

5. Metadata and Metadata Standards

IS 202 - 13 September 2005

Copyright © 2005 Robert J. Glushko

Plan for IO & IR Lecture #5

- What is metadata?
 - Types of metadata
 - Metadata standards
 - Metadata incompatibility and interoperability
-

Plan for lectures 5-13 - Defining What Something Means

- Metadata [Lecture 5 - today]
 - Controlled names and controlled vocabularies [Lecture 6]
 - Classification [Lecture 7]
 - Formal assertions and ontologies [Lectures 8 and 9]
 - Metadata for multimedia [Lecture 10]
 - Models of documents (document types) [Lecture 11]
 - Models of data collections (database schemas) [Lectures 12-13]
-

What Is Metadata?

- Literally "data about data"
 - "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"
 - Find
 - Identify
 - Select
 - Obtain
-

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
-

An Expanded Definition

- 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.pdf)
- Metadata is that set of descriptive properties which serves one of more of the following functions:

- Uniquely characterizes an entity and for which values associated with the descriptive properties allow a user (human or machine) to discriminate between one entity and another
 - 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
 - Contains 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 machines 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
-

Types of Metadata

- 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"

- 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
-

Metadata for Versions and Editions

- 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
-

Metadata for Finite vs Continuing Resources

- Traditional library distinction between monographs and serials
 - MONOGRAPH is something that is complete, finished
 - SERIAL is resource that is expected to have more information added to it (like a journal where we expect regular issues)
 - But with digital information this distinction gets muddier
-

Administrative Metadata

- Location information
 - Acquisition information
 - Preservation metadata
 - Ownership, rights, permission, reproduction information
 - Usage information
-

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
-

Metadata Location

- Metadata with the object:
 - In the "header"
 - In the "body" as one of the components of the object
 - Metadata separate from the object it describes
 - Metadata repositories
-

Levels of Metadata, or Whose 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
-

Bibliographic Relationships - Tillet's Taxonomy [1]

- 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
-

Bibliographic Relationships - Tillet'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.

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
- And there are some Meta-metadata standards - standards for how to define metadata standards

The Same Item in Different Metadata Models

- MARC (MACHine-Readable Catalog) Record
- International Standard Bibliographic Description (ISBD)
- RFC 1807
- Text Encoding Initiative (TEI) Header
- Dublin Core

MARC Record

```

• 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   PSI:0   ILC:a   II:1
MMD:     OR:     POL:   DM:     RR:     COL:     EML:     GEN: BSE:
010      9124851
020      0872878112 (cloth)>
020      0872879674 (paper)
040      DLC$DLC$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 cataloging rules.
700 10   Taylor, Arlene G.,$d1941-

```

ISBD Syntax

- Title Proper (GMD) = Parallel title : other title info / First statement of responsibility ; others. -- Edition information. -- Material. -- Place of Publication : Publisher Name, Date. -- Material designation and extent ; Dimensions of item. -- (Title of Series / Statement of responsibility). -- Notes. -- Standard numbers: terms of availability (qualifications).

ISBD Instance

- Introduction to cataloging and classification / Bohdan S. Wynar. -- 8th ed. / Arlene G. Taylor. -- Englewood, Colo. : Libraries Unlimited, 1992. -- (Library science text series).

RFC 1807

- BIB-VERSION:: CS-TR-v2.1
ID:: UCB/123456
ENTRY:: September 9, 1997
TYPE:: BOOK
TITLE:: Introduction to cataloging and classification
AUTHOR:: Wynar, Bohdan S.
AUTHOR:: Taylor, Arlene G.
DATE:: 1992
PAGES:: 633
COPYRIGHT:: Libraries Unlimited, 1992

SERIES:: Library Science Text Series
 END:: UCB//123456

TEI Header (Minimal)

```

• <teiHeader>
  <fileDesc>
    <titleStmt>
      <title> Introduction to cataloging and classification</title>
      <respStmt><name>Bohdan S. Wynar<resp> 8th edition by</resp>
        <name>Arlene G. Taylor</name>
      </respStmt>
    </titleStmt>
    <publicationStmt>
      <distributor>Libraries Unlimited</distributor>
    </publicationStmt>
    <sourceDesc>
      <bibl> Introduction to cataloging and classification / Bohdan S. Wynar. -- 8th ed. / Arlene G. Taylor. -- Englewood, Colo. : Libraries Unlin
      </bibl>
    </sourceDesc>
  </fileDesc>
</teiHeader>

```

Dublin Core

```

• <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>
  ...

```

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

Achieving Metadata Interoperability

- "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

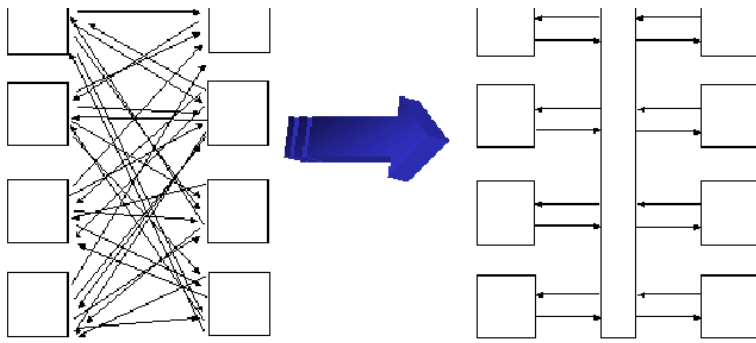
Metadata Encoding and Transmission Standard (METS)

- <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
- If this equivalence can be established, converting one implementation to another is a necessary but often trivial thing to do
- But isn't always possible to establish equivalence, and even when you can, it may not be possible to automate the transformation

Interchange Formats



- Ideally, any two metadata standards could interoperate by transforming them into a common interchange format
- This would reduce the $N \times N$ requirement for crosswalks from any model to another to the simpler $2 \times N$ task of transforming each to and from the interchange format

Readings for IO & IR Lecture #6

- Taylor Chapter 9 (241-247), Chapter 10 (261-282)
 - Brian Farish, What's in a name?
 - Karl Fast, Fred Liese, and Mike Steckel. What is a controlled vocabulary?
 - Karl Fast, Fred Liese, and Mike Steckel. Creating a controlled vocabulary.
 - Karl Fast, Fred Liese, and Mike Steckel. Synonym rings and authority files.
 - Karl Fast, Fred Liese, and Mike Steckel. Controlled vocabularies: A glosso-thesaurus.
 - Rick Jelliffe, W3C XML Schema Datatypes Reference
-