Constructing a Metadata Database to Enhance Internet Retrieval of Educational Materials *

Sam-Gyun Oh**

contents

ABSTRACT

This paper reports the GEM (Gateway to Educational Materials) project whose goal is to develop an operational framework to provide the K-12 teachers in the world with "one-stop/any-stop" access to thousands of lesson plans, curriculum units and other Internet-based educational resources. To the 15-element Dublin Core base package, the GEM project added an 8-element, domain-specific GEM package. The GEM project employed the conceptual data modeling approach to designing the GEM database, used the Sybase relational database management system (RDBMS) to construct the backend database for storing the metadata of educational resources, and also employed the active server page (ASP) technology to provide Web interfaces to that database. The consortium members catalog lesson plans and other Internet-based educational resources using a cataloging module program that produces HTML meta tags. A harvest program collects these meta tags across the Internet and outputs an ASCII file that conforms to the standard agreed by the consortium members. A parser program processes this file to enter meta tags automatically into appropriate relational tables in the Sybase database. The conceptual/logical schemas of Dublin Core and GEM profile are presented. The advantages of conceptual modeling approach to manage metadata are discussed. A prototype system that provides access to the GEM metadata is available at http://lis.skku.ac.kr/gem/

초 록

이 논문은 미국 초중고교 교사들을 이용 대상자로 인터넷 상에 산재해 있는 강의안 및 교육자료의 메타데이터 DB를 구축한 GEM 프로젝트에 대한 보고이다. GEM 프로젝트에서는 현재 거의 표준으로 간주되는 더블린 코어의 15개 요소(Elements)를 채택하였고, 여기에 8개 요소를 첨가하여 검색을 원활히 하고자 하였다. GEM 메타데이터 DB의 구축에는 메타데이터 요소들간의 관계를 좀 더 명확히 표현할 수 있는 개념적 데이터 마들링을 사용하였고, 메타데이터는 Sybase라는 관계형 데이터베이스에 저장했으며, 이 DB에 웹 인터페이스를 장착하는 데에는 Microsoft 액티브 서버 페이지 (ASP) 기술을 이용하였다. GEM 메타데이터의 실제목록은 미국 전역에서 참가하고있는 컨소시엄 회원들에 의해서 이뤄지고 있으며, 그 결과는 인터넷을 통해 Sybase 관계형 데이터베이스에 자동적으로 입력된다. 이 논문에서는 더블린코어, GEM의 개념 및 논리 스키마들을 제시하는 한편, 메타데이터 DB의 구축에 개념적 데이터 마들링을 사용함으로써 얻어지는 장점들을 논하였다. GEM 프로토타입 시스템이 가동되고 있는 URL은 다음과 같다:
http://lis.skku.ac.kr/gem/

* The Author was in charge of the Web-Based database construction of this project which was funded by the US Department of Education while he was a faculty member at the University of Washington.

** Sung Kyun Kwan University, Department of Library and Information Science, Assistant Professor

접수일자 1998년 9월 1일

1. INTRODUCTION

Many lesson plans, curriculum units and other educational materials are available on the Internet. These valuable resources are difficult for most teachers to find in an efficient and effective manner. The goal of the GEM Project is to substantially alleviate this resource discovery problem. The US Department of Education funded this project while the author was a faculty member at the University of Washington. The GEM project adopted the Dublin Core (DC) for its base metadata elements to describe the lesson plans and other Internet-based educational resources. The GEM project team conducted a user survey of K-12 teachers to find whether the DC-elements would be sufficient to provide all the access points needed by teachers and found that the following additional 8-elements were needed to provide K-12 teachers with better access to those resources:

- Audience : The GEM audience type element contains information from a GEM controlled vocabulary that most closely identifies the specific audience of the resource being described.
- Cataloging Agency : The cataloging agency provides basic information about the agency that created the GEM catalog record.
- Duration : Theduration of the activity or lesson.
- Essential Resources : Resources essential to the effective use of the educational resource by the teacher.
- Educational Level : Grade, grade span, or educational level of the resource's audience
- Pedagogy : The teaching methods, student instructional groupings, and assessment methods.
- Quality Assessments : The following scheme (accuracy, appropriateness, clarity, completeness, motivation, and organization) for assessing the quality of instructional materials is employed.
- Educational Standards : State and/or national academic standards mapped to the educational resource being described.

