Conceptual Design and Deployment of a Metadata Framework for Educational Resources on the Internet

Stuart A. Sutton
Syracuse University

Follow this and additional works at: https://surface.syr.edu/istpub

Part of the Databases and Information Systems Commons

Recommended Citation
https://surface.syr.edu/istpub/92

This Article is brought to you for free and open access by the School of Information Studies (iSchool) at SURFACE. It has been accepted for inclusion in School of Information Studies: Faculty Scholarship by an authorized administrator of SURFACE. For more information, please contact surface@syr.edu.
The metadata framework described in this article stems from a growing concern of the U.S. Department of Education and its National Library of Education that teachers, students, and parents are encountering increasing difficulty in accessing educational resources on the Internet even as those resources are becoming more abundant. This concern is joined by the realization that as the Internet matures as a publishing environment, the successful management of resource repositories will hinge to a great extent on the intelligent use of metadata. We first explicate the conceptual foundations for the Gateway to Educational Materials (GEM) framework including the adoption of the Dublin Core Element Set as its base referent, and the extension of that set to meet the needs of the domain. We then discuss the complex of decisions that must be made regarding selection of the units of description and the structuring of an information space. The article concludes with a discussion of metadata generation, the association of metadata to the objects described, and a general description of the GEM system architecture.

Introduction

Enhanced access to educational materials on the Internet for the nation’s teachers and students is one of President Clinton’s second-term goals. In pursuit of that goal, the U.S. National Library of Education (NLE) Advisory Task Force identified lesson plans and teacher guides as a critical area in which library and information science expertise should be applied in order to improve the organization and accessibility of large collections of educational materials that are already available on various federal, state, university, non-profit, and commercial Internet sites. The U.S. Department of Education and the NLE charged the ERIC Clearinghouse on Information and Technology at Syracuse University with the task of spearheading a project to develop an operational framework to provide the nation’s teachers with “one-stop/any-stop” access to this vast pool of educational materials on the Internet. These valuable resources are difficult for most teachers, students, and parents to find and evaluate in an efficient, effective manner. The goal of the Gateway to Educational Materials (GEM) project is to alleviate this resource discovery problem through development and deployment of a metadata element set and accompanying procedures for its use.

Dempsey and Heery (1997) state:

It is recognized that in an indefinitely large resource space, effective management of networked information will increasingly rely on effective management of metadata. The need for metadata services is already clear in the current Internet environment. As the Internet develops into a mixed information economy deploying multiple application protocols and formats, this need will be all the greater. Metadata is not only key to discovery, it will also be fundamental to effective use of found resources (by establishing the technical or business frameworks in which they can be used) and to interoperability across protocol domains.

To develop a mechanism for the design and development of a metadata framework rooted in constituency needs, a stakeholders’ meeting was held at Syracuse University in November, 1996. In addition to seeking advice from agencies with repositories of Internet-based educational resources and individuals, organizations with subject expertise, and individuals with experience in the development of metadata schemes, input was sought from front-line teachers. In the months following the stakeholders’ meeting, the group of participants served as a GEM Working Group (GWG) to develop and refine the work begun at Syracuse. The plan was for the GWG to develop the foundation for the GEM framework through the conceptual phase of the project. Subsequently, a broader constituency in the form of the GEM Consortium was engaged in the full-scale deployment of GEM metadata through its application to educational materials across the Internet.

The five major tasks addressed by the GEM project are:

© 1999 John Wiley & Sons, Inc.
(1) Define a semantically rich metadata profile and domain-specific controlled vocabularies necessary to the description of educational materials on the Internet;

(2) Develop a concrete syntax and well-specified practices for its application using the current HTML specifications relied on by existing Internet browsers while preparing for emerging standards such as Resource Description Framework (RDF) and eXtensible Markup Language (XML);

(3) Design and implement a set of harvesting tools for retrieving GEM metadata distributed in repositories across the Internet;

(4) Encourage the design of a number of prototype interfaces to GEM metadata; and

(5) Develop a governance mechanism for GEM.

This paper explores the first four of these tasks in depth, and touches only tangentially on the fifth.

