2001

Converting a Controlled Vocabulary into an Ontology: the Case of GEM

Jian Qin
Syracuse University, jqin@syr.edu

Stephen Paling
Syracuse University, swpaling@syr.edu

Follow this and additional works at: https://surface.syr.edu/istpub
Part of the Library and Information Science Commons

Recommended Citation

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.
Converting a controlled vocabulary into an ontology: the case of GEM

Jian Qin & Stephen Paling
School of Information Studies
Syracuse University
Syracuse, New York, USA

Abstract
The prevalence of digital information raised issues regarding the suitability of conventional library tools for organizing information. The multi-dimensionality of digital resources requires a more versatile and flexible representation to accommodate intelligent information representation and retrieval. Ontologies are used as a solution to such issues in many application domains, mainly due to their ability explicitly to specify the semantics and relations and to express them in a computer understandable language. Conventional knowledge organization tools such as classifications and thesauri resemble ontologies in a way that they define concepts and relationships in a systematic manner, but they are less expressive than ontologies when it comes to machine language. This paper used the controlled vocabulary at the Gateway to Educational Materials (GEM) as an example to address the issues in representing digital resources. The theoretical and methodological framework in this paper serves as the rationale and guideline for converting the GEM controlled vocabulary into an ontology. Compared to the original semantic model of GEM controlled vocabulary, the major difference between the two models lies in the values added through deeper semantics in describing digital objects, both conceptually and relationally.

Introduction
Ontologies, as a form of knowledge representation, are defined as a systematic account of existence, a specification of a conceptualization (Gruber, 1993a). They represent a domain of discourse, and allow for relationships such as the definition of classes, relations, and functions. Despite their high level of specification of these classes and relationships, ontologies also allow a great deal of flexibility. Among the more flexible possibilities are the sharing and reuse of ontologies, and the ability to accommodate varying descriptive terms. In the context of metadata for digital library objects, ontologies are specified in the form of definitions of representational vocabulary. Conventional representational vocabularies used in libraries include cataloguing codes as embodied in MAchine Readable Cataloguing (MARC), thesauri or subject heading lists, and classification schemes. There is a clear functional distinction between such vocabularies. Thesauri and classifications, on the one hand, are used to represent the subject content of a book, journal article, a film, or any form of recorded knowledge, though they also contain a limited number of terms for document formats and genre. Semantic relationships among different concepts are reflected through broader terms, narrower terms, and related terms in thesauri, and a hierarchical structure in classification schemes. On the other hand, because a thesaurus is a subject content rich vocabulary, it does not handle descriptive data (title, author, publisher, etc.) object well. Libraries use separate representational vocabularies for the descriptive data such as Anglo-American Cataloging Rules 2nd ed. (AACR2) to meet the need for descriptive representation.

The prevalence of digital information resources raised several issues for conventional representational vocabularies. Firstly, digital objects, as those in the Gateway to Educational Materials (GEM, http://www.thegateway.org/) collections, encompass multiple dimensions of characteristics and such characteristics often play important roles for users in search for precise information in an efficient manner. For example, a lesson plan of arithmetic for fifth graders may contain information on pedagogy, class activities, math games, test samples, educational standards for fifth grade mathematics curriculum; as a digital object, it may include text, images, audio/video clips; and it may be a parent object of a collection of
child objects or a child object of a parent object. A conventional cataloguing code will be inadequate to
describe these details in a lesson plan, for many of these elements do not even exist in the vocabulary.
Secondly, different metadata schemes are used to represent digital objects in different domains. When an
information repository has heterogeneous digital objects, interoperability becomes the first obstacle to
provide access to these resources. A library catalogue system may not be able to translate a record from a
museum or a library MARC record can not be recognized by an Internet-based information gateway
system. Finally, more and more systems today use agents and other intelligent applications to improve
information search performance. While such capabilities are hard to implement directly within the metadata
records, it seems to make more sense if we create agents that bear the concept models with deeper
semantics. Such effort does not have to start from scratch, but rather, can make use of the knowledge
structure and vocabularies in classifications and thesauri and add values to these representations.