The basic DC package and additional 8 GEM elements will be referred to as the GEM element set in this paper. General information on the GEM project can be found at http://gem.syr.edu.

This paper will focus on the database-driven implementation of the GEM project. The paper will discuss the approaches taken to construct a backend

database that stores GEM data and to build the Web-based interfaces that provide flexible and effective access (searching and browsing) to the GEM database.

2. THE CONCEPTUAL MODELING APPROACH

The conceptual model represents a global view of the data and is the basis for the identification and description of the main data objects (Rob & Coronel, 1997). The most widely used conceptual model is the Entity-Relationship (ER) model. Using the ER model yields the conceptual schema, which is the basic database blueprint. The conceptual model offers some important advantages. First, it forms the basis for the conceptual schema, which provides a relatively easily understood view of the data environment. Second, the conceptual model is independent of both software and hardware. The ER model was employed to develop a conceptual schema that captures the diverse relationships among the entities associated with lesson plans and other educational resources. The GEM project team contends that a MARC-type data structure in describing lesson plans will not be sufficient to handle the complexities and sophisticated demands of Web user queries because it does not capture the relationships among entities in a structured way. This weakness prevents information retrieval (IR) systems based on the MARC data structure from grouping the retrieved sets of the resources in meaningful ways for users. For example, many subjects are assigned to a particular lesson plan and a particular subject can be assigned to many different lesson plans. Creating a list, which provides the user with ability to browse the resources by subjects, would be very difficult and inefficient if the MARC-type data structure is employed. The ER model has the power to capture all of these relationships and also possesses the requirements of a good semantic data model because it has the following important characteristics (Elmasri & Navathe, 1994; Oh, 1995):

- it is expressive enough to point out commonly occurring distinctions between types of data, relationships, and constraints;
- it is simple and its concepts can be easily understood;
- it has a small number of basic concepts that are distinct and non-overlapping in meaning; and
- it has a diagrammatic notation for displaying a conceptual schema that is easy to interpret.

The idea that conceptual data modeling can be beneficial to IR design is not new. Agosti et al (1989) expressed the need for a conceptual paradigm for the design of advanced IR applications, and Ingwersen (1992) also asserted that IR systems should provide users with different ways of accessing documents because users have different cognitive views. The conceptual database design approach makes it easier to provide various types of access to Internet-based educational resources. In addition to various search capabilities, for example, dynamic browsable lists can be generated directly from the backend database based on the relationships between entities: including the lists by subject, keyword, creator, contributor, publisher, coverage, etc.

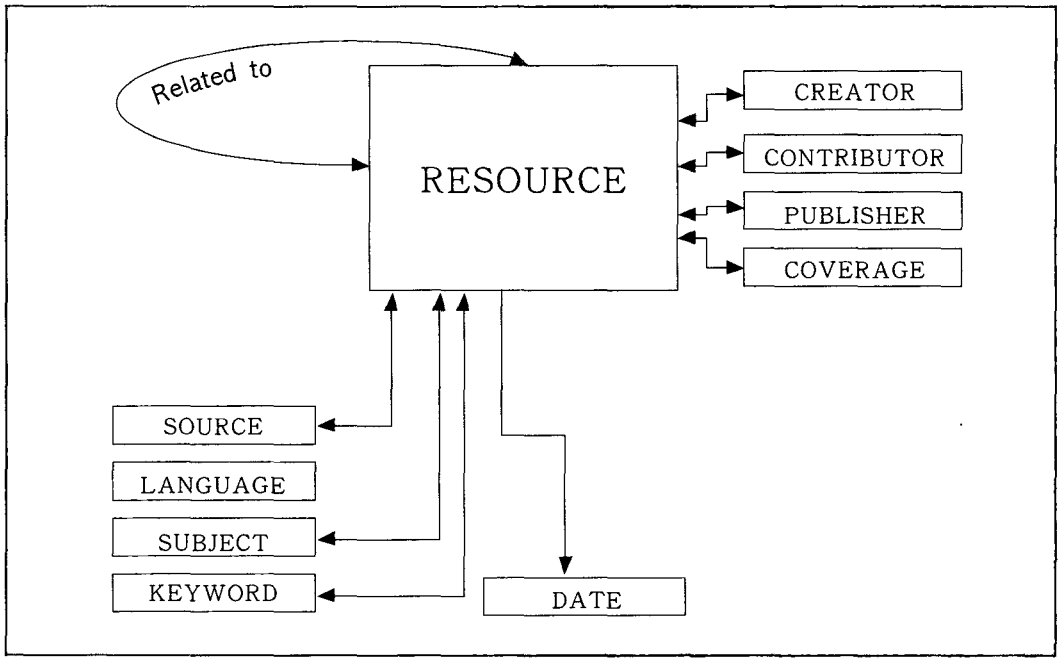
3. DUBLIN CORE CONCEPTUAL AND LOGICAL SCHMAS