**GEM Foundations**

Two decisions of the GWG, made at its initial meeting, were pivotal in shaping the subsequent direction of GEM development. The first key decision was to accept the then nascent Dublin Core Element Set (DC) (and its metadata model) as the project’s base referent. The second key decision was that the long-term work on GEM was to be the function of a consortium of Internet repositories of educational resources, subject experts, and end-user representatives—what would evolve into the GEM Consortium. From this second decision stems the concept of the Gem Metadata Union (GMU). Due to the fundamental significance of these two decisions, each is chronicled at some length in the following paragraphs.

**Dublin Core**

From the outset, the GWG wanted to develop the various aspects of GEM around emerging standards for networked information discovery and retrieval (NIDR). As noted above, the decision was made that GEM would assume DC as its base referent due to what then appeared to be its growing national and international recognition, acceptance, and support. At the time this decision was made (1996), DC was very much in its formative stages. At the time of this writing (1998), it is moving toward formal standardization in its unqualified form through submission of the first of five Requests for Comments (RFC) to the Internet Engineering Task Force (Weibel, Kunze, & Lagoze, 1998). Subsequent RFCs will consider (a) conventions for embedding unqualified DC elements in HTML files, (b) principles for “qualifying” DC elements, (c) encoding qualified DC elements in HTML, and (d) encoding qualified DC elements in RDF (Resource Description Framework). NISO and ISO standardization initiatives are also under consideration.

Table 1 briefly describes DC’s 15 elements.

**Table 1. Dublin Core Element Set**

<table>
<thead>
<tr>
<th>Dublin Core Element</th>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Title</td>
<td>The name given the resource by the creator or publisher</td>
</tr>
<tr>
<td>Author or Creator</td>
<td>Creator</td>
<td>The person/organization primarily responsible for the intellectual content</td>
</tr>
<tr>
<td>Subjects and Keywords</td>
<td>Subject</td>
<td>The topic of the resource</td>
</tr>
<tr>
<td>Description</td>
<td>Description</td>
<td>A textual description of the content of the resource</td>
</tr>
<tr>
<td>Publisher</td>
<td>Publisher</td>
<td>The entity responsible for making the resource available in its present form</td>
</tr>
<tr>
<td>Other Contributor</td>
<td>Contributor</td>
<td>Secondary contributors to the intellectual content</td>
</tr>
<tr>
<td>Date</td>
<td>Date</td>
<td>The date the resource was made available in its present form</td>
</tr>
<tr>
<td>Resource Type</td>
<td>Type</td>
<td>The category of the resource</td>
</tr>
<tr>
<td>Format</td>
<td>Format</td>
<td>The data format of the resource</td>
</tr>
<tr>
<td>Resource Identifier</td>
<td>Identifier</td>
<td>A string or number that uniquely identifies the resource</td>
</tr>
<tr>
<td>Source</td>
<td>Source</td>
<td>A string or number used to uniquely identify the work from which this resource was derived</td>
</tr>
<tr>
<td>Language</td>
<td>Language</td>
<td>Language of the intellectual content of the resource</td>
</tr>
<tr>
<td>Relation</td>
<td>Relation</td>
<td>The relationship of this resource to other resources</td>
</tr>
<tr>
<td>Coverage</td>
<td>Coverage</td>
<td>The spatial and/or temporal characteristics of the resource</td>
</tr>
<tr>
<td>Rights Management</td>
<td>Rights</td>
<td>A link to copyright and/or use restriction statements</td>
</tr>
</tbody>
</table>
pect and practice domains evolve, the existing (and future) Web-crawling programs, such as AltaVista, Excite, InfoSeek, and Lycos, will become increasingly metadata-aware, and, therefore, more efficient and effective in providing access to information. In addition to these general retrieval services, a number of services fashioned to meet the needs of specific domains are emerging (Dempsey, 1996). Readily available tools make local “harvesting” of metadata possible and their extension to multiple Web sites serving a specific community has been demonstrated (Beckett & Smith, 1996).

