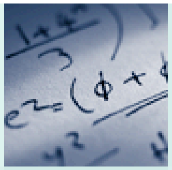
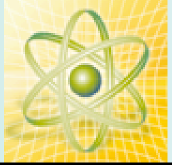
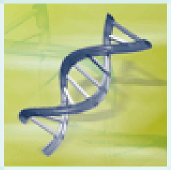


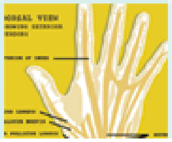


NSDL

THE NATIONAL SCIENCE DIGITAL LIBRARY

Developing Vocabularies for the NSDL





Vocabulary Issues in the NSDL

- ***No Controlled vocabularies are used.*** Many metadata implementations do not use controlled vocabularies in generating metadata values
- ***Vocabularies used are not identified.*** Even when controlled vocabularies are used, they may not be identified in instance metadata
- ***Vocabularies are not publicly accessible.*** Even where the vocabulary is identified in instance metadata, public access to the vocabulary by humans and/or machines is unavailable

Definition

- Controlled vocabulary (CV): A finite set of distinct values for a metadata property
    - Different from a “metadata vocabulary” which defines a finite set of properties (i.e., *a format or schema*)
    - In the metadata statement “dc:subject=cybernetics”, we are concerned only with the controls placed on the right-hand side of the statement (i.e.,the *scheme* or value space)

Vocabularies are not just for Subject!

- With the exception of properties with uncontrolled value strings in Dublin Core (e.g., dc:description), all properties can successfully use controlled vocabularies to increase precision and enhance meaning
  - E.g., the DCMIType vocabulary for use with the dc:type element is a “controlled vocabulary”

So far …

- The NSDL has recommended use of Qualified Dublin Core to allow exposure of standard controlled vocabularies already in use
- Problem: For audience (mediator and educationLevel) there have been no accepted standard vocabularies in general use
- Solution: Develop and recommend specific vocabularies

Workshop focus:

- Audience (including Mediator and Education Level)
- Resource Types:
  - Educational materials (at a more granular level than DCMIType)
  - Pedagogy*
- Interactivity Level

**NOTE: New element “instructionalMethod” approved by the DC Usage Board in 2004*

Workshop Outcomes

- Recommended strategy for developing standard controlled vocabularies for NSDL
- Increased support and guidance for declaring specific vocabularies within instance data for distribution within NSDL
- To Come: Guidelines for creation, management and exposure of local vocabularies used by NSDL projects

Determining Future Strategy

Property	Priority Ranking (1=high; 5=low)	Difficulty to Create
<<audience>>	3	High
<<educationLevel>>	1	Relatively Easy
<<interactivityLevel>>	5	Medium+
<<mediator>>	3+	Medium
<<resourceType>>	2	Medium
<<pedagogy>>	4	High

Creating a New Controlled Vocabulary

- Construct Vocabulary
  - Identify terms
    - Pull together synonyms, disambiguate homographs
    - Any term that is not “official” can be used as an aid in search (expanding the query to direct it to the “official” term
  - Identify relationships between terms
    - What are the relationships between terms that will aid the user during search and retrieval?
    - Make those relationships explicit in your metadata.

These actions make a controlled vocabulary “controlled”

NSDL Education Level Vocabulary

- Three levels of hierarchy
  - Supports Collection level expression as well as specific grade levels
  - Terms chosen from a range of projects, primarily ed.gov, the Dept. of Education website
  - Will include references from other known vocabularies, to allow crosswalking

Goals

- Develop EdLevel vocabulary as model “webized” controlled vocabulary
- Crosswalk all education level terms in the Metadata Repository
- Expose both incoming terms and crosswalked standard terms to services harvesting from the NSDL MR
- EdLevel vocabulary already available within new NSDL Resource Recommendation System

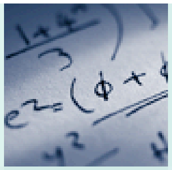
Webized Controlled Vocabularies

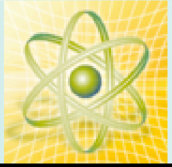
"Webized" controlled vocabularies *and* vocabulary terms are:

- Persistently and uniquely **identified**
  - URIs for names/tokens/identifiers
- Formally **declared** by means of a schema language
  - Represented in XML or RDF/XML
- Made Web-available by being **published**
  - Published through a Web-accessible registry

Continuing efforts

- Develop technical infrastructure for NSDL "Webized" vocabularies
  - Registry
  - Crosswalking capabilities
- Pursue development of additional vocabularies, as per the Workshop priorities
- Develop stronger community support and consensus around vocabulary creation and use

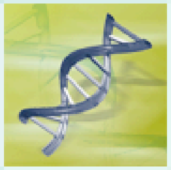




NSDL

THE NATIONAL SCIENCE DIGITAL LIBRARY

Perspectives on NSDL Vocabulary Development



The NSDL Vocabulary Workshop

- Interoperability needs
  - Controlled Vocabularies
  - Granularity
  - Applicability
- Small group process
  - Focusing questions
  - Multiple perspectives

ENC's Involvement In Vocabulary Development

- Math and Science Subject Vocabularies
- NSDL Middle School Pathway at ENC
  - http://nsdl.enc.org/
  - Augment Learning Resource Type (LRT), Educational Level, and Subject metadata

Choosing Learning Resource Type (LRT) Vocabulary

- Reviewed existing LRT vocabularies
- Considerations
  - Terms need a learning aspect
  - Terms need to be distinct from the media types
  - Terms need to be useful in education digital libraries

Early Concerns

- Applicability to digital libraries with different audiences
- Terms that are close to Media Type (video)
- Missing terms (tool)
- Unused terms

Early Concerns (cont.)