A conceptual database design approach to the GEM metadata is desirable because it:

- enables the GEM IR system to group the retrieved sets of educational resources based on relationships among entities;
- helps us to focus our attention on clear understanding of entities and relationships associated with educational resources independent of particular implementations;
- allows us to implement the conceptual schema using different technologies so we can compare the differences in performance; and
- gives us flexibility and modularity so that entities and relationships may be easily added and removed as we find out what metadata elements are really needed by users.

The following is a conceptual schema of the Dublin Core elements.

In the ER model, an entity is denoted by a rectangular box with its name written inside. A relationship between entities is denoted by a line connecting them. The relationship type can be specified either above or the below the line linking the two entities. A relationship's degree indicates the number of associated entities or participants. A unary relationship exists when an association is maintained within a single entity. In Figure 1, the "Related to" relationship is unary. For example, a course within the COURSE entity is a prerequisite for another course within that entity. COURSE has a "Prerequisite" relationship with itself. Such a relationship is also known as a recursive relationship. A



〈Figure 1〉 Conceptual Schema of Dublin core

binary relationship exists when two entities are associated. In Figure 1, the relationship between RESOURCE and DATE is an example of a binary relationship. A ternary relationship exists when three entities are associated. Binary relationships are most common. In fact, to simplify the conceptual design, most higher-order (ternary and higher) relationships are decomposed into appropriate equivalent binary relationships whenever possible. Lastly, an attribute is a specific piece of information which describes a property of an entity. Attributes are not specified in the conceptual schema, but in the logical schema.

In Figure 1, the entity RESOURCE represents lesson plans and other educational materials available on the Internet. A lesson plan (RESOURCE) can be "based on" more than one printed material (SOURCE) and a particular piece of printed material can also be "utilized" in more than one lesson plan (RESOURCE). For example, the print version of "Paradise Lost" can be the basis for many lesson plans. The phrase "based on" is a description of the relationship type and would appear either above or below the connecting line. We omit relationship type descriptions here in order to simplify the conceptual schema. The

arrows and straight lines denote the specific number of entity occurrences associated with one occurrence of the related entity. The arrow denotes many occurrences and the straight line denotes one occurrence. Thus, relationships may be one-to-one, one-to-many, and many-to-many.

A cataloger or creator can assign many subjects to a lesson plan (RESOURCE) in the form of keywords (KEYWORD) or as postings from a controlled subject vocabulary (SUBJECT) and either a subject or a keyword can be assigned to more than one lesson plan (RESOURCE).

The "related to" relationship occurs within the single entity (RESOURCE) and is a recursive relationship. For example, a lesson plan (RESOURCE) can be "associated with" many other related lesson plans (RESOURCE) and a related lesson plan (RESOURCE) can also be "associated with" many other lesson plans (RESOURCE). This is a recursive relationship because the relationship occurs within a single entity (RESOURCE).

A number of relevant dates (DATE) can be associated with a lesson plan (RESOURCE) and a particular date (DATE) is associated with only one lesson plan (RESOURCE). This is an example of one-to-many relationship between the two entities (RESOURCE and DATE). There are three different dates that can be recorded for a particular lesson plan: Creation Date, Modified Date, Placed Online Date.

A lesson plan (RESOURCE) can also be written in more than one language (LANGUAGE) and a particular language (LANGUAGE) can be used in many other lesson plans (RESOURCE).

