SCHOLARLY TECHNOLOGY GROUP

Retrieval of Education Materials on the World Wide Web: Overview of Projects, Issues, and Directions--1998

Retrieval of Education Materials on the World Wide Web: Overview of Projects, Issues, and Directions--1998

Introduction

The US Department of Education supports numerous public access Web sites that provide

information to educators, administrators, policy makers, and students. The WALDO project is one of several responses to an initiative calling for the provision of "one stop" access to materials on any of these sites. This report, part of Brown University's (Scholarly Technology Group) component of WALDO, reviews the issues involved in information retrieval from the Web, with particular reference to cataloging resources for the education community. Recommendations specify ways to exploit features of existing metadata sets and authority lists to accommodate the full range of educational materials on the web.

Approaches to retrieval of electronic information

The growing body of educational material on the world wide web cannot be located efficiently at present by potential users [HRST], [DC1]. Two approaches to this problem are currently being pursued: (1) searching content and (2) searching specific bibliographical information or "metadata." In the first approach, query terms are matched to terms in indexes created by a program from target documents. Documents whose terms match the query terms are presented as items of possible interest to the searcher. Such automatic indexing programs have been implemented by the Office of Educational Research and Improvement's Cross-Site Indexing project and by the Northeast and Islands Regional Educational Lab [CSI], [LAB]. This approach typically presents the searcher with far too many items, since a randomly selected word in a document is unlikely to provide much information about its focus of interest. Many relevant documents may not be recovered at all. Statistical methods can be used to make better use of vocabulary information ([SAL], [IR]), but in the absence of information that distinguishes terms of the title, author, subject, or other crucial roles from the uses of words in general, such methods are limited.

It seems necessary, therefore, to pursue the second approach: to characterize the contents of web documents by means of an electronic cataloging procedure. The bibliographic information about a document required for efficient retrieval is known as "metadata." Metadata records contain information about a resource in a structured format as values of labeled attributes. This information is much like that in existing catalogs such as the ERIC database [ERIC] or ENC's Resource Finder [ENC]. Metadata from a variety of sites can be harvested, indexed, and then searched in response to a request for information that refers to specific attributes or identifies values that are of interest. All and only the materials whose records are appropriate matches to the query are returned.

Use of metadata for more efficient retrieval

A variety of metadata formats designed for similar purposes (MARC records, the GILS system, TEI headers, IAFA templates) are being used for bibliographic control and other information management projects (see [H&D] and [RUS] for a review and evaluation of these systems). But many of these formats cannot provide useful models for retrieval of educational materials because they need to be implemented by a staff of trained catalogers and are not suited for the huge number of documents on the web. Such formats are designed for substantial information management tasks not pertinent to retrieval of the educational materials with which this report is concerned. One metadata model is significantly more useful, largely because it is simple enough to be used by non-professional catalogers. This is the Dublin Core, a basic set of 15 metadata elements needed for retrieval of web documents.

The Dublin Core as a developing standard

The original plan for the Dublin Core emerged from a workshop held in Dublin, Ohio, in March 1995. The goal of the meeting was to achieve consensus among librarians, scholars, and internet standards makers on "a list of metadata elements that would yield simple descriptions of data in a wide range of subject areas" [DC1]. Subsequent workshops proposed design principles that would keep the Core small and simple by specifying how extensions and modifications could be handled by individual implementers. All Dublin Core elements are optional and repeatable in principle, though they may be otherwise constrained by individual implementers.

A basic 'container' architecture was developed to provide for the representation of additional metadata element sets. This "Warwick Framework" [DC-2] underlies the evolving Resource Description Framework that will extend the metadata model to accommodate specialized information management requirements without altering the integrity of the basic Core.