The GWG decided that it would pursue a framework for access to GEM metadata along the lines of both of these general Web resource discovery models. First, GEM would promote the wide deployment of GEM metadata across the Internet in a form accessible by the general service Web-crawling applications as they become more metadata-aware. To this end, GWG would make the tools necessary to the creation of GEM metadata and the association of that metadata to described resources publicly and freely available. However, in addition to this widespread deployment of GEM metadata, the GWG would create a service fashioned to meet the specific needs of teachers and students by guaranteeing “one-stop/any-stop” access to quality educational materials. To achieve this second goal, the GWG created a consortium for high-integrity repositories of educational materials known as the GEM Consortium Repositories (GCR). In addition to establishing the GCR, a set of tools was created to harvest GEM metadata from the GCR to build the Gem Metadata Union (GMU). Currently, the GMU is accessible through any GCR; in the near future, the GMU will be Z39.50 compliant.

The rationale for the GMU can be found in the following observation of Lagoze, Lynch, and Daniel in their exploration of issues surrounding the Dublin Core (1996, p. 6):

[T]he use of the Dublin Core in a limited context might produce very positive results. For example, assume a set of “high-integrity sites.” Administrators at such sites might tag their documents…with Dublin Core metadata elements using a set of well-specified practices that include relatively controlled vocabularies and regular syntax. Retrieval effectiveness across these high-integrity sites would probably be significantly better (assuming harvesting and retrieval tools that make use of the metadata) than the unstructured searches available now through Lycos and Alta Vista.

The growing consortium of GEM repositories is just such a set of “high-integrity sites.”

Given the context provided by these two key decisions, the GWG began work on the GEM framework to which we now turn.

GEM Extensions to the Dublin Core

Having assumed a base referent in DC, and before consideration could be given to potential extensions to DC, fundamental questions of the role(s) the GEM element set was to play had to be addressed. Dempsey & Heery (1997) describe the broad range of roles and other factors affecting the development of metadata schemes:

There is a variety of types of metadata. There is traditional descriptive information of the kind found in library catalogues, which typically includes such attributes as author, title, some indication of intellectual content and so on. There is information that might help a client application make a decision based on format (where certain local browser equipment is available) or on location (to save bandwidth). There are different types of user: a user as customer wishes to know the terms under which an object is available; a user as researcher may wish to have some extended documentation about a particular resource, its provenance for example. There are different types of resource. Some resources may have a fugitive existence, existing to satisfy some temporary need and only ever minimally described if at all; some are important and valuable scholarly or commercial resources, where the value of extensive description is recognized. Some resources may be simple; some may be complex in various ways. There will be many different information providers, some commercial “yellow pages” type services, some scholarly or research-oriented services, in different organizational configurations with different target audiences and products. Metadata may be closely coupled with the object it describes as an intrinsic part of its composition; or it may have no intrinsic link with it at all.

The metadata continuum in Figure 1 (freely adapted from Murphy, 1998, and Dempsey, 1996a) frame the complexity identified by Dempsey and Heery in terms of (a) three general metadata roles (location, discovery, and documentation), (b) the degrees of complexity of metadata at various points along the continuum, and (c) the varying amenability of that metadata to mechanized generation. Figure 1 also illustrates how the GWG finally envisioned GEM’s role in the resource discovery and documentation processes and its relationship to DC and richer, more complex schemes such as MARC.

The metadata continuum ranges from simple, terse descriptions found in the indexes of existing Web-crawling retrieval engines, to complex, rich descriptions such as those found in fully cataloged bibliographic records using MARC. Functions along the continuum range from (a) simple location of items meeting a fixed set of known criteria, (b) through the more complex process of discovery of resources of possible utility to an end-user, to (c) full documentation embodying information not contained explicitly within the resource being described (e.g., information about provenance and the relation of the resource to other resources). DC is envisioned by the DC “minimalists” (DC adherents opposing extensive qualification of the base elements) as resting in the middle of this continuum and playing a single role — discovery. The DC “structuralists” (DC adherents embracing modest-to-rich qualification of the base elements) envision a larger DC role that pushes DC
along the continuum toward documentation. Beyond agreement on the methods for qualifying the DC element set through the Canberra Qualifiers (see Weibel & Iannella, 1997), there appears to be little agreement in the DC community on whether DC should be extended to accommodate functions on the right side of the continuum.