A lesson plan (RESOURCE) may have many spatial aspects (London, Japan, Korea) and many temporal aspects (13th century, 20th century). These kinds of data are kept in the coverage entity (COVERAGE). There is a many-to-many relationship between the RESOURCE and COVERAGE entities.

Lastly, there can be a number of people (CREATOR, CONTRIBUTOR) or organizations (PUBLISHER) involved in the creation and publication of a particular lesson plan and a particular person (CREATOR, CONTRIBUTOR) or an organization (PUBLISHER) may create more than one lesson plan (RESOURCE).

The following logical schema 〈Table 1〉corresponds to the conceptual schema

〈Figure 1〉 presented above, and describes the attributes defined for each entity. This logical schema is the basis for the relational implementation constructed for the project.

When a many-to-many relationship (e.g., RESOURCE and CREATOR; RESOURCE and KEYWORD; RESOURCE and LANGUAGE, etc.) exists between the two entities, a bridge entity is necessary to keep track of relationships between them. For example, a particular keyword (KEYWORD) can be assigned to many different lesson plans and a lesson plan may have several keywords as index terms. So there has to be a link between the RESOURCE entity and the KEYWORD entity. The "ResourceKeyword" entity in the logical schema below provides this

〈Table 1〉 Logical Schema of the Dublin Core

ENTITY	ATTRIBUTES
Resource	{ResourceID, Format, Description, ResourceType, RightsURL, Title}
Creator	{CreatorID, Name, Affiliation, Contact, Email, Postal, Phone, Fax, HomePageURL}
Contributor	{ContributorID, Name, Affiliation, Contact, Email, Postal, Phone, Fax, HomePageURL}
Publisher	{PublisherID, Name, Affiliation, Contact, Email, Postal, Phone, Fax, HomePageURL}
*RelatedResource	{**ResourceID, RelatedResourceID**, Relationship}
*ResourceCreator	{**ResourceID, CreatorID**}
*ResourceContributor	{**ResourceID, ContributorID**}
*ResourcePublisher	{**ResourceID, PublisherID**}
Coverage	{CoverageID, CoverageScheme, CoverageType, Text}
*ResourceCoverage	{**ResourceID, CoverageID**}
Source	{SourceID, SourceType, FormattedData, Text}
*ResourceSource	{**ResourceID, SourceID**}
Language	{LanguageCode, LanguageScheme, Text}
*ResourceLanguage	{**ResourceID, LanguageCode**}
Subject	{SubjectID, SubjectScheme, SubjectType, Text}
*ResourceSubject	{**ResourceID, SubjectID**}
Date	{DateType, DateScheme, Date, **ResourceID**}
Keyword	{Keyword, KeywordScheme}
*ResourceKeyword	{**ResourceID, Keyword**}

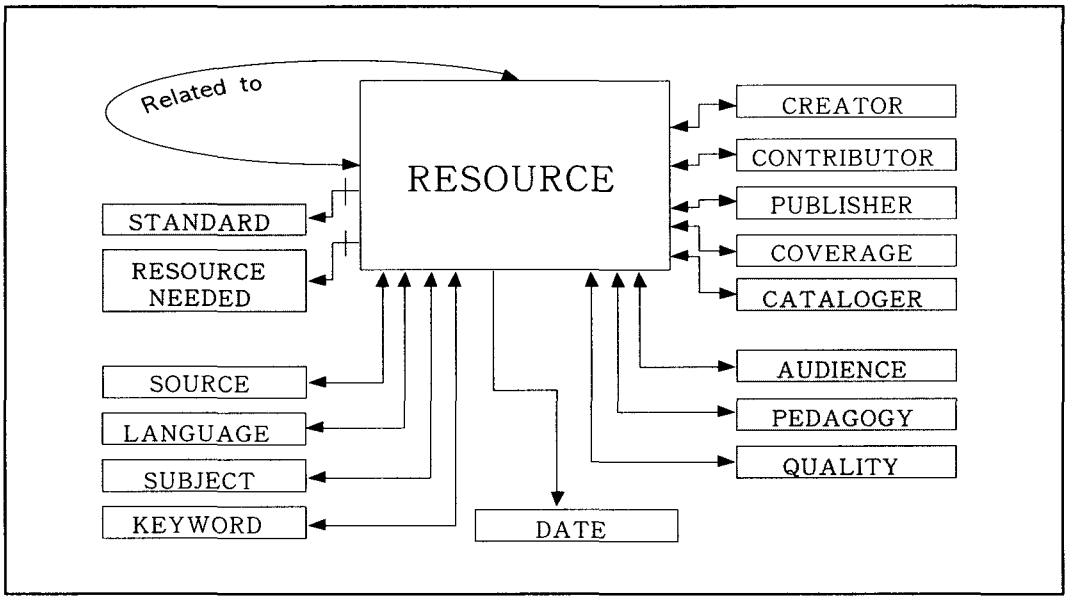
Legend: Primary Key: Underlined, Foreign Key: Boldfaced, * Bridge Tabie