The fourth metadata workshop, and subsequent discussions on the meta2 list, have developed the notion of qualifiers as a mechanism to refine the elements. The "Canberra Qualifiers" [DC-4] extend the basic elements by allowing them to be qualified in three ways:

- Scheme: the identification of the authority list or standard from which the value of an element is obtained (such as an indication that subject terms are selected from the ERIC thesaurus, or that date formats follow the ISO 8601 specification)
- Lang: indication of the language of the content of the metadata element (a document in Spanish would have a DC.Language value of "Spanish", but the Lang qualifier would make it possible to indicate that a set of Subject values was in English).
- Subelement: a mechanism for more narrowly specifying the meaning of a given element (for instance, to indicate whether the value of the Date element represents the date of creation of the document or the date that it was cataloged, etc.). Qualification with subelements makes it possible to characterize resources much more accurately, but it also introduces the possibility that the metadata records will become too complex to be created by non-specialists and too divergent to be interoperable. The needs of implementers varies and the model must work for both 'minimalists' concerned with ease of use and 'structuralists' concerned with accuracy of description.

One of the major outcomes of the fifth and most recent workshop (Helskinki, October 1997) was stabilization of the set of 15 Core elements. A plan was drawn up for documenting both Simple and Qualified Dublin Core in a series of Internet Drafts in order to move towards formal standardization and international acceptance. The first document [RFC1] has been submitted. It discusses the conventions for use of the basic Dublin Core elements, as they are described in the reference description [DC-REF], without qualifiers. These elements can be divided into three groups:

- Elements related mainly to the Content of the resource (Title, Subject, Description, Source, Language, Relation, Coverage)
- Elements related mainly to the resource when viewed as Intellectual Property (Creator, Publisher, Contributor, Rights)
- Elements related mainly to the Instantiation of the resource (Date, Type, Format, Identifier)

Four additional RFC documents will describe (a) conventions for embedding unqualified Dublin Core metadata in an HTML file, (b) recommended qualifiers and principles for qualifying DC elements, (c) conventions for embedding qualified metadata in an HTML

file, and (d) a description for encoding qualified DC metadata in compliance with the Resource Description Framework (RDF).

The combination of the basic simplicity of the Dublin Core with a capacity for extension through use of qualifiers or collection-specific elements has motivated implementation of DC metadata in a great variety of disciplines. Implementers must determine for themselves how to balance the need for richer bibliographic description against the need for efficient cataloging.

What are the resource discovery needs of the US Department of Education?

ED-sponsored Web sites of particular interest for the WALDO project are the Regional Technology in Education Consortia (RTECs), the Regional Educational Labs (for research on school improvement), and the Eisenhower National Clearinghouse for Math and Science. The materials for educators at these sites are primarily concerned with public schools (K-12), with some attention to programs for vocational training and adult literacy. They include:

- curricular materials (lesson plans, descriptions of activities, multimedia presentations)
- documents (research reports, guidelines, regulations, policies, professional development materials, news)
- databases of information about off-line books, articles, films, etc.,
- lists of links (internal links to components of a document or project; directories of external links to sites or resources on specific topics)
- software packages (tools for assessment, course development, etc.)

What cataloging systems for such materials are currently in use?

To identify the kinds of metadata needed for cross-site discovery of educational resources, it is useful to consider the kinds of cataloging systems that are familiar to educators. The ENC Resource finder represents a familiar system for classifying **curricular materials**; the ERIC database is the standard for classification of **documents** and **journals** dealing with education issues.

ENC Resource Finder: <http://www.enc.org/rf/index.htm>

The Eisenhower National Clearinghouse catalogs and evaluates curricular materials for mathematics and science education. Most resources cataloged are not in electronic form, and those that are available on the web cannot be directly accessed, in that a search would take the user to the ENC Resource Finder database rather than to the resource proper. Resource Finder records are created by expert catalogers and incorporate a rich classification system designed to facilitate ordering of materials and assessment of their suitability for a given educational purpose. In addition to the kinds of information represented in the Dublin Core set, these records contain fields for grade and audience, vendor, physical description, cost, product identification number, national or local standards met, evaluators' judgments, and extensive description of the conceptual substance of the resource, including abstract and table of contents. A decision needs to be

made as to which of these many fields should be represented as metadata for cross-site searching and which should remain as local information available once the resource has been selected. The most obviously useful extensions to the Dublin Core for materials targeted by the Resource Finder would be elements for **grade** and **audience**. An appropriate **subject** element for curricular resources would need to distinguish academic discipline (English, biology, etc.) from concepts or key terms that characterize the resource more narrowly (epic poetry, frogs, etc.).

