6. Metadata and Metadata Standards

INFO 202 - 17 September 2008

Bob Glushko

Plan for INFO 202 Lecture #6

What is metadata?

Types of metadata

How much metadata?

Plan for lectures 6-9 - Defining What Something Means

Metadata [Lecture 6 - today]

Controlled names and controlled vocabularies [Lecture 7]

Classification [Lecture 8]

Ontology [Lecture 9]

What Is Metadata? [1]

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)

What Is Metadata? [2]

"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

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

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

The MARC Record

1968 - When the Library of Congress began to use computers in the 1960s, it devised the LC Machine Readable Catalog Format, a system of using brief numbers, letters, and symbols within the cataloging record itself to mark different types of information.

MARC mandates a rich description with strong datatyping and vocabulary control for the values of its metadata elements

In the 1980s (and revised in 2002) the Anglo-American Cataloguing Rules (AACR) extended the MARC standard so that it could describe music and various other kinds of "non-book" entities

This "integration" causes some substantial technical and theoretical concerns

MARC is often criticized for being unsuited to the modern computing environment

The MARC Record [Example]

```
ID:DCLC9124851-B            RTYP:c    ST:p    FRN:    MS:c   EL: AD:06-20-91
CC:9110  BLT:am      DCF:a   CSC:     MOD:     SNR:      ATC: UD:04-11-92
CP:cou     L:eng     INT:    GPC:     BIO:     FIC:0     CON:b
PC:s      PD:1992/           REP:     CPI:0    FSI:0     ILC:a      II:1
MMD:      OR:    POL:     DM:     RR:         COL:      EML:      GEN: BSE:
010      9124851
020      0872878112 (cloth)>
020      0872879674 (paper)
040      DLC$cDLC$dDLC
050 00   Z693$b.W94 1991
082 00   025.3$220
100 1    Wynar, Bohdan S.
245 10   Introduction to cataloging and classification /$cBohdan S. Wynar.
250      8th ed. /$bArlene G. Taylor.
260      Englewood, Colo. :$bLibraries Unlimited,$c1992.
300      xvii, 633 p. :$bill. ;$c24 cm.
440  0   Library science text series
504      Includes bibliographical references (p. 591-599) and index.
650  0   Cataloging.
650  0   Subject cataloging.
650  0   Classification$xBooks.
630 00   Anglo-American cataloguing rules.
700 10   Taylor, Arlene G.,$d1941-
```

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

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 - Pragmatics and Problems

"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

There are other metadata standards: ISBD, RFC 1807, TEI header, ...

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

Achieving Metadata Interoperability [1]

"We do not need a bibliographic record format. We need a bibliographic metadata infrastructure... Our systems must be able to accommodate a great diversity of record formats to provide us with the flexibility and power that only such diversity can provide" (Tennant)

Interoperability doesn't require that two systems be identical in design or implementation, only that they can exchange information and use the information they exchange.

Interoperability requires that the information being exchanged is conceptually equivalent

Achieving Metadata Interoperability [2]

If conceptual equivalence can be established, converting one implementation to another is a necessary but often trivial thing to do

But it isn't always possible to establish equivalence, and it is often not bi-directional because one model is "smarter" or "richer" than anoteher

And even when you can, it may not be possible to automate the transformation

Metadata Encoding and Transmission Standard (METS)

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

Developed by the Digital Library Federation as an implementation strategy for preservation metadata (needed to periodically refresh and migrate the data,)

Specifies an XML syntax for packaging metadata adhering to different standards as parts in a container and associating it with the same object

METS doesn't address the problem that the metadata standards are different; it just defines a standard way to package a set of them

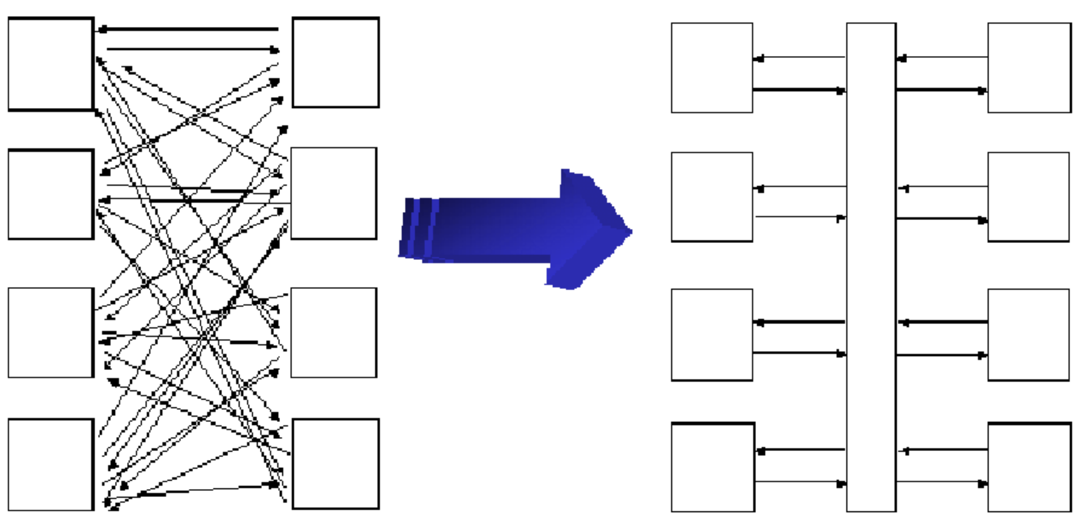
Crosswalks

A transformation that re-encodes, renames, rearranges, or restructures information from one metadata standard to another is sometimes called a CROSSWALK

First you need to establish the conceptual equivalence of information in the source and target models

It is sometimes useful to define equivalences for subsets or profiles of different metadata models and settle for a partial crosswalk

Interchange Formats



Ideally, any two metadata standards could interoperate by transforming them into a common interchange format

Doctorow on Metadata

People lie

People are lazy

People are stupid

People delude themselves

Metadata metrics distort it

Metadata suffers from "the vocabulary problem"

Readings for INFO 202 Lecture #7

Svenonius Chapter 6, Chapter 8 (127-132)

George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais, "The Vocabulary Problem in Human-System Communication"

Karen Coyle, "Identifiers: Unique, Persistent, Global"

L. Karl Branting, "Name Matching in Law Enforcement and Counter-Terrorism"