bridging. This kind of bridge entity is needed for each many-to-many relationship.

4. GEM CONCEPUTAL AND LOGICAL SCHMAS

Because of the inability of the Dublin Core element set to adequately describe lesson plans and other Internet-based educational materials for the K-12 teachers, the GEM working group added 8 elements. The following conceptual schema 〈Figure 2〉incorporates the GEM additions to the DC elements.

To the DC Core Element Set, GEM adds elements to make it possible to capture the following domain-specific information about the educational resource being cataloged: (1) whether the target audience for the resource has special needs or forms a descrete demographic group (AUDIENCE), (2) the nature of any identified pedagogical methods employed (PEDAGOGY), (3) indicators of resource quality (QUALITY), (4) academic standards mapped to the resource (STANDARD), (5) the person who cataloged the educational resource (CATLOGER), and (6) other resources necessary to the successful use of the resource (RESOURCE NEEDED).



〈Figure 2〉 Conceptual Schema of GEM Profile (GEM addition is shaded)

Also, information regarding grade or educational level of the target audience of the resource is also captured in the resource entity (RESOURCE).

The following logical schema 〈Table 2〉corresponds to the integrated conceptual schema 〈Figure 2〉 presented above. Attribute names are defined in the context of each entity. The added metadata elements for the GEM are shaded in the logical schema 〈Table 2〉 of the GEM element set.

〈Table 2〉 Logical Schema of the GEM Profile

ENTITY	ATTRIBUTES
Resource	{ResourceID, SID, SDN, GEMversion, Duration, Format, Description, ResourceType, RightsURL, Title, GradeBegin, GradeEnd, Compliance}
Creator	{CreatorID, Name, Affiliation, Contact, Email, Postal, Phone, Fax, HomePageURL}
Contributor	{ContributorID, Name, Affiliation, Contact, Email, Postal, Phone, Fax, HomePageURL}
Publisher	{PublisherID, Name, Affiliation, Contact, Email, Postal, Phone, Fax, HomePageURL}
Cataloger	{CatalogerID, Name, Affiliation, Contact, Email, Postal, Phone, Fax, HomePageURL}
*RelatedResource	{**ResourceID, RelatedResourceID**, Relationship}
*ResourceCreator	{**ResourceID, CreatorID**}
*ResourceContributor	{**ResourceID, ContributorID**}
*ResourcePublisher	{**ResourceID, PublisherID**}
*ResourceCataloger	{**ResourceID, CatalogerID**}
Coverage	{CoverageID, CoverageScheme, CoverageType, Text}
*ResourceCoverage	{**ResourceID, CoverageID**}
Source	{SourceID, SourceType, FormattedData, Text}
*ResourceSource	{**ResourceID, SourceID**}
Language	{LanguageCode, Text, LanguageScheme}
*ResourceLanguage	{**ResourceID, LanguageCode**}
Subject	{SubjectID, SubjectScheme, SubjectType, Text}
*ResourceSubject	{**ResourceID, SubjectID**}
Date	{DateType, DateScheme, Date, **ResourceID**}
Audience	{AudienceID, AudienceScheme, AudienceType, Text}
*ResourceAudience	{**ResourceID, AudienceID**}
Pedagogy	{PedagogyID, PedagogyType, Text}
*ResourcePedagogy	{**ResourceID, PedagogyID**}
Quality	{QualityID, QualityScheme, Authority, Scale, Category, DetailURL}
*QaulityIndicator	{Criteria, Value, **QualityID**}
*ResourceQuality	{**ResourceID, QualityID**}
Keyword	{Keyword, KeywordScheme}
*ResourceKeyword	{**ResourceID, Keyword**}
*ResourceNeeded	{Resource, ResourceID}
Standard	{StandardID, Authority, Correlation, Text, StandardScheme, Code, Topic, Grade, Main, Subordinate, **ResourceID**}