Sutton and Oh (1997, p. 22) describe the GWG decision reached regarding the roles GEM would play, as well as the relationship between the DC and GEM element sets as follows:

The general goals of GEM and DC are similar; however, in many ways, they are not congruent. DC is designed to serve NIDR through a fielded surrogate supposedly simple enough to be applied to resources by authors and Internet providers untrained in the complexities of cataloging necessary to the creation of more richly structured surrogates (e.g., the MARC record). While its simplicity serves coarse-grained NIDR across a broad range of networked information, DC is ill-equipped for more fine-grained NIDR of resources necessary to particular discourse or practice communities, such as the nation’s teachers. GEM is intended to serve NIDR needs of this constituency along a continuum that begins with what is achievable with a simple, unqualified, fielded surrogate as set out by the DC “minimalists” to a surrogate coming closer to (but never reaching) the richly structured surrogate. In addition, GEM assumes that the [element set] will be applied by a range of organizations with a higher level of commitment and expertise than that assumed by DC.

To meet this need to serve a broader range of functions along the Figure 1 continuum than the narrow function envisioned for DC, the GWG added an eight-element, domain-specific GEM package and a rich array of element qualifiers. Because the primary goal of GEM remains the discovery of resources, a consortium collection holder with limited means may elect to catalog at “level 1” capturing just enough information for effective NIDR. Basic NIDR functions are served at level 1 with valid, unqualified content for DC elements Title, Subject, Format, Resource Type, Publisher (Online Provider), and Description, and for a single GEM element, Grade (discussed below). However, collection holders wishing to use GEM for purposes of both discovery and documentation may do so by engaging in “level 2” cataloging. Level 2 is achieved by the addition of information beyond that required for level 1.

As Figure 1 implies, GEM metadata is intended to serve the differing metadata needs of a range of users including those of teachers, parents, students, and metadata repository managers. At the most basic level of discovery, the element set provides students and parents with substantial topical information as well as grade levels for retrieval that is more precise. The element set provides the teacher with additional information in order to tailor a search for resources meeting specific pedagogical methods or learning objectives. GEM’s capacity to capture information necessary to a fairly precise documentation of a resource serves the needs of those repositories wishing to use the element set as a primary means of managing resources.

Exploratory research was performed by the GEM project team at Syracuse (Small et al., 1999) in order to determine both the information-seeking behavior of teachers searching for educational resources on the Internet and the language teachers used in their queries. The study found that there were a number of descriptive elements used by teachers...
when they searched for instructional resources that are not found in DC. The result was the addition of a GEM package containing five description elements, two evaluative elements, and one meta-metadata element.

Description Elements

The following elements were necessary to capture basic descriptive data for many (but certainly, not all) resources in the domain:

- **Audience**: The audience element identifies the specific audience of the resource being described. The element has a controlled vocabulary and several types or subelements that identify two conceptually distinct aspects of audience. The first aspect denotes the type of person that the resource directly targets. For example, is the resource intended for direct use by a teacher, a student, or a parent? The second aspect denotes the ultimate beneficiary of the resource. In other words, while the resource may be targeted to teachers, the ultimate beneficiaries of its use may be students with disabilities, their parents, or even other teachers.

- **Duration**: The duration element contains an uncontrolled statement of the time needed to effectively use the resource and might include statements such as: “Two one-hour class sessions” or “50 minutes.”

- **Essential Resources**: The essential resources element is used to briefly describe any resources beyond the resource being described that must be present for its effective use. For example, a lesson plan might require the use of a computer in the classroom equipped with a specific software package. Both the computer and the software would be listed in the metadata as essential resources. This allows the user to determine whether he or she should “visit” the actual resource being described if a computer and the necessary software are unavailable in the classroom.

- **Education Level**: The educational level element describes the grade, grade span, or education level of the object’s ultimate beneficiary.

- **Pedagogy**: The pedagogy element has four types or subelements under vocabulary control denoting any teaching methods used by the resource, any student instructional groupings, methods used to assess student performance, and any learning prerequisites to effective use.

Evaluation Elements

The GEM element set contains two elements for evaluative data. The first provides for evaluation of the quality of the resource and the second for assessing the utility of the resource in terms of individual state or national academic standards. These assessment metadata may be either internally referenced or externally referenced. In discussing the Warwick Framework, Manola (1998) defines these two states as follows: “An internally referenced metadata container is metadata that may be created and maintained by an authority separate from the creator or maintainer of the content object. In fact, the creator of the object may not even be aware of this metadata.” Figure 2 illustrates the logical aggregation of such externally referenced metadata packages in a Warwick Framework-like “container.”

