

6. Metadata and Metadata Standards [2]

IS 202 - 14 September 2006

Bob Glushko

Plan for IO & IR Lecture #6

The same item using different metadata models and syntaxes

MARC Record

Dublin Core

Metadata incompatibility and interoperability

"Metadata" {and,or,vs} "Vocabulary"

Cory Doctorow's Rant

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

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-

```

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>
    <istributor>Libraries Unlimited</istributor>
  </publicationStmt>
  <sourceDesc>
    <bibl> Introduction to cataloging and classification /
Bohdan S. Wynar. -- 8th ed. / Arlene G. Taylor. -- Englewood, Colo.
: Libraries Unlimited, 1992.
    </bibl>
  </sourceDesc>
</fileDesc>
<teiHeader>
```

Dublin Core

Proposed in 1995 as a standard set of metadata elements, simple enough to 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

FORMAT -- the physical or digital manifestation 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

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 another

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

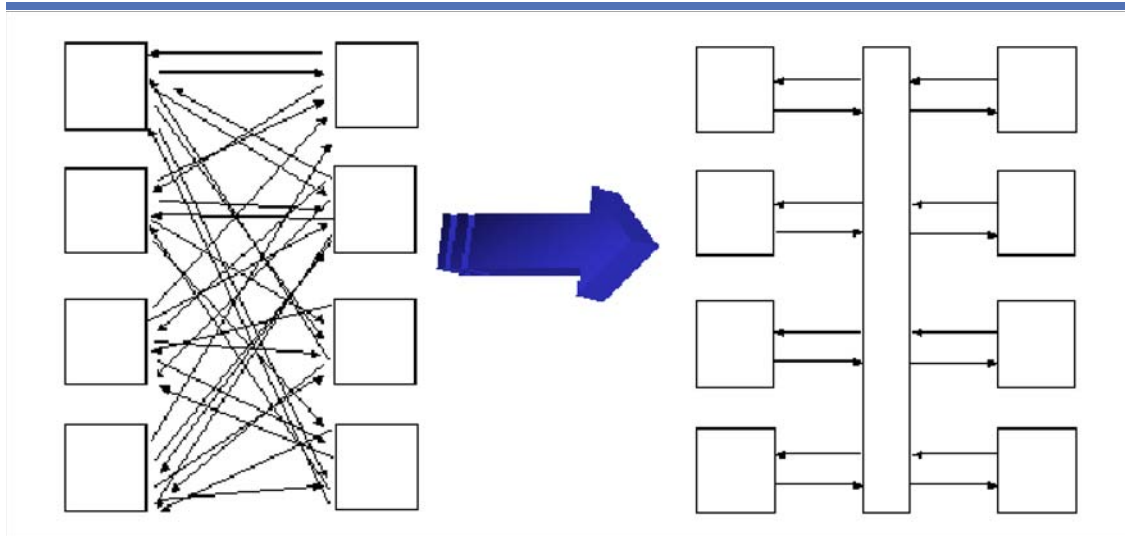
Crosswalks

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

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

"Metadata" {and, or, vs} "Vocabulary"

"Metadata" usually means description information about some content or entity

- Often general-purpose or "horizontal"

"Vocabulary" means the set of terms needed to encode the semantics in some content domain

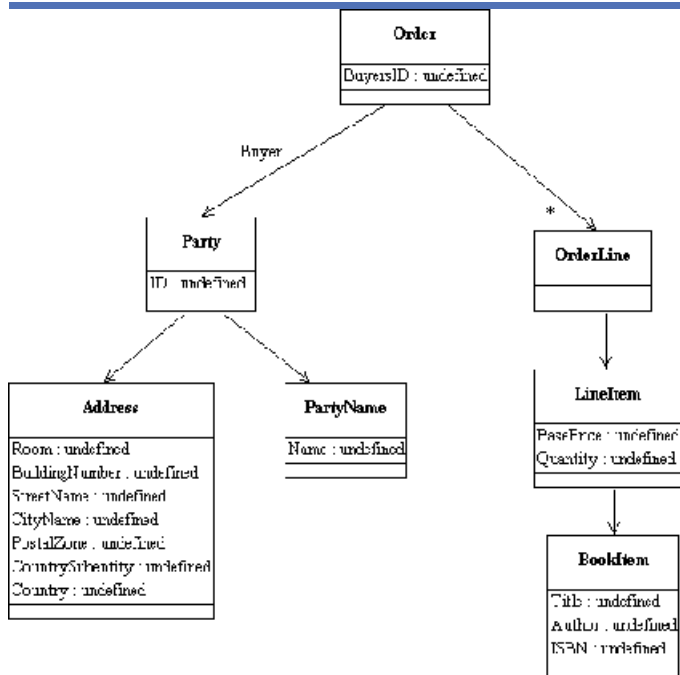
- Can be horizontal, but often "domain-specific" or "vertical"
- A "document type" model is defined by its "vocabulary"

