separate dimensions on which to distinguish what the metadata is associated with:

- Abstraction hierarchy
- Granularity
The "Abstraction Hierarchy" of the "Work" (Svenonius Ch. 3)

WORK - an abstract entity; the distinct intellectual or artistic creation; it has no single material manifestation

EXPRESSION - the multiple realizations of a work in some particular medium or notation, where it can actually be perceived

MANIFESTATION - each of the formats of an expression that have the same appearance; but not necessarily the same implementation

ITEM - a single exemplar of a manifestation; if we distinguish this level it is because otherwise identical manifestations have some differentiation
Metadata Granularity

An object can be described at various levels of contexts/containers/collections in which it occurs.

Physical objects are more easily bounded than information objects.

For information objects the boundaries between levels of description are less clear.

And it can seem a little circular because we can define "information object" as anything that can be addressed and manipulated by a person or system as a discrete entity.
Physical objects are more or less permanent or their "condition" changes very slowly (like deterioration)

Digital objects can be changed readily, often w/o notice, and so issues of versioning/edition arise
Administrative Metadata

- Location information
- Acquisition information
- Preservation metadata
- Ownership, rights, permission, reproduction information
- Usage information
The User can be a Process, Not just a Person

from Ken Laskey, "Metadata Concepts to Support a Net-Centric Data Environment"
http://www.mitre.org/work/tech_papers/tech_papers_05/04_1279/04_1279.

Metadata describes how the entity and its content can be accessed (both procedurally and the terms of access) in either a read or write mode or executed if the entity comprises processing instructions.

It can contain pointers to information not explicitly part of a given metadata set but which is required as processing or control inputs by other applications or services.
Implications of the Expanded Definition

Broader contexts of use that explicitly acknowledge the use of metadata by processes/services as well as people

Considers information services, not just information objects

Implies the possible existence of multiple metadata sets, one for each context

The metadata description must be expressed in a universally accessible format

The information consumer must be able to access the content or invoke processing on it without knowing APIs or other implementation details about the resource

The information provider needs information about the consumer to determine if access is authorized
Structural Metadata

Information about the structure, format, and composition of the thing being described

This might include data format, file size, running time, digitization or compression specifications, encryption - other characteristics related to the technology realization of the object

Could include hardware or software requirements for using the information
Contextual Metadata

The category of CONTEXTUAL metadata is an alternative and "trendy" category that cuts across the ADMINISTRATIVE and STRUCTURAL ones:

- Metadata about the context in which some content was "captured" -- usually by automated means
- Location, time, other people or things present are basic elements, but there are many more
- This kind of information has often been collected, but not usually analyzed and applied to description until afterwards
Metadata Location

EMBEDDED -- Metadata with the object
  - In the "header"
  - In the "body" as one of the components of the object

EXTERNAL -- Metadata separate from the object it describes
  - Metadata repositories
Few information objects exist in isolation, and it is helpful in resource discovery and retrieval if the relationships among them are encoded in descriptive and structural metadata:

**EQUIVALENCE** - relates copies, facsimiles, reprints, microforms, record/tape/disc, etc

**DERIVATIVE** - relates editions, revisions, adaptations

**DESCRIPTIVE** - description, criticism, evaluation, review of a work
Relationships Among "Bibliographic Entities" - Tillett's Taxonomy [2]

WHOLE-PART - relates a work to a larger work of which it is a part; selections from anthologies, collections, journals

ACCOMPANYING - relates a work to complementary works

SEQUENTIAL - relates a work to preceding or successive parts, prequels and sequels

SHARED CHARACTERISTICS - relates a work to works by same author, etc.
But Svenonius Warns Us...

Because the choice objective is capable of spawning description *ad infinitum*, it is economically untenable (p. 23)

Attempts to cope with the unwanted economic consequences of open-ended objectives surface periodically as a rethinking of a "core" set of essential metadata to be used in description (p. 23)

An important question is whether the bibliographic universe can be organized both intelligently (to meet the traditional bibliographic objectives) and automatically (p. 25)

Any task that requires an organizing intelligence to engage in research is costly (p. 26)
How Much Metadata, What Kind, and by Whom?

