Use of Learning Object Vocabulary in GEM Queries

Jian Qin
School of Information Studies
Syracuse University
Tel: +1 315 443 5642
Fax: +1 315 443 5806
jqin@syr.edu

Javier Calzada Prado
Departamento de Biblioteconomía y Documentación
Universidad Carlos III de Madrid
Tel: 916248600
fcalzada@bib.uc3m.es

Abstract:

Metadata applications have developed local controlled vocabulary to meet information needs of users, but little is known about what vocabularies users use in searching for information. This paper reports the findings from an analysis of a digital library's query log. The analysis addresses questions of to what extent users use controlled vocabulary in resource discovery and what non-controlled vocabulary users use in their resource discovery. The authors discuss what is missing between the controlled and non-controlled vocabulary and how we can integrate user query terms into a learning object vocabulary for improving learning object representation and discovery.

Keywords:

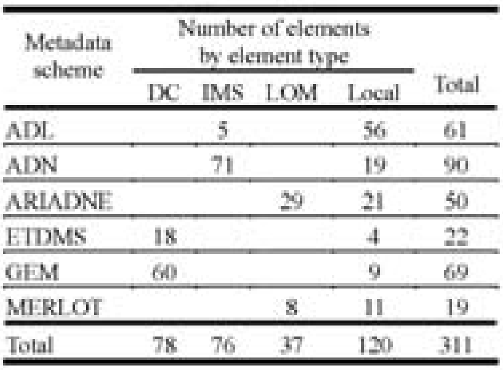
Educational metadata, Controlled vocabulary, Query log analysis

1. Educational Metadata Elements and Vocabulary

The Semantic Web (SW) puts the development of ontologies as the central task. The RDF model and Web Ontology Language (OWL) standards impose a challenge to developers, that is, capturing the semantics and representing it in the SW model and language. For the educational digital library community, this challenge translates into the semantics for metadata elements and element values.

In a recent study of learning object metadata application profiles, Godby (2004) points out that large variation exists in elements present in metadata application profiles as well as the naming differences in these elements. Due to the lack of subject classification schemes for learning objects, resource "discovery strategies will probably be restricted to known-item search" [1]. Her observation echoes a research we conducted at about the same time of her survey, in which we examined the content elements in six metadata application profiles and compared their conformance with Dublin Core, the Learning Object Metadata (LOM) standard, and the IMS LOM binding [2]. Table 1 contains the data about the application profiles and their adoption of metadata standards, as well as the number of local elements. The focus was on the vocabulary used in these profiles, among other things. One of our observations was the paucity of vocabulary related to learning objects. This reflects in the low percentages of learning related elements and the vocabulary available for these elements (Figure 1).

Table 1. Number of elements by element type and schema



Metadata scheme	Number of elements by element type				Total
	DC	IMS	LOM	Local	
ADL		5		56	61
ADN		71		19	90
ARIADNE			29	21	50
ETDMS	18			4	22
GEM	60			9	69
MERLOT			8	11	19
Total	78	76	37	120	311

A large part of the reasons for insufficient representation of learning related content is the lack of a vocabulary. Although some metadata applications have developed their own controlled vocabularies, the

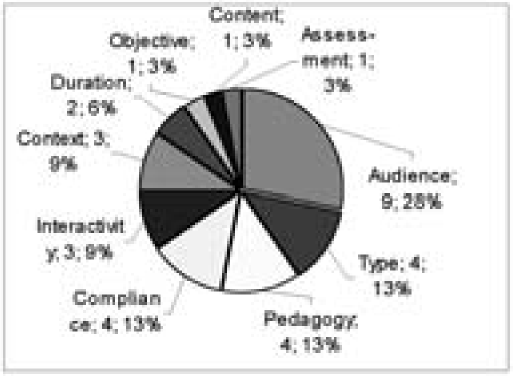


Figure 1. Educational metadata elements

scope and facets covered are often limited to local needs. Moreover, little is known what vocabulary users use in searching for information, which often leaves learning object creators and catalogers to struggle with choosing the right vocabulary or simply give up all together. This paper reports the findings from an analysis of a digital library's query log. The analysis addresses three questions: To what extent did users use controlled vocabulary in resource discovery? What non-controlled vocabulary was used in their resource discovery? How can we integrate user query terms into a learning object vocabulary for improving learning object representation and discovery?