The need for a framework to achieve such aggregation of multiple metadata objects was recognized early in the work of the Dublin Core, because the disparate metadata needs of various communities would surely result in multiple metadata sets—many of which reference the same resource. Reporting on the second Dublin Core workshop at Warwick University in England, Dempsey and Weibel (1996; see also Lagoze, Lynch, & Daniel (1996a)) describe the Warwick Framework:

Satisfying the need for competing, overlapping, and complementary metadata models requires an architecture that will accommodate a wide variety of separately maintained metadata models. It was concluded that an architecture for the interchange of metadata packages was required. A package is conceived as a metadata object specialized for a particular purpose. A Dublin Core-based record might be one package, a MARC record another, and so on. Such discrete packages might be numerous and varied in content and even source. Users or software agents would need the ability to aggregate these discrete metadata packages in a conceptual container (a metadata basket of sorts), hence the notion of a container-package architecture. This architecture would be modular, to allow for differently typed metadata objects; extensible, to allow for new metadata types; distributed, to allow external metadata objects to be referenced; and recursive, to allow metadata objects to be treated as “information content” and have metadata objects associated with them.

FIG. 2. Aggregating metadata packages.
As Figure 2 denotes, the presence of at least three types of GEM metadata packages is currently assumed: (a) descriptive metadata packages that may or may not contain their own quality assessment and academic standards mapping information; (b) academic standards mapping metadata packages; and (c) resource quality assessment packages.

Meta-Metadata Element

The GEM Cataloging Agency element was added to rectify what some believe to be a major shortcoming of the Dublin Core element set—the lack of any statement of metadata “authority.” Clarke (1997) rightly notes that DC provides no means for capturing data about metadata (i.e., meta-metadata) such as the “identity and affiliation of the author of the meta-data (as distinct from the author of the object itself).….Without such information: [1] the end user is left uninformed about the source of the meta-data, and hence cannot make a judgment about its reliability; [2] there is no ability to distinguish between multiple instances of meta-data about the same object; and [3] there is no ability to choose among alternative instances of meta-data about the same object.” (Emphasis added)

The GWG thought such identification would be crucial in an open, distributed system that makes it possible for any agency or service bureau to create separate GEM metadata packages (description, quality assessment, and academic standards mapping packages) as externally referenced metadata packages and to deploy those packages on the Web. Figure 3 illustrates the potential existence of multiple metadata packages referencing the same primary object.

In Figure 3, we see metadata packages from six different agencies or service bureaus—three that have done standards mappings and three packages of quality assessments. The arrows from the primary source to the Agency B and Agency F packages denote direct internal references to those packages using the DC relation element, and imply approval of their contents by the creator of the internal metadata for the resource.

We already see such GEM quality assessments emerging. For example, the University of Georgia School of Education has done quality assessments of approximately 100 lesson plans in the AskERIC Virtual Library at the ERIC Clearinghouse on Information and Technology using the GEM quality indicator scheme. It is preparing to do an assessment of educational resources made available through the U.S. Department of Education. In keeping with an emerging object model of the Web (see Manola, 1998), each of these assessments will exist as externally referenced metadata containers. Knowing the identity of the creator of those assessments (or the person’s agency affiliation) will aid the end-user in evaluating that assessment’s reliability and credibility. The minimal level of compliance with this element is to include the metadata-creator’s name or affiliation.

Unit of Descriptive Analysis

GEM is designed to handle the creation of metadata for educational resources at any arbitrary level of granularity. In implementing GEM, a repository of resources must determine its unit(s) of descriptive analysis. Figure 4 creates a hypothetical information space for the U.S. Presidency and illustrates a range of possible choices in determining the resource objects to be described.