You must consider the tradeoffs between organization and retrieval

Not all documents / resources need the same amount of metadata

The same metadata elements or attributes might need different amounts of semantic precision for different document types or contexts ("A laxer form of vocabulary control"-- Svenonius p. 26)
Levels / Sources of Metadata

SIMPLE metadata, unstructured, existing in or extracted from the contents of an information object / document / instance

- But "without formal rules, metadata description is no better than keyword access" (Taylor, p. 142)

STRUCTURED metadata, possibly following a template or schema (a metamodel) created by the author or other person who isn't a professional "producer of metadata"

RICH or BIBLIOGRAPHIC metadata, created by professional "producers of metadata" according to standard models that may vary by domain or discipline

- This is sometimes called "cataloguing" to distinguish it from "populist" metadata
Metadata Standards

Metadata ELEMENTS are the individual categories/ fields/ tags that contain the separate pieces of the description of some information object.

Metadata STANDARDS specify the sets of elements that meet the requirements of some community or context, the rules by which they are arranged.

Metadata standards might also specify the encoding SYNTAX.

Metamodels or metadata schemas don't always dictate the CONTENT of the metadata elements - these are specified in content standards and controlled vocabularies.

Metadata standards are sometimes called metamodels or metadata schemas.
Dublin Core

Proposed in 1995 as a standard set of metadata elements, simple enough be be supplied by a document's author rather than by a professional metadata-maker

DC is the set of elements, described abstractly and all optional

The semantics of DC were established by an international, cross-disciplinary group of professionals from librarianship, computer science, text encoding, the museum community, and other related fields

There are specifications of how to use it in numerous syntaxes (especially XML and RDF) and languages
The Dublin Core Elements [1]

TITLE -- the name given to the resource

IDENTIFIER -- an unambiguous reference to the resource within a given context

SUBJECT -- the topic of the resource's content; key words or classification phrases

CREATOR -- an entity primarily responsible for making the content of the resource

CONTRIBUTOR -- An entity responsible for making contributions to the content of the resource

PUBLISHER -- the entity primarily responsible for making the resource available

DATE -- a date associated with an event in the life cycle of the resource
The Dublin Core Elements [2]

DESCRIPTION -- an account of the content of the resource; abstract, TOC, etc.

LANGUAGE -- a language of the intellectual content of the resource

TYPE -- the nature or genre of the content of the resource

RIGHTS -- information about rights held in and over the resource

SOURCE -- reference to a resource from which the present resource is derived

RELATION -- reference to a related resource

COVERAGE -- the extent or scope of the content of the resource

AUDIENCE -- a class of entity for which the resource is intended or useful
Dublin Core [Example]

<dc:title>Introduction to cataloging and classification</dc:title>
<dc:creator>Taylor, Arlene G.</dc:creator>
<dc:contributor>Wynar, Bohdan S.</dc:contributor>
<dc:date>1992</dc:date>
<dc:format>book</dc:format>

...
Using the Dublin Core

"Some information may appear to belong in more than one metadata element"

"There is potential semantic overlap between some elements"

"There will occasionally be some judgment required from the person assigning the metadata"
Metadata Incompatibility

All of these metadata models and syntax co-exist but they are not completely compatible.

Some of this incompatibility reflects the different purposes and audiences for which the standard was created.

This is reflected in different scopes and granularity of the metadata elements.

There are also no guarantees of semantic equivalence among the seemingly corresponding metadata elements.
Doctorow on Metadata

People lie
People are lazy
People are stupid
People delude themselves
Metadata metrics distort it
Metadata suffers from "the vocabulary problem"
Assignment 2: Designing a Vocabulary

Develop a vocabulary for describing some aspects of sports or some aspects of music - choose the domain that interests you the most.

Identify and define the terms or semantic components needed in the vocabulary

Test the adequacy of the coverage of your sports or music vocabulary by using it to describe a real or hypothetical event in one existing sport or music category of your choosing

This does not require any XML

Due on next Monday 17 September before class
Readings for INFO 202 Lecture #6

Svenonius Chapter 6, Chapter 8 (127-132)

Karl Fast, Fred Liese, and Mike Steckel. What is a controlled vocabulary?

Karl Fast, Fred Liese, and Mike Steckel. Creating a controlled vocabulary.

Glushko and McGrath, Document Engineering, part of Chapter 12 (399-406)