Distinction not always clear or important; both metadata and vocabularies are MODELS of what they describe

Interoperability, crosswalks, interchange hubs, etc concepts apply to both metadata and vocabularies

"Vocabulary" Interoperability

Example -- The Target Model for an Order



The XSD Schema for the Expected Order [1]

```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified">

  <xs:element name="Order" type="OrderType"/>
  <xs:complexType name="OrderType">
    <xs:sequence>
      <xs:element name="BuyersID" type="xs:string"/>
      <xs:element name="BuyerParty" type="PartyType"/>
      <xs:element name="OrderLine" type="OrderLineType"
        maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="PartyType">
    <xs:sequence>
      <xs:element name="ID" type="xs:string"/>
      <xs:element name="PartyName" type="PartyNameType"/>
      <xs:element name="Address" type="AddressType"/>
    </xs:sequence>
  </xs:complexType>

  <xs:complexType name="PartyNameType">
    <xs:sequence>
      <xs:element name="Name" type="xs:string" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
```

The XSD Schema for the Expected Order [2]

```
<xs:complexType name="AddressType">
  <xs:sequence>
    <xs:element name="Room" type="xs:string"/>
    <xs:element name="BuildingNumber" type="xs:string"/>
    <xs:element name="StreetName" type="xs:string"/>
    <xs:element name="CityName" type="xs:string"/>
    <xs:element name="PostalZone" type="xs:string"/>
    <xs:element name="CountrySubentity" type="xs:string"/>
    <xs:element name="Country" type="xs:string"/>
  </xs:sequence>
</xs:complexType>

<xs:complexType name="OrderLineType">
  <xs:sequence>
    <xs:element name="LineItem" type="LineItemType"/>
  </xs:sequence>
</xs:complexType>

<xs:complexType name="LineItemType">
  <xs:sequence>
    <xs:element name="BookItem" type="BookItemType"/>
    <xs:element name="BasePrice" type="xs:decimal"/>
    <xs:element name="Quantity" type="xs:int"/>
  </xs:sequence>
</xs:complexType>

<xs:complexType name="BookItemType">
  <xs:sequence>
    <xs:element name="Title" type="xs:string"/>
    <xs:element name="Author" type="xs:string"/>
    <xs:element name="ISBN" type="xs:string"/>
  </xs:sequence>
</xs:complexType>
</xs:schema>
```

The Expected Instance

```
<Order>
  <BuyersID>91604</BuyersID>
  <BuyerParty>
    <ID>KEEN</ID>
    <PartyName>
      <Name>Maynard James Keenan</Name>
    </PartyName>
    <Address>
      <Room>505</Room>
      <BuildingNumber>11271</BuildingNumber>
      <StreetName>Ventura Blvd.</StreetName>
      <CityName>Studio City</CityName>
      <PostalZone>91604</PostalZone>
      <CountrySubentity>California</CountrySubentity>
      <Country>USA</Country>
    </Address>
  </BuyerParty>
  <OrderLine>
    <LineItem>
      <BookItem>
        <Title>Foucault's Pendulum</Title>
        <Author>Umberto Eco</Author>
        <ISBN>0345368754</ISBN>
      </BookItem>
      <BasePrice>7.99</BasePrice>
      <Quantity>1</Quantity>
    </LineItem>
  </OrderLine>
</Order>
```