In Figure 4, resources at three levels of granularity are denoted A, B, and C. Each resource has either internal or external metadata and all may reside on different servers or on a single server. At level C is found two images containing portraits of President Carter and two of President Nixon. Similar treatment may be imagined for the remainder of the U.S. Presidents. Each image has externally referenced metadata. The two sets of images are referenced by a page at level B dealing with the history of the term of office for each President in a series named “History of the U.S. Presidency.” In like fashion, the images are referenced by a second set of level B pages in a portraiture series called “U.S. Presidential Portraits.” Each level B resource carries internal metadata. The two level A resources are meta-objects functioning as top-level index pages for the two series—each contains its own metadata. The dashed arrows represent traditional in-line HTML links. The solid arrows represent several explicit relationships among a number of the pages: (a) links binding the image metadata to their respective images at level C; (b) links binding level B children to their level A parents using the DC relation element; and (c) links between siblings using the DC relation element at level B. GEM uses the relation element to provide typed linkages to a rich array of related objects well beyond the familial, including, among others, links to companion data sets, peer reviews, sponsorship organizations, site criteria for resource selection, external standards mappings, and external quality ratings (as illustrated in Figure 3).
Figure 4 defines a hypothetical information space in which every primary object directly accessible by an end-user has associated metadata. A simple search for either President would retrieve the associated objects at all three levels of granularity. Thus, a teacher wanting a complete resource covering the historical aspects of the Presidency has access at two levels of granularity—levels A and B. A teacher wanting to build such a resource has access to fundamental building blocks at level C.

In reality (and unlike the environment illustrated in Figure 4), it is unlikely for a range of reasons (e.g., cost and perceived purpose and utility) that metadata will be created at all levels by any single repository. In some instances, metadata may be created only for level A meta-objects. Given sufficient descriptive information regarding the series content that these meta-objects represent, discovery and retrieval is possible without providing direct access to the space’s functional leaves at level B. However, in another instance, fairly fine-grained NIDR could be provided by creating metadata solely at level B with relations to higher-level meta-objects being provided using the DC relation element. In such a case, sibling relationships may not be explicitly defined, relying instead on implicit linkages through the parent meta-object.

While the GEM deployment of the DC relation element in its metadata-generating tools provides the mechanism for creating highly integrated information spaces, GEM requires neither the use of the element nor, when used, that it be done at any particular level of granularity. Such policy decisions are left to the GEM repositories. In current practice, it appears that repositories favor the creation of metadata solely at the functional leaves (level B).

Creating GEM Metadata

While GEM metadata can be created with any text editor, a cross-platform metadata-generating application called GEMCat was developed in Java to ease the creation process by allowing the creator to focus solely on content. Figure 5 illustrates the general architecture of the GEMCat implementation. GEMCat is intended for use by repositories needing to create metadata for a substantial number of valuable, stable Web resources.

Figure 5 denotes three system components or modules (A, B, and C), and two metadata output formats (D and E). In the following paragraphs, we briefly examine each of these components and formats.

Site Profile

The module A site profile is designed to achieve two primary functions:

1. To reduce the repetitive rekeying of redundant and default data by allowing such information to be entered once at the site administration level with appropriate session and record overrides; and
2. To provide a mechanism for generalizing the use of
schemes (as defined in the DC metadata model) in order to allow an agency to substitute one controlled vocabulary for another according to specific needs.

A number of the GEM elements require the use of controlled vocabularies including the subject and resource type elements. While GEM maintains its own subject vocabulary, it is intentionally terse (approximately 250 terms). It is designed to be easy to apply and capable of getting the end user into the right “subject neighborhood” and to cluster educational materials in ways commonly understood by the education community. However, it is recognized that this minimalist vocabulary is inadequate to meet the needs of many subdomains in the education community. Therefore, GEM provides for the addition of subject terms from any appropriate controlled vocabulary such as the ERIC and NICEM thesauri for education terminology, or specialized subject thesauri such as the Art and Architecture Thesaurus.

The site profile makes it possible for an agency to identify such vocabularies and to configure GEMCat to include appropriate scheme identifiers in the metadata. In like fashion, an agency with particular resource types may configure GEMCat to accept an alternative scheme for the required resource type element.