- How to define the terms clearly?
  - Words get in the way
  - Keep in mind the aboutness as opposed to the description (Demonstration)
  - Make clear distinctions between terms (Forum or Discussion, Message Board, and Weblog)
  - Make sure that progressions are clearly defined (Lessons and Activities → Project → Course → Curriculum)

Early Concerns (cont.)

- How many terms do we apply to each resource?
  - Make sure relationships are enduring (Nonfiction book, Article, and Reference)
  - Consider related terms, broader terms, and narrower terms

First Attempt Available for Review

- LRT vocabulary can be reviewed at http://metamanagement.comm.nsdl.org/Learning_Resource_Type.html
- Metamanagement-vocabularies mailing list
  - Metamanagement-vocabularies@comm.nsdl.org
  - http://comm.nsdl.org/mailman/listinfo/metamanagement-vocabularies
- Contact Judy Ridgway at jridgway@enc.org

Using, Choosing, and Creating Vocabularies in the NSDL Context

NSDL Annual Meeting

November 16, 2004

Discipline Specific Subject Vocabularies

Darin Burleigh
Journal of Chemical Eduction Digital Library

Introduction

Controlled vocabularies (CVs) are an important part of a digital library.

> "An essential part of any metadata plan is whether to use controlled vocabularies, and, if so, which one(s). Using controlled language terminology ensures more consistent description and better retrieval results. "
>
> NSDL Metadata Primer
> http://metamanagement.comm.nsdlib.org/creating2.html#thesauri;
> accessed 2004-07-14

CVs are important for interoperability.

> "Have you selected values from enumerated lists recommended to assist in cross-domain searching? If not, please recognize that interoperability will be degraded and records will be harder

> to maintain."
>
> > Guide to Best Practice: Dublin Core Version 1.1
> > Consortium for the Computer Interchange of Museum Information (CIMI)
> > http://www.cimi.org/public_docs/meta_bestprac_v1_1_210400.pdf
> > ;accessed 10/24/2004

> "Analysis of these heterogeneous collections indicates that controlled vocabularies and values are widely used in most repositories. Usage is extremely variable, however.
>
> ... The lack of interoperability is one of the significant problems facing digital libraries."
>
> > [Liu2002]

How does one choose a Subject vocabulary?

Goals

Promote discussion of semantic interoperability of metadata as it applies to NSDL.

Generate tips, hints, guidelines, recommendations for digital library projects.

Interoperability

What is it? Miller [Miller2000] identifies the following 'flavours' of interoperability:

> Technical, Semantic, Political/ Human,Inter-community, Legal, International

NSDL on interoperability [Arms2002]

> "The goal of interoperability is to build coherent services for users, from components that are technically different and managed by different organizations. This requires agreements to cooperate at three levels: technical, content and organizational.
>
> Content agreements cover the data and metadata, and include semantic agreements on the interpretation of the information.
>
> In 1998 Sarantos Kapidakis suggested that interoperability could be analyzed by comparing cost against functionality [Kapidakis 1998]. The following model, based on that suggestion, was first described in [Arms 1999]. "

Challenges for Subject Element

Variability is a challenge to interoperability.

While established or proposed vocabularies for some metadata elements already exists, and are widely applicable (e.g. Resource Type, Audience, Format), Subject element is by its nature specific to a given discipline.

The problem is inherent in the semantics, not the scheme (same problem in MARC, METS LOM...)

Also depends on depth, breadth, and granularity of digital library.

- More terms -> greater specificity in retrieval
- Fewer terms -> lower maintenance costs

Specificity of search: Physical chemistry > Kinetics > Rate law

Challenging questions:

- How do I find an appropriate Subject vocabulary for my digital library?
- How do I create one if it does not exist?
- What are the criteria to think about that will ensure interoperability?
- What are the options for relating terms in different vocabularies?
- Should we worry about it?

Examples from JCE DLib

Limitations on size of list

- Managable by editors: Metadata integrated with workflow.
- Managable by authors: we rely on them to supply primary metadata

Revised existing keyword list.

- Gold standard for chemistry: Chemical Abstracts - Too large, possibly encumbered by copyright
- LCSH: too general

- Selected terms from textbook chapters (and sub-chapters)
- Expanded coverage of Organic and Biochemistry
- Terms managed with web-based thesaurus
- Rated by JCE Reviewers

Terms with definitions: http://jce.divched.org/Journal/Authors/keywords.html
Local link

Interoperability: Chemistry, "the central science", overlaps with Physics, Biochemistry, Earth science, and Mathematics.