ERIC database of books and articles: http://www.aspensys.com/eric/

The Education Resources Information Center has been abstracting and disseminating educational publications for forty years. Publications are cataloged by subject specialists at clearinghouses dedicated to specific topics of interest. The fields of the ERIC database provide a straightforward, minimal, and intuitive bibliographical description. They could be represented as Dublin Core metadata without significant extension of the element set. The ERIC activities requiring specialists for implementation are creation of abstracts or digests, assignment of subject descriptors from the ERIC thesaurus, and determination of publication type. Creation of abstracts and digests does not concern us here, as it is not directly relevant to metadata design. But even the simplest metadata system must confront the problems of selecting appropriate descriptive terminology. ERIC's thesaurus and list of publication types are much too extensive for use by nonspecialist catalogers, incorporating detail that might in another system be represented as discrete fields of the record, such as Audience or Instructional Method. Metadata developers must address the need for simpler subject and resource type vocabularies dedicated to web search requirements. This might be accomplished in part by associating a somewhat wider set of elements with smaller, more manageable vocabulary lists.

What element sets are used for metadata projects in the field of education?

As noted above, the Dublin Core has emerged as the dominant architecture for classification of web materials. The major education metadata projects analyzed below either were defined as Dublin Core projects from the outset or were adapted to the Dublin Core model in the course of development.

Education Network of Australia (EdNA)

< http://www.edna.edu.au/edna/owa/info.getpage/?sp=auto&pagecode=5210>

- Scope of project: EdNA is a collaboration between Australian states and territories and all sectors of education and training: schools, vocational education and training, adult community education, and higher education. Cataloging of education resources is decentralized, being undertaken by educational institutions themselves. Over 71,000 documents have been cataloged.
- Materials being cataloged: Participating primary and secondary schools have so far cataloged resources that are mostly 'external'--i.e., descriptions of educational components of government or corporate websites, such as NASA or the Smithsonian. Vocational and higher education participants have cataloged 'internal' curricular and training materials.
- Metadata elements: The EdNA metadata record is based on the Dublin Core but does not use the Relation, Source, or Contributor elements.
    1. Most additional elements are related to the management of the records

themselves (meta-metadata):
- Entered (date of creation of metadata record)
- Approve (approver of item for inclusion)
- Suggestor (suggestor of item)
- Reassessment (months until resource should be reassessed)
- Categories (classification system in directory of resources)

2. Two additional EdNA elements recognize the unique hyperlinked character of web resources and specify constraints on the harvesting of documents to index:
   - IndexLevel (number of levels of links to follow)
   - IndexSites (number of servers to access when following links)

3. The only elements added to the Dublin Core set that apply specifically to the educational character of the resources are:
   - Review (third-party review of the resource)
   - UserLevel (controlled vocabulary of user and school level)
     Comment: The EdNA metadata set is among the simplest in use for education resources. Subject terms are uncontrolled key words and the vocabulary of users is limited to 'students' and 'teachers'. It is not clear if this set will provide sufficient description for discovery of the full range of resources for the education community.