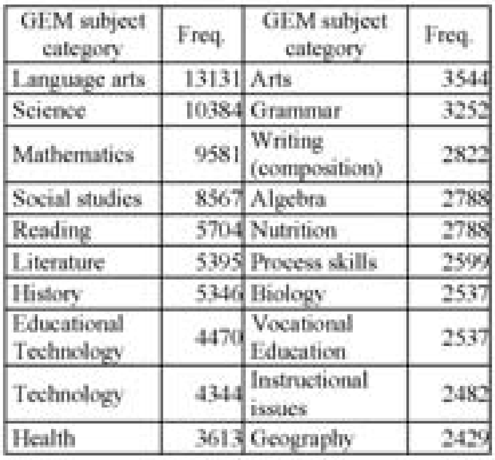
2. Data

In fall 2003, we collected query logs from the Gateway to Educational Materials (GEM). The query log data covered a four-month period in 2003 (February, March, April, and August). This period generated 411,898 queries. We wrote SQL programs to parse the queries in order to obtain a master list of query components. The master list was then cleaned and coded for counting frequencies of terms and query fields, which resulted in 1,044,043 query components that consisted of grade numbers and/or terms, topical keywords, book/movie titles, person names, geographical names, organization names, year, and other categories. There were also a good number of meaningless components, such as punctuations and unrecognizable codes and symbols, which were ignored in analysis.

3. Use of GEM Controlled Vocabulary

Table 2 contains the top 20 terms in the GEM controlled vocabulary. The first four subject categories were used most frequently among the top terms. In all 272 GEM subject categories at both first and second level, 206 appeared in the queries. Among the 206, 24 occurred in more than 2,000 times, 20 in between 1,000 and 2,000 times, 98 in between 100 and 1,000 times, and 61 occurred between 1-100 times. We examined the first four subject categories in relation to the number of hits and found a wide range of numbers for this measure. The highest number of hits for Language arts was 14,639 and there were four occurrences with more than 10,000 hits. As much as high numbers of hits appeared frequently, there were also a large number of occurrences of zero hit involving queries containing Language arts. Similar patterns can also be observed in other GEM subject categories that occurred highly frequently.

Table 2. Top 20 terms in GEM controlled vocabulary

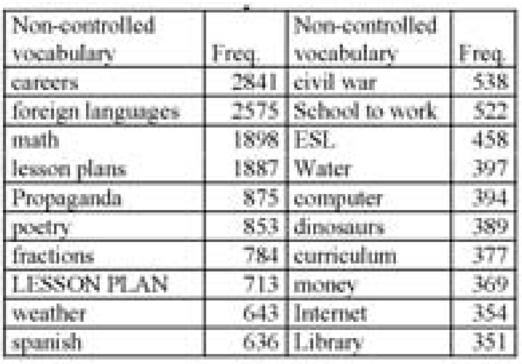


GEM subject category	Freq.	GEM subject category	Freq.
Language arts	13131	Arts	3544
Science	10384	Grammar	3252
Mathematics	9581	Writing (composition)	2822
Social studies	8567	Algebra	2788
Reading	5704	Nutrition	2788
Literature	5395	Process skills	2599
History	5346	Biology	2537
Educational Technology	4470	Vocational Education	2537
Technology	4344	Instructional issues	2482
Health	3613	Geography	2429

4. Use of Non-GEM Vocabulary

Compared to the GEM controlled vocabulary, the top 20 non-GEM subject categories present a few interesting perspectives. The first observation is that the top terms had much lower frequencies than the top GEM subject categories did. Second, forms of a term caused overlap in semantics, e.g., "math" is equivalent to Mathematics in the controlled vocabulary, and "foreign language" is a branch of Language arts. The different forms in "lesson plan" resulted in two different frequency numbers in Table 3, so is it for the poetry and fractions, which are part of the first level concepts Language arts and Mathematics respectively. Another phenomenon is the large number of unique terms with smaller frequencies—over 800,000 occurrences for non-GEM terms while only 177,906 for GEM subject categories.