Legend: Primary Key: Underlined, Foreign Key: Boldfaced, * Bridge Table

5. GEM SYSTEM IMPLEMENTATION

A prototype system was developed using the conceptual and logical schema presented above.

Microsoft active server page (ASP) technology and Microsoft Internet Information Server were employed to provide users with Web interfaces to the GEM metadata database. The ASP technology lets one work with any ODBC (Open DataBase Connectivity) compliant database, for example, Oracle, Sybase, SQL Server, Access, etc. The GEM project selected Sybase as the backend database system.

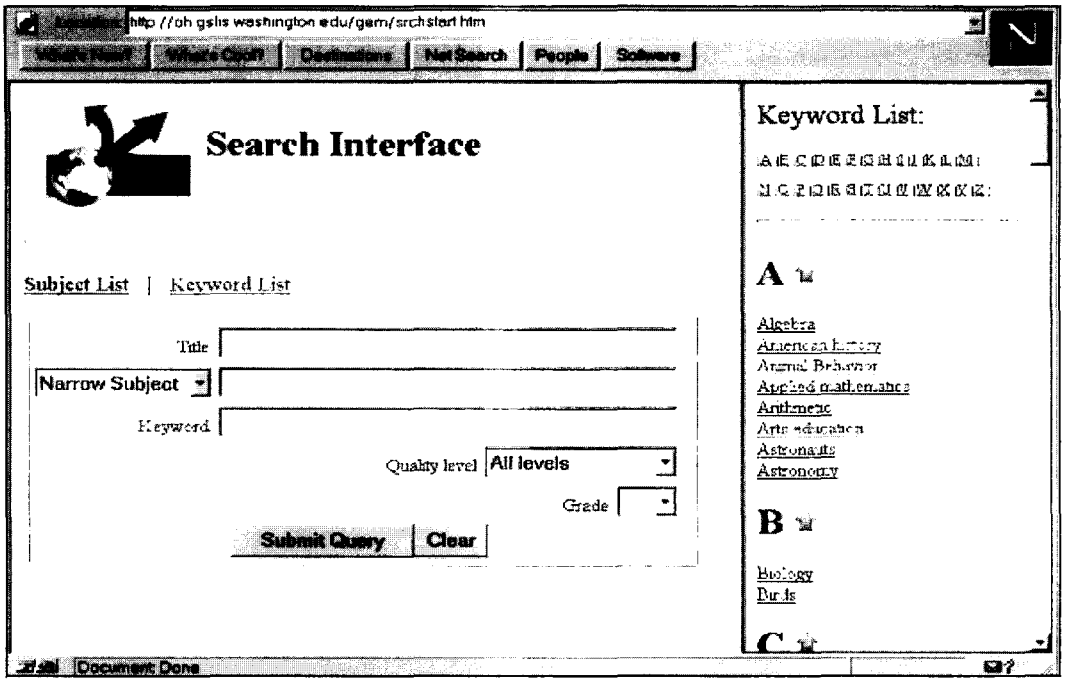
The following is a screen shot of the GEM search interface with provision for browsing an alphabetical listing of assigned keywords on the right. The keyword and subject lists are created dynamically from the backend Sybase database. The user can also access educational resources by browsing assigned subjects in the same manner. In addition, other alphabetical lists (creator, publisher, contributor, coverage, etc.) can be easily generated dynamically if desired. This is one advantage of employing the conceptual data modeling and relational approach to construct the GEM metadata database.

The prototype system currently allows users to search by title, broad and narrow subject terms, keywords, quality level, and grade, but other search criteria can be easily added as needed. It also allows users to browse by keywords or subjects using the right side of the window. The dynamic keyword list is organized A-Z at the top which allows users to choose an alphabetical section to browse. One can come back to the top of the A-Z list instantly by pressing the up-arrow icon 〈Figure 3〉.