ARIADNE

<http://ariadne.unil.ch>

- Scope of project: Ariadne is part of the education and training program of the European Union Telematics Application Program. Its goal is to foster the share and reuse of electronic pedagogical material by universities and corporations.
- Materials being cataloged: The Knowledge Pool System is a database of reusable pedagogical materials and metadata records describing these materials. It is currently distributed at eight sites throughout Europe.
- Metadata elements: The Dublin Core elements and additional Ariadne elements are grouped into several categories:
  1. General information about the resource (Dublin Core Identifier, Title, Creator, Date, Language, Publisher, Source).
  2. Semantics of the resource. The DC.Subject element comprises four subelements, the academic discipline to be distinguished from the main concept, its synonyms, and other possible subject terms.
  3. Pedagogical attributes. These elements designate metadata specific to the educational character of Ariadne resources.
     - user type, either 'learner' or 'author'
     - document type, either 'expositive' (learning from instruction or study) or 'active' (learning by doing)
     - document format, from a list of values depending on document type
     - usage (optional) freetext comments on how to use the resource

     The following elements are applicable to the metadata record only if the value of **user** is 'learner', i.e., the resource is for a student rather than for use by a creator of educational resources
     - didactical context, values from a list of learning styles
     - course level, a pair of values: country and educational level as specified in that country (e.g., 'US, K-3')
     - difficulty level, 'low', 'medium', or 'high' for designated course level
     - interaction quality (for 'active' resource) semantic density (for expositive

resource), 'low', 'medium', or 'high'
- pedagogical duration, minutes needed by an average learner to use the resource

4. Technical characteristics (document handle for resource retrieval, format, file size, and installation information)
5. Conditions for use (DC.Rights element which can have the values 'free' or 'not free' and elements for price and acquisition of not free resources).
6. Meta-metadata (creator of the metadata, date, language, and revision information about the metadata record itself)

Comment: The Ariadne metadata record provides detailed features for cataloging materials targeted for students, including evaluation components whose values must be provided by experts. Other materials are not explicitly provided for: it is not clear how one would catalog information resources for teachers, administrators, education researchers, etc. who are neither 'learners' nor 'authors'.

Information Management System (Educom)

<<http://www.imsproject.org/metadata>>

- Scope of project: IMS is a cataloging system sponsored by software corporations, publishers, and institutions of higher learning developing education resources on the internet.
- Materials being cataloged: Education resources produced by sponsoring institutions are the priority materials, but the system is designed to be extensible for a wide variety of resources.
- Metadata elements: An important design feature of IMS metadata is the definition of distinct sets of metadata elements to characterize different kinds of resources. Apart from a small base set, there is no attempt to impose the same metadata on all resources. IMS uses 'containers' to define the set of metadata appropriate for a particular object (item to be cataloged).
  1. The Base Set of metadata contains the minimal elements for description of all resources cataloged in IMS. These are basic Dublin Core elements including title, publisher, date, description, format, identifier, and subject. The base set also includes 'meta-metadata' (author of the metadata, creation date, date of last modification, validator, and container type).
  2. There are two types of values for the Subject element in the Base Set: the **descriptor,** a term from a controlled vocabulary paired with an identifier naming the source of the the term (e.g., ERIC, LCSH); and **key word,** for uncontrolled terms that describe the subject, such as a proper nouns and new terminology.
  3. The container types that augment the Base Set are: **Item**, **Module**, and **Tool**. The Item container is used for a unitary resource such as a text or image. The Module container is for a learning resource with a specific educational value or purpose, such as a course, topic, assessment, assignment, or activity. The Tool container is for a learning resource that provides a function for the user, such as a word processor, calculator, statistical analysis package, or composition guide.
  4. The Module container employs the full set of IMS metadata elements; the Item and Tool containers employ subsets of this set. The Item set adds only 'author', 'price code' and 'rights' to the Base set, providing minimal elements for the description of education resources that are not well-defined curricular materials. The Tool set includes the elements

'user support' and 'platform'. The Module container includes a particularly detailed set of elements for educational methods and objectives, including objectives mandated by government agencies.