Table 3. Frequency distribution of non-GEM controlled vocabulary



Non-controlled vocabulary	Freq.	Non-controlled vocabulary	Freq.
careers	2841	civil war	538
foreign languages	2575	School to work	522
math	1898	ESL	458
lesson plans	1887	Water	397
Propaganda	875	computer	394
poetry	853	dinosaurs	389
fractions	784	curriculum	377
LESSON PLAN	713	money	369
weather	643	Internet	354
spanish	636	Library	351

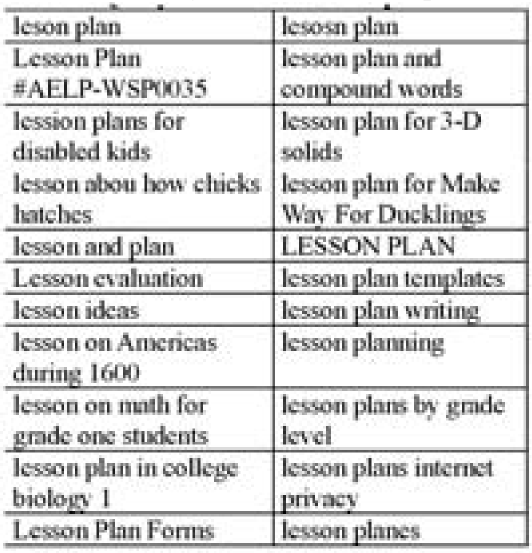
5. The Missing Link

While the use of both GEM and non-GEM vocabulary has complex causes and requires more exploration, we observe a missing link between the controlled and non-controlled vocabularies. Let us use an example from the

non-GEM terms in the query log to demonstrate metadata application provides only a list of

what the missing link is. Table 4 contains some query terms regarding lesson plans:

Table 4. Query terms for "Lesson plans"



leson plan	lesosn plan
Lesson Plan #AELP-WSP0035	lesson plan and compound words
lession plans for disabled kids	lesson plan for 3-D solids
lesson abou how chicks hatches	lesson plan for Make Way For Ducklings
lesson and plan	LESSON PLAN
Lesson evaluation	lesson plan templates
lesson ideas	lesson plan writing
lesson on Americas during 1600	lesson planning
lesson on math for grade one students	lesson plans by grade level
lesson plan in college biology 1	lesson plans internet privacy
Lesson Plan Forms	lesson planes

Regardless of the typos in these examples (which the technology has the capability of dealing with), we can detect two types of patterns among them:

1) *Linguistic patterns:* lesson plan + preposition phrase, lesson plan + noun phrase, lesson + nouns, etc.
2) *Semantic patterns:* lesson plan + historical topic, lesson plan + language arts topic, lesson plan + creation tools topic, lesson plan + specific materials (e.g., a book or movie title), lesson plan + audience, etc.

Most search systems allow users to combine multiple terms and limit search fields when enter a query, but this is often done without a sufficient vocabulary support. By "sufficient vocabulary support," we mean that when a user enters "lesson plan templates," the system knows "lesson plan" is a learning object (or instructional object) and "templates" is a creation tool for lesson plan. Or, in the case of "lesson plan for Make Way For Ducklings," the system knows that it is the title of a book. From an ontological point of view, what prevents the system from making such judgments is the absence of a knowledge base (built from an ontology) that contains concepts and relationships among the concepts. By linking instances with concepts, we can create inference rules to help the system make judgments in situations like lesson plan queries. Most controlled vocabulary developed for a terms with two or three levels. As the size of digital library collection grows and the content becomes more diversified, simple list of controlled vocabulary would lose its advantages quickly. The exceptionally large amounts of search results for GEM controlled vocabulary are a warning sign for this danger. We are further analyzing the GEM query log and generalizing more patterns from the user query terms. Our goal is to use the result from query log analysis to enhance the learning object vocabulary built in the last two years (http://web.syr.edu/~jqin/LO/LOV2/) and to test it with the Web Ontology Language (OWL), a Semantic Web standard developed at the World Wide Web Consortium (W3C). The capability of OWL in representing class relationships and constraints will allow the learning object vocabulary to fill in the missing link between the controlled vocabulary and non-controlled terms.

6. Conclusion

The query log analysis provides valuable insights into what terms users used in resource discovery as well as a rich source for building a well-defined knowledge system. As part of a larger project the findings will be incorporated into the learning object ontology and used for developing ontology-based metadata tools and for browsing, retrieving, and post-searching processing.

References

1. C. J. Godby. What do application profiles reveal about the learning object metadata standards?

ARIADNE, 2004, 41, http://www.ariadne.ac.uk/issue41/godby/

2. J. Qin and J.C. Prado. The semantic and syntactic model of metadata. Working paper. http://web.syr.edu/~jqin/papers/Metadata_model.pdf