Identical Model with Different Tag Names [1]

```
<Customer>
<Number>KEEN</Number>
<Name>
  <BusinessName>Maynard James Keenan</BusinessName>
</Name>

<Location>
  <Unit>505</Unit>
  <StreetNumber>11271</StreetNumber>
  <Street>Ventura Blvd.</Street>
  <City>Studio City</City>
  <ZipCode>91604</ZipCode>
  <State>California</State>
  <Country>USA</Country>
</Location>
</Customer>
```

Identical Model with Different Tag Names [2]

```
<Acheteur>
<ID>KEEN</ID>
<Nom>
  <NomCommercial>Maynard James Keenan</NomCommercial>
</Nom>
<Adresse>
  <Appartement>505</Appartement>
  <Bâtiment>11271</Bâtiment>
  <Rue>Ventura Blvd.</Rue>
  <Ville>Studio City</Ville>
  <CodePostal>91604</CodePostal>
  <Etat>California</Etat>
  <Pays>USA</Pays>
</Adresse>
</Acheteur>
```

Same Model, Attributes Instead of Elements

```
<BuyerParty
  ID="KEEN"
  Name="Maynard James Keenan"
  Room="505" BuildingNumber="11271"
  StreetName="Ventura Blvd."
  City="Studio City"
  State="California"
  PostalCode="91604"
>
```

Granularity Conflicts

```
<Address>
  <StreetAddress>11271 Ventura Blvd. #505</StreetAddress>
  <City>Studio City 91604</City>
  <CountrySubentity>California</CountrySubentity>
  <Country>USA</Country>
</Address>

<PartyName>
  <FamilyName>Keenan</FamilyName>
  <MiddleName>James</MiddleName>
  <FirstName>Maynard</FirstName>
</PartyName>
```

Assembly Mismatch - Separate Customer and Order Documents [1]

```
<BuyerParty>
<ID>KEEN</ID>
<PartyName>
  <Name>Maynard James Keenan</Name>
</PartyName>
<Address>
  <Room>505</Room>
  <BuildingNumber>11271</BuildingNumber>
  <StreetName>Ventura Blvd.</StreetName>
  <CityName>Studio City</CityName>
  <PostalZone>91604</PostalZone>
  <CountrySubentity>California</CountrySubentity>
  <Country>USA</Country>
</Address>
</BuyerParty>
```

Assembly Mismatch - Separate Customer and Order Documents [2]

```
<Order>
<BuyersID>91604</BuyersID>
<BuyerParty>
  <ID>KEEN</ID>
</BuyerParty>
<OrderLine>
<LineItem>
  <BookItem>
    <Title>Foucault's Pendulum</Title>
    <Author>Umberto Eco</Author>
    <ISBN>0345368754</ISBN>
  </BookItem>
  <BasePrice>7.99</BasePrice>
  <Quantity>1</Quantity>
</LineItem>
</OrderLine>
</Order>
```

Conceptual Incompatibility

```
<Address>  
  <Latitude direction="N">37.871</Latitude>  
  <Longitude direction="W">-122.271</Longitude>  
</Address>
```

The "Not So Fast" Cases that Might Even Validate

The names are the same but the semantics aren't

```
<BuyerParty>  
<ID>555-22-1234</ID>
```

Validation Does Not Imply Interoperability

Suppose the document validates against the recipient's schema

- The semantics can still be different in important ways (the ID SSN example) – the strongest level of validation can fall short of establishing that the "same tags" have exactly the "same meaning" to the sender and recipient
- Furthermore, the recipient may not be able to validate all of the business rules that are important
- This is a good argument for industry standards / reference models / in your conceptual models or using XML vocabularies that represent them in authoritative ways

Doctorow on Metadata

People lie

People are lazy

People are stupid

People delude themselves

Metadata metrics distort it

Metadata suffers from "the vocabulary problem"

Graded Assignment 1: 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 at least 10 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 Tuesday 19 September before class

Readings for IO & IR Lecture #7

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, Chapter 12 (399-406)

<http://www.sims.berkeley.edu/~glushko/DocumentEngineering>