5. In this product, **learning level** is indicated by a pair of values which describe the academic grade and skill level for which the resource is appropriate. For example, 8-9:3 would be used to represent a resource appropriate for ages 8 to 9 with a difficulty level of 3 on a scale of
6. -5, 5 being most difficult.
7. The IMS **pedagogy** element corresponds to Ariadne's **document type.** It has two possible values: "expository" (= "expositive" in Ariadne) and "discovery" (= "active" in Ariadne).
8. The IMS **resource type** is a Dublin Core element, but the controlled vocabulary associated with this element is peculiar to IMS.
9. The IMS **use time** element, measured in minutes, corresponds to **pedagogical duration** in Ariadne.

Comment: IMS employment of containers with distinct sets of metadata elements for different kinds of resources offers the prospect of simplified cataloging to the extent that unnecessary elements can be excluded from a record structure. Some IMS values, such as the paired values for learning level, are not directly meaningful to a searcher and would require cataloging expertise to use properly. For more efficient searches, it seems desirable to map the searcher's familiar vocabulary (such as "grade level" in the US) to IMS's neutral values of age in years.

Gateway to Educational Materials (GEM)

<<http://gem.syr.edu/Workbench/index.html> >

- Scope of project: GEM is a US Department of Education-sponsored project which has developed a metadata structure and cataloging program for distributed cataloging of K-12 resources.
- Materials being cataloged: The project targeted lesson plans and other curricular materials as the highest priority for cataloging, as is reflected in the metadata element set.
- Metadata elements: GEM uses all the Dublin Core elements. The GEM controlled vocabulary for the Dublin Core Subject element is well developed for curricular materials, providing two levels of classification, one for general academic subject area, and a second for specific topic. The ERIC controlled vocabulary can be designated for non-curricular subject values. A subelement of the Subject element is available for uncontrolled key words. The Dublin Core is extended with a variety of elements for K-12 curricular resources:
    1. An **Audience** element added to the Dublin Core is further qualified to designate both the immediate user of the resource (Tool For) and the student population to be served (Beneficiary). Each of these subelements has its own GEM controlled vocabulary.
    2. A **Grade** element is qualified to designate the K-12 grade level of the beneficiary or, in a **Level** subelement, beneficiaries outside the K-12 range.
    3. A **Pedagogy** element has three subelements: (1) teaching (instructional method); (2) grouping (of students in classrooms); and (3) assessment.
    4. **Quality** and **Standard** elements are available to represent evaluations by outside agencies.
    5. A **Duration** element corresponds to **use time** in IMS and **pedagogical**

          duration in Ariadne.

6. An elaborate set of qualifiers for the **Relation** element is used to associate the many evaluative components of a curricular resource (e.g., isRevisionHistory, isContentRating, isPeer Review). Reference to these components in a full description of the resource may be valuable once a particular resource has been selected for consideration, but it is not clear what role these components would play in initial resource discovery.

Comment: Although the full set of GEM elements are designed primarily for curricular materials, it is possible to identify an appropriate subset that can be used to catalogue non-curricular materials, such as those targeted by WALDO. A new controlled vocabulary needs to be designed for each such element. The GEM architecture allows for addition of new controlled vocabularies by means of the **Scheme** qualifier. GEM comes closer to addressing the needs of non-specialist catalogers than do the other systems reviewed, with element names and definitions providing a fairly close match to the likely interests of searchers.

Metadata for a broad range of education resources

Materials on sites of concern to the WALDO project include both curricular and non-curricular resources. Curricular resources are, for the most part, provided through clearinghouses (such as ENC) or other organizations that create records for databases to support adoption decisions and inventory and ordering needs in addition to basic resource discovery. The components of these records that correspond to GEM's set of metadata elements can be mapped fairly directly to the GEM record without an additional cataloging procedure, though terminology for values of elements like Audience and Resource Type may not match the GEM scheme.