Ontologies are in a right position to address these issues. The fundamental difference between an ontology
and a conventional representational vocabulary is the level of abstraction and relationships among concept.
In this paper, we report the conversion of the GEM controlled vocabulary into an ontology, which is the
first step in a larger project towards an ontology-driven information gateway for educational materials. The
current study discusses the preliminary planning and steps involved in converting the existing GEM
vocabulary to an ontology. The purpose of the conversion is not only to reduce the duplication of effort
involved in building an ontology from scratch by using the existing vocabulary, but also to establish a
mechanism for allowing differing vocabularies to be mapped onto the ontology. Soergel (1999) drew
connections between ontologies and other methods of classification, pointing out that ontological and
lexical structures underpin much scientific and scholarly work. He criticises scholars in scientific
communities for not communicating more effectively with each other about their classifications, and
laments the fact that the intellectual capital of classification schemes and thesauri is often ignored. This
project also attempts to avoid this problem. It uses the existing capital represented by the GEM scheme, and
makes it available not just for automated processing at GEM, but also to other communities who choose to
take up the structure of the ontology. The following sections include an introduction to GEM's background,
a literature review of formal ontologies and applications, conversion considerations, and the constructs of
GEM ontology.

Background of GEM

GEM is an initiative of the US Department of Education's National Library of Education (NLE) to expand
educators' capability to access Internet-based lesson plans, curriculum units and other educational materials
(Sutton, 1999). There are over 14,500 metadata records in the GEM database as of this writing. Its
resources come from more than 100 collections, including AskERIC Virtual Library, Math Forum,
Microsoft Encarta, North Carolina Department of Public Instruction, and U.S. Department of Education.
GEM acquires indexing records through distributed cataloguing that uses a proprietary cataloguing system.
The software contains an interface that allows GEM cataloguers to enter metadata under the 15 Dublin
Core Metadata Elements (short for Dublin Core or DC) from the Online Computer Library Center (OCLC)
and 8 local elements that are designed specifically for the resources covered by GEM (Table 1). Elements
with an asterisk in Table 1 describe the subject content of a resource, among which the Subject element
allows two types of vocabulary: the GEM controlled vocabulary and the keywords selected by metadata
creators. Box 1 shows a typical GEM cataloguing record. Note that the Description, Subject, and Title
elements contain information on what the resource is about. In the Subject element, GEM requires metadata
creators to include a level one subject from the GEM Subject Vocabulary (Morgan, 1999). Other elements
such as Audience, Grade, and Pedagogy include information on what and whom the resource is for, which
can play important roles in narrowing down a search when querying the database. Box 1 shows an example
of a GEM metadata record, where the two-levelled subject terms were selected from the GEM Controlled
Vocabulary, and the keywords extracted by metadata creators from the resource.
Elements from Dublin Core | Elements from GEM
--- | ---
**DC.Contributor** | **GEM.Audience**
**DC.Coverage** | **GEM.Cataloging**
**DC.Creator** | **GEM.Duration**
**DC.Date** | **GEM.EssentialResources**
**DC.Description** | **GEM.Grade**
**DC.Format** | **GEM.Pedagogy**
**DC.Identifier** | **GEM.Quality**
**DC.Language** | **GEM.Standards**

Table 1: Data elements for GEM indexing records

* Indicates elements that are used to tell what a resource is about.

| **Title:** Eyes on the Sky, Feet on the Ground: Hands On Astronomy Activities for Kids
resource contains hundreds of fun explorations into astronomy as a classroom tool for learning how to theorise, experiment, and analyze data. The activities are fully illustrated and contain detailed, step-by-step instructions as well as suggested discussion topics. |
| **Pedagogy:** Teaching Method: Hands-on learning
Multiple activities

| **Resource Type:** Collection
**Format:** text/HTML
**Relation:**
**Rights Management:**
Cost: free
**Date:** Record Created: 2000-04-13T01:45:39-5:00
**Publisher:**
Name: Harvard-Smithsonian Center for Astrophysics, High Energy Astrophysics Division
Role: onlineProvider
Homepage: http://hea-www.harvard.edu/

| **GEM Subjects:** Science--Astronomy
Science--Physics
Science--Space Sciences
Science--Technology
**Keywords:** Orbit, Time, Calendar,
Maps, Solar system, Moon, Rotation
**GEM Audience:**
Tool For: Teachers
Beneficiary: Students
**cataloguing Agency:**
Name: GEM
Version: 3.2
Homepage: http://www.geminfo.org

Box 1. A sample GEM metadata record

The GEM controlled vocabulary is resource-oriented (Figure 1), i.e., it defines the categories (classes) and subcategories (subclasses) of information that are included in educational resources. Among other classes that describe the many physical/medium attributes for a resource, the subject class encompasses a typical curriculum in the US elementary and secondary schools, which has a two level subdivisions—subclass and concept names. Currently, the GEM controlled vocabulary is used as a metadata scheme for representing the multiple dimensions for digital educational objects. The GEM system provides access to these multi-dimensions by using Boolean operators through the search interface. Though this controlled vocabulary integrates the descriptive data elements with subject categories, it is difficult to achieve a higher level of representation of these digital objects (such as games, test samples, and classroom activities in a math lesson plan) and the relationships among and in the digital objects. This difficulty manifests a need for a more systematic abstraction that will explicitly specify the concepts as well as the relationship between these concepts.
Formal Ontologies and Ontological Applications