Metadata Generation

The metadata-generating module B is a cross-platform Java application with three basic functions: (a) accepting metadata input from both the creator of the metadata and the site profile, (b) validating the metadata where possible, and (c) passing it to the syntax engine. The current generation of module B provides simple pull-down menus for GEM-controlled vocabularies. The next generation will provide for the automatic inclusion of alternative controlled vocab-

HTML Syntax

The changes in HTML’s ability to effectively accommodate richly structured metadata through meta tags has been chronicled elsewhere (Weibel & Iannella, 1997) and is discussed only briefly here because this evolution has significant impact on projects seeking to implement. From the beginning of the DC dialog, it has been the consensus that only the simplest of implementations of unqualified DC could be accommodated effectively by the limited HTML 2.0 meta tag (see Weibel & Iannella, 1997). Given the HTML 2.0 limitation of relevant elements to NAME and CONTENT, there was no other means of dealing with DC’s qualifying information (schemes and subelements) other than through “overloading content” (Weibel & Iannella, 1997), which takes the form:

```
<META NAME="DC.subject" CONTENT="(SCHEME=LCSH) (LANG=en) Electronic data processing"
```

Weibel and Iannella (1997) note that “[t]his approach has the significant disadvantage of cluttering the CONTENT field with qualifier information. Since the goal of embedded metadata is in part to make the metadata more readily visible to harvesters, the obfuscation of the CONTENT attribute is suboptimal to say the least. A smart harvester might parse the CONTENT attribute intelligently and use (or ignore) the SCHEME and LANG identifiers sensibly, but it is unreasonable to expect this to be the norm.”

What was needed with HTML 2.0 was the ability to define more structure through a broader array of meta tag elements. This could be achieved by simply adding a SCHEME element to the meta tag and trusting that it would be ignored by most Web agents. However, without formal specification, this “additional attribute” approach with HTML 2.0 might result in a failure to validate by some
parsing and undefined behavior by editing software (Weibel & Iannella, 1997).

The HTML 4.0 Specification (Raggett, 1997) somewhat alleviates the problem through the addition of the SCHEME and LANG elements to the list of meta tags. When these additions are combined with the appending of “type” or subelement information to the NAME value, the “overloading content” problem is eliminated. The following GEM metadata example illustrates the integration of scheme and subelement information in HTML 4.0 meta tags:

```html
<META NAME="DC.package.begin" CONTENT="Version 1.0">
<META NAME="DC.Identifier" SCHEME="URL" CONTENT="gem.syr.edu/Stuff/Whales.html">
<META NAME="DC.date.recordCreated" SCHEME="" CONTENT="19971127">
<META NAME="DC.subject.level1.1" SCHEME="GEM" CONTENT="Science">
<META NAME="DC.subject.level2.1" SCHEME="GEM" CONTENT="Biological sciences">
<META NAME="DC.subject.level2.1" SCHEME="GEM" CONTENT="Life sciences">
<META NAME="DC.subject.level2.1" SCHEME="GEM" CONTENT="Technology">
<META NAME="DC.format.contentType" SCHEME="" CONTENT="text/HTML">
<META NAME="DC.publisher.role.2" CONTENT="OnlineProvider">
<META NAME="DC.publisher.name.2" CONTENT="Discovery Communications">
<META NAME="DC.publisher.email.2" CONTENT="info@school.discovery.com">
<META NAME="DC.publisher.homePage.2" SCHEME="URL" CONTENT="http://school.discovery.com/">
<META NAME="DC.type" SCHEME="GEM" CONTENT="Lesson plan">
<META NAME="DC.title" CONTENT="H2Oceans: In the Company of Whales">
<META NAME="DC.package.end" CONTENT="Version 1.0">
<META NAME="GEM.package.begin" CONTENT="Version 2.0">
<META NAME="GEM.cataloging.name" CONTENT="ERIC/IT">
<META NAME="GEM.grade.grades" scheme="GEM" CONTENT="6, 7, 8, 9, 10, 11">
<META NAME="GEM.package.end" CONTENT="Version 2.0">
```

In this example, the metadata is divided into DC and GEM “packages.” Related tags are aggregated through the use of an arbitrary grouping number (e.g., the numbers appended to the subject (“1”) and publisher (“2”) elements). In addition to aggregation, the grouping number renders unique the NAME values for multiple occurrences of the same element (e.g., multiple creators and subjects).

**Associating Metadata**

When the metadata creation process for a resource is complete, the creator has several options available for associating that metadata with the resource described. In discussing metadata in the context of the RDF working draft, Manola (1998) describes four mechanisms for associating metadata with the resource described:

1. It may be contained within the resource (embedded);
2. It may be external to the resource but supplied by the transfer mechanism in the same retrieval transaction as that which returns the resource (along-with);
3. It may be retrieved independently from the resource, including from a different source (service bureau);
4. It may contain the resource (wrapped).