Non-curricular resources such as reports, training guidelines, directories and reference lists, interactive forums and listservs, etc., are not provided through agencies with cataloging capability and do not require extensive or detailed bibliographic description. Catalog records for such resources are critical primarily to improve access to the information they contain. The WALDO prototype of the element set for such resources includes the 15 Dublin Core elements, though only 7 are mandatory:

- Date (of metadata record creation)
- Identifier
- Format
- Language
- Resource Type
- Title
- Subject
- Publisher (Online Provider)

Values for other Dublin Core elements may not be readily obtainable from the resource "in hand" or may be less critical to initial discovery of relevant resources. These elements are optional, but should be used if the values are given in the resource being cataloged.

- Description (recommended)

Creator
- Coverage
- Relation
- Source
- Contributor
- Rights

Additional elements are proposed for a catalog record to support the discovery of relevant resources for the education community:

- Audience (mandatory)
- Grade/Education Level (optional)
- Essential Resource (optional)

The elements proposed for cataloging WALDO materials correspond to a subset of the elements developed by the GEM project. The two projects are concerned with materials addressing similar populations and educational issues. It seems appropriate to use a container architecture like that of the IMS product, making the elements of the WALDO prototype available in a module of GEM for non-curricular materials where the full set of GEM elements is available for curricular resources.

The two projects differ in the use of subelements to refine particular elements (especially Relation) and in the vocabularies associated with identification of Subject, Audience, and Resource Type. To the extent that these vocabularies describe different conceptual domains, the Scheme qualifier can be used to specify different vocabulary authorities. For example, a curricular resource would use the GEM vocabulary of K-12 academic disciplines. This information would be represented as follows (using the conventions for embedding in HTML documents with the META tag):

```
<META NAME="DC.Subject.Level1" SCHEME="GEM" CONTENT="arts">

<META NAME="DC.Subject.Level2" SCHEME="GEM" CONTENT="music">
```

A non-curricular resource dealing with funding for music instruction in elementary grades would be cataloged with subject terms selected from the ERIC thesaurus or another suitable authority list:

```
<META NAME="DC.Subject" SCHEME="ERIC" CONTENT="music education; fund
raising; philanthropic foundations; ...">
```

In other cases the differences in vocabulary may need to be resolved by seeking consensus among the broader community of users and producers of education resources. See discussion below.

Prospects for interoperability of education metadata projects

Most projects are substantially in agreement about the critical elements needed for education materials, such as 'academic discipline' and 'user level' (i.e., material for students must indicate grade or learning level, other materials may be targeted to teachers, administrators, parents, etc.). There are unnecessary differences from one project to another, however, in the terminology of element labels and in the refinement of elements with qualifiers. Other problematic differences involve smaller-scale elements for details such as price code, standards mapping, or

software requirements. Such elements of a full biblographic record may be useful once a relevant resource has been located but are not likely to play a crucial role in the initial search and make cataloging more difficult for nonspecialists. Much of this detail might be more useful in a separate database under an entry for a particular document, thus keeping the core scheme simple. Alternatively, the mechanism of separate "containers" could be employed to isolate metadata for simple resource discovery from elements of a fuller record.

It is likely that these projects will work towards greater interoperability as they participate in the ongoing development of the Dublin Core. The recent memorandum of understanding between IMS, Ariadne, and GEM is a particularly hopeful development. A viable search procedure for educational documents will best be created through consensus about how to derive the necessary components from a selection of the existing categories that have required so much effort and expert knowledge to devise.

Controlled vocabularies for education metadata projects

Controlled vocabularies are critical to any metadata system because diversity in catalogers' choices of terms reduces retrieval effectiveness. Serious discrepancies have already arisen among education metadata projects with respect to the lists of terms that specify values for a given metadata element. Although all projects have elements for resource type, subject, and audience, no two projects use the same list of terms for the values assigned to any one of these elements.