Researchers in the areas of knowledge representation have debated the nature of ontologies in the last few decades. In general, "an ontology is an explicit specification of a conceptualization" (Gruber, 1993). While Gruber's definition is widely cited, there have been questions about the meaning of conceptualization. Guarino (1997) argues that an ontology is "not a specification of a conceptualization, but a (possibly incomplete) agreement about a conceptualization." He further states that:

An ontology is a logical theory accounting for the intended meaning of a formal vocabulary, i.e. its ontological commitment to a particular conceptualization of the world. The intended models of a logical language using such a vocabulary are constrained by its ontological commitment. An ontology indirectly reflects this commitment (and the underlying conceptualization) by approximating these intended models. (Guarino, 1998)

When ontologies are used at system development time, they get transformed and translated into an information system component, thus reduce the costs of conceptual analysis in system development; when used at run time, ontologies enables communication between software agents (ontology-driven communication). Guarino also demonstrates in his paper how ontologies affect the architecture of information systems. Thus, to be more accurate, an ontology is a partial specification of a shared conceptualization, to be used for formulating knowledge-level theories about a domain (Domingue and Motta, 1999).

Various formal analysis methods provide a wide range of choices for modelling the "world." One of the main purposes of formal analysis is to investigate the logical properties of concepts and their relations. Goasdoue and Reynaud (1999) used semantic concept skeletons to identify and model relevant concepts. The process involves the use of the Entity-Relationship model to generate semantic concepts, which, in turn, is used to group entities and relationships. Conceptual structures can also be modelled as a hierarchical network in the form of a mathematical lattice (Priss, 1998). The main purpose of formal concept analysis is to define a formal model in which there exists a set of formal objects \( G \), a set of formal attributes \( M \), and a relation \( I \) between \( G \) and \( M \). For example, in the context of a lesson plan, \{arithmetic, test\} = \{accelerated\} is a formal concept. A formal concept does not necessarily define a topic (i.e., subject); it may well be used to specify a process. The University Michigan Digital Library (UMDL) ontology
Ontology projects. The conversion of a
This study maintains that while these criteria serve as useful guides, they do not apply uniformly to all
ontology development. These criteria are:

Different kinds of ontologies are developed to serve different levels of generality. Guarino (1998) proposes
that "top-level" ontologies describe all general concepts such as space, time, matter, object, event, action,
and so forth; domain ontologies and task ontologies describe, respectively, the vocabulary related to a
generic domain (like medicine, or automobiles) or a generic task or activity (like diagnosing or selling), by
specializing the terms introduced in the top-level ontology; and application ontologies describe concepts
depending both on a particular domain and task, which are often specializations of both the related
ontologies. The ontology library at Stanford University Knowledge Systems Laboratory (http://www-
ksl.stanford.edu/) contains many examples of these kinds.

The ultimate goal of developing ontologies is to enhance information representation and information
retrieval systems. The Web provides a perfect test-bed for ontology-driven applications ranging from
information extraction to information retrieval. Embley, et al. (1998) built an ontology-based system for
extracting information from data-rich, unstructured documents. They define "data-rich" documents as those
having a number of identifiable constants such as dates, names, account numbers, ID numbers, part
numbers, times, and currency values. With such a narrow ontology breadth, an application ontology can be
modelled by object-oriented analysis. The application ontology was mainly used to generate SQL tables. In
this sense, their application ontology functions as a database schema. Another system dealing with the
similar type of documents is OntoSeek developed by Guarino, et al. (1999), which targets content-based
information retrieval from online yellow pages and product catalogues. Unlike Embley et al.'s information
extraction system, they used linguistic tools such as WordNet (http://www.cogsci.princeton.edu/~wn/) and
Penman to avoid building the ontology from scratch. OntoSeek utilizes ontology not only at the encoding
level but also the end-user level. Ontology has also been used to build Web agents (Luke, et al., 1997) to
add semantics to HTML in order to define attributes and relationships in machine-readable form, and to
map Web sites for information exploration (Zhu, et al., 1999).

The above discussion of formal ontologies and ontological applications reveals an intrinsic relationship of
ontological commitments with thesauri and classifications that have long tradition in the library and
information science communities. While classifications play an increasingly important role in knowledge
representation and discovery (Kwasnik, 1999), a practical question for converting a controlled vocabulary
into an ontology is: What is the added value of an ontology compared to the knowledge representation
vocabulary used in libraries and information industries? The answer is given in the next section.