At this time, GEM metadata may be associated with the resource described using the first three mechanisms—the first mechanism is embodied in the Figure 5 format E, while the second and third are embodied in Figure 5’s format D. Manola goes on to note that “[a]ll resources will not support all association methods (e.g., many resource types will not support embedding.” Thus, writing GEM metadata to a separate file for along-with and service bureau associations is required when the file format of the original precludes embedding (e.g., sound, video, applications, and image files) or the person or agency creating the metadata has no direct access to the resource being described.

**GEM System Architecture**

The processes used in the generation of GEM metadata are but one part of a larger system architecture illustrated in Figure 6.

Figure 6 is divided into a number of “zones” (A, B, and C) and a series of functions (1–4). Zone A represents the Internet site of a GCR (GEM Consortium Repository) with its attendant educational materials and a local site index, which consists of the metadata associated with those resources. Zone B denotes the presence of the GMU (GEM Metadata Union), which resides at ERIC/IT. And Zone C represents value-added products derived from the GMU for searching and navigating educational resources. The four functions that overlay the zones are discussed in the following paragraphs.

**Creation of Metadata**

In the framework, each Consortium member is responsible for the creation and storage of its own metadata. Figure 6 illustrates the two available options discussed earlier: (a) embedding of the metadata directly into the site’s HTML resources, and (b) storing the metadata separate from the resource being described. Of course, different GCRs handle the storage of educational resources and GEM metadata in different ways. Because one GEM project goal is to be unintrusive when it comes to issues of local implementation, how a Consortium member handles its GEM metadata is its own concern so long as it is capable of generating the local site index described in the second function.
Local Site Index

The second function represents the harvesting of a Consortium site’s metadata and the creation of a site index in the form of a simple ASCII file. The harvesting is accomplished through use of a publicly available, cross-platform GEM harvester implemented in Java. It is the responsibility of the Consortium member to periodically generate its local site index of new materials based on the frequency with which curriculum resources are added, deleted, or moved.

As noted above, local implementations may take a completely different route to the creation of the local site index than the one illustrated in Figure 6. For example, the Eisenhower National Clearinghouse on Science and Mathematics (ENC) maintains its resources in a relational database management system from which HTML pages are generated dynamically in response to user queries. ENC’s GEM metadata results from a mapping of the ENC data structures directly to the GEM element set and a site index.

Aggregation

With the exception of metadata generated by a service bureau, the first two functions in Figure 6 (metadata generation and harvesting) occur at the Consortium member repository. In the third function, the focus shifts from the individual repositories to the ERIC Clearinghouse on Information and Technology where the local site indexes from the repositories are systematically merged to form the logical Union by being parsed into a database, as illustrated in the Figure’s function four.

Resource Discovery

The ERIC Clearinghouse on Information and Technology has been charged by the GWG with providing access to the Union through one or more readily available interfaces. However, it is not assumed that ERIC/IT will provide the only such interface. All Consortium members have a right to the Union and are at liberty to develop their own interfaces to either the entire Union or some select subset of it.

In the project’s first phase, two prototype interfaces to the GMU were developed. At Syracuse, a search and browse environment was built using PLWeb, a full-text, relevance-ranking search engine by Personal Library Software. The PLWeb implementation currently provides public access to GEM metadata. A relational database implementation was developed, which will form the backbone of the Union as the number of repositories grows.

In order to extend access, the database will be made available through the Z39.50 protocol (Z39.50 Maintenance Agency, 1995) using the GILS profile (Christian, 1997) and a Web gateway. At this time, GILS does not support the full GEM element set; therefore, the GEM simple and advanced search interfaces are maintained with the option for the user to expand a search to the larger domain served by the Z39.50 protocol.

Conclusion

As the World Wide Web grows exponentially, discovery and retrieval of useful educational materials grows more problematic. The framework developed by the GEM project seeks to meet the needs of educators, students, and parents through development and wide deployment of the GEM standard in the form of a metadata element set, an accompanying set of controlled vocabularies, and a well-defined set of practices in their application. In addition, the GEM Metadata Union promises enhanced “one stop/any stop” access to Internet-based educational materials.

References