Vocabulary lists for the content values of resource types and subjects need further development. The Waldo prototype has explored the usefulness of the thesaurus of ERIC descriptors as a controlled vocabulary of subject terms, but much additional work will be necessary to facilitate subject identification by non-specialists. For example, interfaces for catalogers should be developed that make it easier to locate existing terms by linking synonyms to a relatively small number of standard vocabulary items. This procedure differs significantly from the one employed in the ERIC thesaurus, which presents catalogers with a multiplicity of narrow alternatives to a more general term.

Problems that arise in local development of controlled vocabularies are illustrated below by comparison of lists for resource types in the Dublin Core, EdNA, Ariadne, IMS and GEM.

The Dublin Core Resource Type working group has recommended a set of primary values for the Resource Type element. This set of terms classifies resources according to the nature of the medium embodying the resource and bears a close resemblance to the general material descriptions of standard library catalogs [AACR2]:

- text
- image
- sound
- data
- software
- interactive
- physical object

A draft 'structuralist version' of resource types subcategorizes these genres with a set of more specific terms in an effort to characterize the resource more fully. The subtypes of "text," for example, are:

- abstract
- advertisement
- article
- correspondence
- dictionary
- form
- homepage
- index
- manual
- manuscript (i.e., unpublished text not described elsewhere)
- minutes
- monograph
- pamphlet
- poem
- preprint
- proceedings
- promotion
- serial
- tech report
- thesis

Some of these terms are counterintuitive. The term "advertisement," for example, refers to announcements of job openings, while "promotion" refers to what American English speakers would think of as advertisements. There is no resource type "announcement," though this is a very common type of Web publication. The recommended procedure for including locally determined subtypes not provided in this list is to precede them with an 'x-', e.g., 'text.x-announcement'.

Developers of the major education metadata systems have each created an idiosyncratic vocabulary of resource types. Comparison of vocabularies for resource type reveals wide disparities in the interpretation of this element despite the fact that all four systems are designed for cataloging of education resources. Inspection of these vocabularies (presented below in alphabetical lists) will also make it clear that they bear little relation to the Dublin Core list.

EdNA

- course offering
- curriculum
- event
- forum
- individual (i.e., home page)
- links
- message
- organization - educational services
- organization - parent
- organization - professional
- project - curriculum
- project - collaborative
- project - research
- project - students

- project - teachers
- report
- school - primary (home page)
- school - secondary (home page)
- university (home page)

Ariadne (examples only, list of suggested terms not available)

- expositive (e.g., hypertext, video)
- active (e.g., exercise, questionnaire, simulation)

IMS

- advertisement
- assessment
- base (undifferentiated)
- collection
- dataset
- document
- example
- exercise
- media resource
- message
- miscellaneous
- reference
- schedule
- simulation
- tool
- tutorial

GEM

- activity
- artifact
- catalog record
- community (e.g., listservs, online forums)
- course
- curriculum
- data set
- environment
- form
- lesson plan
- primary source
- project
- realia
- reference
- research study
- service
- tool
- unit of instruction

Comment: It is not clear that the categories listed above would be useful to educators conducting searches on the web. Educators might well be more interested in categories that refer to the function of an item rather than to the medium or form

relevant for archival purposes. Familiar categories such as regulation, policy, guidelines, or news, and newer categories like directory (list of links) and listserv are essential. Such functional categories are missing from some or all of the education lists above. Developers should review lists used in other education metadata projects to achieve a more consistent set of categories. The list of publication types associated with the ERIC thesaurus is a good source for many of the functional categories relevant to searches by educators, though attention must be paid to creating a set of standard terms for resources unique to the Web.

Conclusion

A metadata record format that facilitates web searches for educational materials and cataloging of such materials by non-specialists can be created from existing technologies and categories of classification. The Dublin Core allows for creation of a manageable set of appropriate classification elements, which can be expanded as necessary by appropriate qualifiers. Established authority lists such as the ERIC thesaurus, which is actively maintained to reflect developing educational usage, provide a valuable source of terminology for controlled vocabularies. But the lack of consistency in selection of vocabularies as values of specific elements poses a substantial problem for metadata creators and searchers.