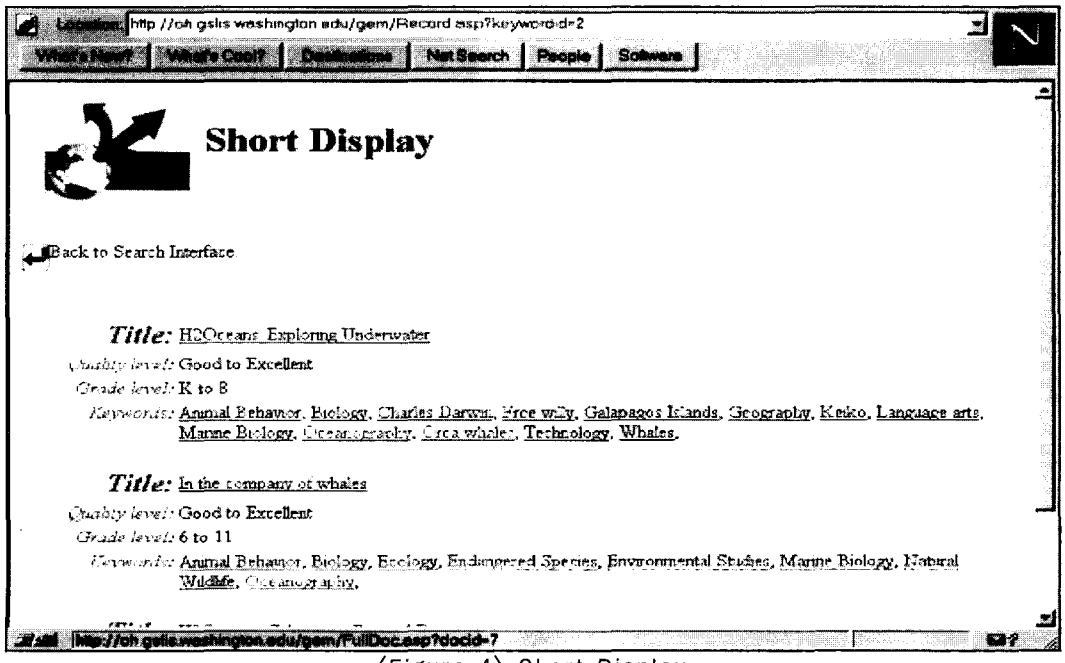
If a user submits a query or selects a keyword or a subject to browse, the following output screen is displayed 〈Figure 4〉:

The title and all the keywords are hypertext links. If the user clicks on one of the keywords listed for a resource, he or she is taken to the short display of all titles to which the clicked keyword has been assigned. If a user clicks a title, the system displays more detailed information (a full GEM description) for the resource. Figure 5 is an example of the full display.

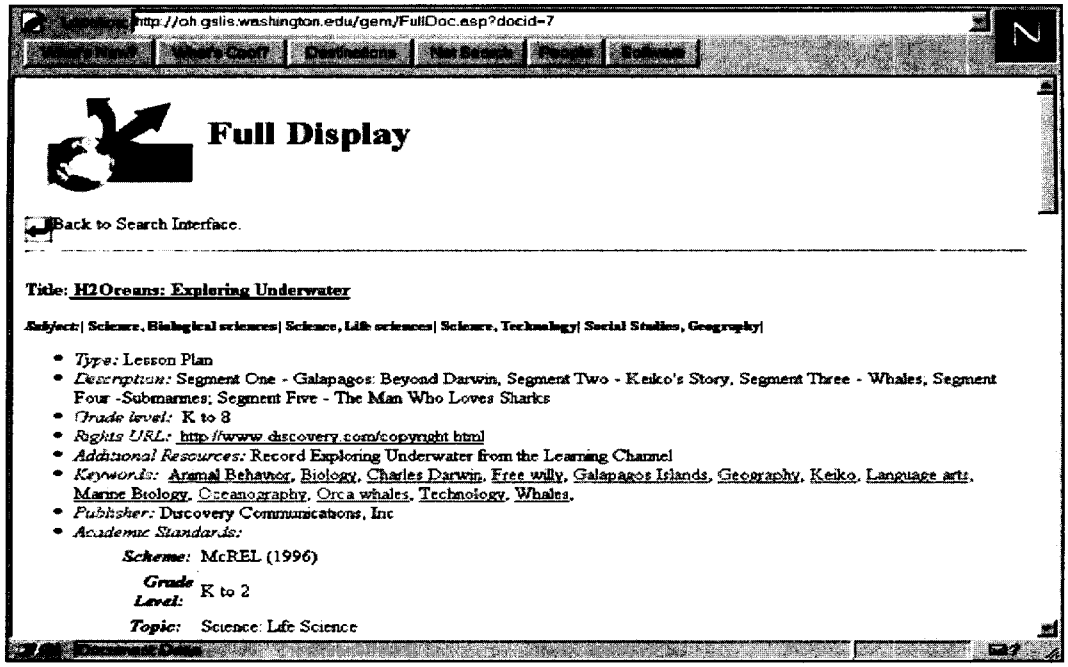
A carefully cataloged resource will provide the user with sufficient information to determine whether or not to visit the actual resource on the remote server. If a user chooses to visit the