**Design Considerations**

Gruber (1993b) discusses ontologies as design artefacts, and outlines a set of design criteria to guide
ontology development. These criteria are:

- **Clarity**: Definitions should be objective, free of social or computational context, and documented
  with natural language.
- **Coherence**: The defining axioms should be logically consistent and sanction inferences consistent
  with the definitions.
- **Extendibility**: The ontology should anticipate new used of the shared vocabulary.
- **Minimal encoding bias**: Encoding bias occurs when a choice of representation is made simply for
  the convenience of notation. The conceptualization should be specified at the knowledge level
  without depending on a particular symbol-level encoding.
- **Minimal ontological commitment**: An agent is said to commit to an ontology when its actions
  are consistent with the ontology’s definitions. An ontology should make a minimal number of
  claims about the world being modelled. (Gruber, 1993b).

This study maintains that while these criteria serve as useful guides, they do not apply uniformly to all
ontology projects. The conversion of an existing vocabulary brings with it certain assumptions that vary
from the criteria. For example, we diverge from the views of clarity being expressed. Because we are converting terms already defined in a social context, we do not believe that they can be defined independently of that context. The terms define different facets of Internet resources used for education, and the definitions are constructed in service of the people who use the resources. Because the facets are relatively independent, they will not be connected by a large number of axioms. Taking the above criteria into account, the ontology must have the following added values compared to the existing vocabulary:

1. Higher levels of conception of descriptive vocabulary;
2. Deeper semantics for class/subclass and cross-class relationships;
3. Ability to express such concepts and relationships in a description language; and
4. Reusability and "share-ability" of the ontological constructs in heterogeneous systems.

**Higher levels of conception.** Current GEM controlled vocabulary represents digital objects in the sense of its "whole", i.e., similar to a library cataloguing record that describes the object by its overall characteristics. But from a GEM user survey, we found that blocks within a digital object whole are frequently the ones that users are looking for. When we represent the information contained in a lesson plan, simply labelling it as "lesson plan" would not give users much valuable information on its test samples, interactive games, in-class activities, or the state standard it relates to. An ontology would enrich the representation vocabulary for such information blocks.

**Deeper semantics for class/subclass and cross-class relationships.** Converting an existing vocabulary into an ontology does not simply mean adding more classes or subclasses to it. The relationships have to be explicitly specified between these classes. When we define a math lesson plan, it is important to indicate that it has game(s) or relates to standard(s), and how these contents are related.

**Description language for concepts and their relationships.** A formal description language is based on Description Logic (DL) that is used to describe situations using various kinds of individuals (as instance in a collection of objects), related by roles (as relation), and grouped into concepts (a primitive specified elsewhere) (Brachman et al., 1999). The real significance of a formal description language is for the representation to have the ability to reason about descriptions.

**Reusability and "share-ability" of the ontology.** If teachers, administrators, and researchers are to share educational knowledge, they must share more than a common vocabulary of previously identified terms; they must delineate the relationships among the objects in the world to which the terms refer. Ontologies describe formal descriptions of objects in the world, the properties of those objects, and the relationships among them. Hence, the value that they add to existing vocabularies such as thesauri or classification schemes is additional information that defines how objects can be classified and related to one another. In addition to these value-added considerations, there are considerations relating to methodology for the conversion. Two aspects of any knowledge representation process have direct implications to this study: structuring and linguistic (Biebow and Szulman, 1999). The structuring aspect refers to the way a concept has been introduced into a model, e.g., bottom-up or top-down, and the linguistic aspect pertains to the linguistic accuracy of a concept. In GEM's situation where a structure is already in place, it makes it a natural choice to use the top-down approach to building the GEM concept model. For the linguistic accuracy, GEM controlled vocabulary is adopted whenever possible. If a new term or concept needs to be added, the educational thesaurus would be consulted.

**The GEM Ontology**

We used the Ontolingua system (http://www-ksl-svc.stanford.edu:5915/) at Stanford's Knowledge Systems Laboratory (http://www-ksl.stanford.edu/) to create the ontology. Ontolingua provides a development environment for ontology construction. In addition to editing tools, it allows the creation of modular, combinable ontologies.