This problem of terminological control cannot be resolved without active collaboration among producers and users of metadata. One alternative to having all terms in a rich but unwieldy authority list like the ERIC thesaurus would be to separate out terms into categories that can serve as values of discrete elements. Instead of having publication types, population groups, pedagogical techniques, student groupings, etc. in a single list of subject terms, these components could be available as values of specific elements that are associated only with appropriate resources.

Another strategy might be to create broad conceptual classes of synonyms and related terms to which a term entered by a cataloger (or a user in a search) would be assigned automatically.

In order to accomplish the task of cataloging and providing access to present and future Web resources it is critical that metadata should be both simple to create and simple to use. Elements of resource description that are not critical to resource discovery should be isolated from the critical set of elements. For WALDO materials, the set of elements required to describe curricular materials should be distinguished from the smaller set that can characterize non-curricular resources quickly and efficiently. This can be accomplished within the GEM cataloging system by using a container architecture like that of the IMS to maintain distinct element sets.

References

[AACR2] Anglo-American Cataloguing Rules, 2d edition, 1988 rev.

[Bibliography] Digital Libraries: Metadata Resources.
http://ifla.inist.fr/ifla/II/metadata.htm

[CSI] US Department of Education, Office of Educational Research and Improvement, Cross-Site Indexing <http://165.224.220.67:8765/csi/>

[DC 1] Weibel, Stuart, Jean Godby, Eric Miller, Ron Daniel. OCLC/NCSA Metadata Workshop Report, 1995. http://www.oclc.org:5047/oclc/research/conferences/metadata/dublin_core_report

[DC 2] Dempsey, Lorcan and Stuart L. Weibel. The Warwick Metadata Workshop: A Framework for the Deployment of Resource Description. D-Lib Magazine, July 1996. http://www.dlib.org/dlib/july96/07weibel.html

[DC-4] Weibel, Stuart, Renato Iannella, and Warwick Cathro. The Fourth Dublin Core Metadata Workshop Report. D-Lib Magazine, June 1997. http://www.dlib.org/dlib/june97/metadata/06weibel.html

[DC-5] Weibel, Stuart. DC-5: The Helsinki Metadata Workshop. D-Lib Magazine, February 1998. http://www.dlib.org/dlib/february98/metadata/02weibel.html

[DC-REF] Weibel, Stuart and Eric Miller. Dublin Core Metadata Element Set: Reference Description. 1997 http://purl.org/metadata/dublin_core_elements

[ENC] Eisenhower National Clearinghouse Resource Finder http://www.enc.org/rf/index.htm

[ERIC] ERIC Database http://www.aspensys.com/eric/searchdb/dbchart.html

[H&D] Heery, Rachel, and Lorcan Dempsey. "A Review of Metadata: a survey of current resource description formats." 1996. http://www.ukoln.ac.uk/metadata/desire/overview

[HRST] Hearst, Marti A. "Interfaces for Searching the Web," Scientific American, March 1997, 68-72.

[IR] Sparck-Jones, Karen and Peter Willett. Readings in Information Retrieval. Morgan Kaufman, 1997.

[LAB] Northeast and Islands Regional Educational Laboratory at Brown University, cross-lab index. <http://www.lab.brown.edu/public/index.shtml>

[NMP] Hakala, Juha, Preben Hansen, Ole Husby, Traugott Koch, and Susanne Thorborg. The Nordic metadata project, final report. 1998. http://linnea.helsinki.fi/meta/nmfinal.htm

[RFC1] Weibel, Stuart L., John A Kunze, and Carl Lagoze. Dublin Core Metadata for Simple Resource Discovery, 1998. ftp://ietf.org/internet-drafts/draft-kunze-dc-02.txt

[RUS] Russom, Jacqueline. Metadata for Information Retrieval on the Internet: Background for WALDO Project.

[SAL] Salton, Gerard. Automatic Text Processing. Addison-Wesley, 1989.