〈Figure 3〉 Search and Browsing Interface



〈Figure 4〉 Short Display



〈Figure 5〉 Full Display

resource, all he or she needs to do is click on the title in the full GEM display.

6. HARVESTING AND DATA POPULATION

Data are being populated for the GEM database in the following manner:

- The consortium members catalog lesson plans and the Internet-based educational materials that they own using the GEMCat cataloging module which creates GEM compatible meta tags.
- A HARVEST program (software) collects all these HTML meta tags from consortium member sites and outputs a standard ASCII file format agreed by the consortium members.
- A PARSER program accepts this ASCII file as input and enters meta tags automatically into appropriate relational tables. As a test, the PARSER processed 175 records successfully. In excess of 10,000 GEM records exist for parsing into the GEM database.

7. CONCLUSION

The use of metadata to enhance networked information discovery and retrieval is an area of growing interest as, with the exponential growth of the World Wide Web, the resource discovery and retrieval grows more problematic. At the same time, as President Clinton's policy focus on education exemplifies, the need for access to web-based materials by the K-12 teachers in the USA is also growing at a rapid pace. GEM seeks to meet the needs of educators through three primary roles:

(1) the 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 to Internet-based educational materials;
(2) the development of a GEM-based union catalog of educational materials on the Internet; and
(3) the development of Web-based interfaces to GEM database of educational materials on the Internet and allowing others to develop interfaces to the GEM union catalog.

The GEM project integrates cataloging, data modeling, and the Web technology to manage ever-increasing Internet resources. The GEM metadata database can be very useful to Korean teachers if the K-12 curriculum is designed to encourage creative thinking. The approach and principles employed in this project can be applied to managing various Internet resources.

Acquiring and managing quality information will be the key ingredient in succeeding whatever we do in the information age.

REFERENCES

Agosti, M. (1989). Towards data modeling in information retrieval. Journal of Information Science, 15, 307-319.

Elmasri, R. & Navathe, S.B. (1994). Fundamentals of database systems. (2nd ed.). Redwood City, CA: The Benjamin/Cimmings Publishing Company, Inc.

Ingwersen, P. (1992). Information retrieval interaction. London: Taylor Graham.

Oh, S. G. (1995). An empirical fact retrieval system: An entity-relationship and relational approach. Ph.D. Dissertation - Syracuse University.

Rob, P. & Coronel, C. (1995). Database systems: Design, implementation, and management. Boyd & Fraser.

URLs:

- Dublin Core Home Page http://purl.oclc.org/metadata/dublin_core
- ERIC Clearinghouse on Information and Technology http://ericir.syr.edu/
- Gateway to Educational Materials (GEM) http://gem.syr.edu
- A GEM Prototype System at the University of Washington http://pange.gslis.washington.edu/gem/srchstart.htm
- U.S. Department of Education http://www.ed.gov
- U.S. National Library of Education http://www.ed.gov/NLE/index.html
- World Wide Wed Consortium http://www.w3.org