Mapping is costly and imperfect: e.g. DLESE -> JCE

Solutions

Wake and Nicholson [Wake2000] note that

> In 1999 Péter Jascó [Jasco1999] wrote that "savvy searchers" are asking for direction. Three years later the scenario he describes, that of searchers cross-searching databases where the subject vocabulary used in each case is different, still rings true.

They identified 5 options for the High-Level Thesaurus Project:

1. Do nothing:
   Artificial Intelligence will solve it in time.
   Big business -- Microsoft or similar -- will solve it.
   It is not important.
   No solution is necessary.
   The problem cannot be solved.
2. Set up a human process intended to lead to a solution in time.
3. Adopt a base-level, gradual approach, with an eye on future

developments.
4. Adopt a single scheme.
5. Mapping service alternatives.

Two strategies are discussed in the literature

Merging Thesauri

Two problems have turned out to be the most difficult:

First, differences in term semantics, semantics of hierarchical relations and term overlap can render the simple combination of concepts from two sources impossible. Resolutions to this problem are usually semiautomatic, see e.g. Constantopoulos and Sintichakis (1997), and can become fairly expensive.

Second, controlled vocabularies are often associated with a large installed base of systems using them, such that migration to a new set of terminology and relations may be virtually impossible; for example, with subject headings used by national libraries (Chan 2000, Landry 2000)

[Doerr2001]

A more fundamental approach involves ontologies, RDF and the Semantic Web

The number of metadata vocabularies will continue to grow as individual communities seek to structure their own information for their own purposes;

Attempts to develop universal metadata vocabularies are misdirected, since “spoken” languages (those used by communities to actively describe content) will inevitably diverge (history is replete with failures to find common spoken languages [25]);

A more useful effort is to attempt to formulate a base “understanding”

> or “processing” language, a core ontology, incorporating basic entities and relationships common across the diverse metadata vocabularies;
>
> Such a core ontology might then be useful for a number of purposes including integrating information from heterogeneous vocabularies and providing base concepts that future metadata initiatives could build on when developing domain specific vocabularies.
>
> [Doerr2002]

Discussion: What should NSDL do?

Top-down approach:

Provide standarized vocabulary(s).

Develop meta-thesaurus.

Develop a core ontology.

Bottom-up approach:

Provide more specific guidelines for choosing a vocabulary - workshop for next year?

Recommend software tools.

Other issues?

Other solutions?

References

[Arms1999] Digital Libraries William Y. Arms *p. 208. MIT Press (1999)* .

[Arms2002] A Spectrum of Interoperability - The Site for Science Prototype for the NSDL William Y. Arms, Diane Hillmann, Carl Lagoze, Dean Kraff, Richard Marisa,John Saylor,Carol Terrizzi, Herbert Van de Sompel *D-Lib Magazine, January 2002, Volume 8 Number 1* http://www.dlib.org/dlib/january02/arms/01arms.html

[Doerr2001] Semantic Problems of Thesaurus Mapping Martin Doerr *Journal of Digital Information, Volume 1 Issue 8 Article No. 52, 2001-03-26* http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/

[Doerr2002] Towards a Core Ontology for Information Integration , Martin Doerr, Jane Hunter, Carl Lagoze *Journal of Digital Information, Volume 4 Issue 1 Article No. 169, 2003-04-09 (2002)* http://jodi.ecs.soton.ac.uk/Articles/v04/i01/Doerr/

[W3CSW] W3C Semantic Web http://www.w3.org/2001/sw/

[Miller2000] Interoperability - What is it and Why should I want it? Paul Miller *Ariadne, 24 June (2000)* http://www.ariadne.ac.uk/issue24/interoperability/intro.html

[Wake2000] HILT - high-level thesaurus project: building consensus for interoperable subject access across communities , S. Wake, D. Nicholson, *D-Lib Magazine vol.7 issue 9(2000),* http://www.dlib.org/dlib/september01/wake/09wake.html

[Jasco1999] Savvy Searchers Do Ask For Direction Péter Jascó, *Online and CD-ROM Review, 23 (2), pp 99-102 (1999).*

[Liu2002] Federated searching interface techniques for heterogeneous OAI repositories Liu, X., Maly, K., Zubair, M., Hong,

Q., Nelson, M.L., Knudson, F., Holtkamp, I., *Journal of Digital Information, 2, 4, (2002),* http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/.

Journal of Chemical Education

Digital Library

> The *Journal of Chemical Education* is the journal of the Division of Chemical Education, Inc. of the American Chemical Society. Published continuously since 1924, *JCE* is the world's premier chemical education journal. Our mission is to help chemistry teachers stay current with research advances as well as share new ideas in teaching methodologies and course organization. A multimedia publisher, *JCE* welcomes materials in print, software, video, and other digital formats.

For more details see About JCE