Because the GEM controlled vocabulary was already in existence, our task was simplified. The following schematic shows the proposed GEM ontology:
A slot represents an attribute of a class in an ontology. For example, a lesson plan class might have title, author, subject, audience, grade level, format, and language as slots. A slot has value and value type. Teacher is one of the values for slot Audience, Mathematics-Geometry is a value for slot Subject, and so forth. These values are typically string type. Theoretically, the value for slots may be a class (such as Audience that may be modelled as a class), a Boolean operator, a number, or a symbol. Apart from the taxonomic relationships between classes, a full-fledged ontology attaches properties or slots to classes, where a subclass inherits the slots from all of its (direct or indirect) super-classes. In a class hierarchy, inheritance proceeds simultaneously along multiple paths. Using a math lesson plan for high school students as an example, the diagram below demonstrates the semantic model of the ontology:
In this example, the lesson plan, Measuring the Earth, inherits all the slots from its super-classes. That is, as a resource, it bears properties of the Resource class, and as a lesson plan, it contains a special set of slots pertaining to lesson plans only. The Subject slot is linked to an external knowledge structure because of the complexity and amount of vocabulary involved.

Most of the descriptions involve simple first-order relationships. Each descriptive element is relatively independent, so the individual subclasses have no necessary relationships expressed in the ontology. Most of the relationships are simple statements about a particular descriptive element, such as Language and Pedagogy. The most complicated is Relation Type. In the GEM controlled vocabulary, the following relation types are available:

- hasBibliographicInformation
- isAgencyReview
- isChildOf
- isContentRatingFor
- isCriticalReviewOf
- isDataFor
- isDerivedFrom
- isOverviewOf
- isParentOf
- isPeerReview
- isQualityScore
- isRevisionHistoryFor
- isSiblingOf
- isSiteCriteria
- isSourceOf
- isSponsoredBy
- isStandardsMappingOf

These terms present a mix of first- and second-order relationships. Expressing this requires three elements: the subclass for the resource itself, a slot for the relation type, and a second, optional, slot for the entity constituting the second half of the relationship. The slots must also have their cardinality defined. Cardinality governs the number of possible values a slot can have, and the cardinality is defined as the minimum and maximum number of values. The first slot has a minimum and maximum cardinality of one, i.e., it cannot be empty, and it can only hold one value. The second slot has a minimum cardinality of zero and a maximum cardinality of one. It can be empty in the case of first-order relationships, and can take a single value if necessary to express a second-order relationship. For example:
### Table 3. Examples of first order and second order relationships

<table>
<thead>
<tr>
<th>Example</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developing Geometry Concepts Using Computer Programming Environments</td>
<td>First-order</td>
</tr>
<tr>
<td>Activity Is-an-instance-of/Activity</td>
<td>First-order</td>
</tr>
<tr>
<td>Activity Is-a-subclass-of Resource</td>
<td>Second-order</td>
</tr>
<tr>
<td>Activity Has-RelationType IsSponsoredBy National Council of Teachers of Mathematics</td>
<td>First-order</td>
</tr>
<tr>
<td>Project Has-RelationType IsSiteCriteria</td>
<td>Second-order</td>
</tr>
<tr>
<td>Course Has-RelationType IsDerivedFrom Exemplary and Promising Mathematics Programs</td>
<td>Second-order</td>
</tr>
</tbody>
</table>

This structure accomplishes a number of goals. It leaves the essential flexibility of the GEM vocabulary intact. The resource serves as the parent class, and subclasses can be repeated to enrich a description. Ontolingua allows an ontology to be exported in LISP code, and this adds an important dimension by formalizing relationships in ways that will make the records more easily manipulable by retrieval agents. The relatively simple structure of the ontology reflects the underlying simplicity of the controlled vocabulary.

### Conclusions and Future Research

The theoretical and methodological framework described in this paper serves as the rationale and guideline for converting the GEM controlled vocabulary into an ontology. Compared to the original semantic model of GEM controlled vocabulary, the major difference between the two models lies in the values added through deeper semantics in describing digital objects, both conceptually and relationally. Unlike most of ontological effort that made no reference to controlled vocabularies created by the librarians and information scientists, this study recognized the importance of connecting existing knowledge structures as well as the difficulty in implementing such structures in ontologies. Since the purpose of this ontology is to describe digital objects in the GEM repository, their characteristics become central to the ontology. From an information architecture point of view, the ontology promotes standardization and reusability of information representation through identifying characteristics common to all digital objects for the top class and subsets of characteristics common only to one particular subclass of objects.

As a continuing effort, GEM ontology components will continue to be built under the current framework. The future research will look into problems such as cross-class relationships (which class is to be used as a slot of which class under what condition), develop templates for representing differing digital object classes, and eventually, implement the ontology into the GEM system for the purposes of enriched representation of digital objects and intelligent information retrieval.

### References