Building a Topic Map Repository

Xia Lin

College of Information Science
and Technology
Drexel University
Philadelphia, PA 19104

Jian Qin

School of Information Studies
Syracuse University
Syracuse, NY 13244

Abstract

This paper presents a research prototype of topic maps repository. The main purpose of this repository is to promote knowledge structure sharing and reuse in representing and organizing information resources. The overall architecture of the repository includes a relational database (using MY-SQL), an XML-enabled application server, a customized XML schema for Topic Maps, a set of XML stylesheets for transforming and displaying topic maps, a set of Java servlets and jsp programs to generate XML files dynamically from the database, and a web interface for browsing and authoring topic maps. Through the interface, a user can create topic maps without having to know the syntaxes of topic maps and XML. The user can add or modify topics, occurrences, associations and other related information directly on the web. The user can also upload an XTM topic map file into the repository or export a topic map to an XTM file.

1. Introduction

As the volume and complexity of information on the Web is growing rapidly, users need not only to have effective search systems to help them locate and retrieve information, but more importantly, to have knowledge tools that will help them build structures of their interest for searching and discovering information resources. Users may use traditional subject structures such as thesauri, subject headings, classification schemes, or taxonomies that exist in some information retrieval systems for help. Nevertheless, such knowledge structures are neither widely available for most of Web-based search tasks, nor specifically enough for search problems of individual users. Topic maps introduced recently address both of these problems. Topic maps are designed for the Web and tailored to individual needs.

However, potentials of topic maps would be very limited without effective tools to help users create, manage and share their topic maps. In this paper, we discuss the design and implementation of a topic map repository prototype to help users build knowledge structures and topic maps. The knowledge repository will help to integrate users' topic maps with general knowledge structures in thesauri, classification schemes, ontologies, or some other forms of knowledge representation. The knowledge repository also has a Web interface to allow individual users to create and edit their own topic maps based on different representations captured in the repository. The prototype targets general users — they can use the topic map editor to do what used to be a professional librarian or indexer's job: creating and modifying knowledge structures, searching information with these structures, and sharing them with other users and reuse them in the future. In the following, we will first review concepts of topic maps and identify potential problems that users may face when they create or use topic maps. We will then describe in details our design and implementation of the knowledge repository.

2. Topic Maps

Recently, topic maps as a new type of knowledge structures for content representation and navigation have quickly caught people's attention. A topic map is a collection of topics with appropriate structures. A topic is considered as an aggregate of topic characteristics, including a topic name and its variances, occurrences, and roles played in associations with other topics (ISO/IEC, 2000). Pepper (2000) used an acronym "TAO" to represent the three most important concepts in topic maps: *T* for Topics, *A* for associations, and *O* for occurrences. Using Topic Maps' syntax, topics and associations can be encoded as XML documents and links established with resources (occurrences) that are relevant to these topics. Unlike thesauri where subject descriptors and resources are completely separate entities, topic maps allow tidily coupling of topics and topic resources through the use of occurrences.

Topic maps promise many uses in organizing and navigating information content. Topic maps link topics together in such a way as to enable navigation between them. This capability can be used for virtual document assembly, and for creating thesaurus-like interfaces to corpora or knowledge bases. Topic maps can also filter an information collection to create views adapted to specific users or purposes. For example, filtering by using topic maps can aid in management of access modes depending on security criteria and delivery of partial views depending on user profiles and/or knowledge domains. One of the most important features is that topic maps can structure unstructured information objects, or to facilitate the creation of topic-oriented user interfaces that provide the effect of merging unstructured information bases with structured ones. Unlike metadata practices that create surrogates of information objects such as indexing terms or abstracts attached to documents, topic maps create an external, arbitrary structure for information objects altering its original form. (ISO/IEC, 2000)

While topic maps are promising, creation of topic maps is not a simple task. Traditional knowledge structures, such as subject headings or thesauri, were constructed by domain experts through lengthy and intellectual intensive processes. Linking subject terms to information objects (documents) is also done through a laborious indexing process typically performed by trained professionals. Obviously, this would be a daunting task for most general users. Tools for creating and editing topic maps must be developed in order to make them practical and useful for the general public.

Thus, a primary goal of our work is to design and develop tools to help users create topic maps. We started with the identification of several major problems that users often encounter in creating knowledge structures and then designed solutions around these problems. For example,

Problem 1: Creating knowledge structures is difficult for many users.
Solution 1: Provide examples from standardized thesauri and ontologies. Allow direct drag-and-drop to copy from the thesauri and other topic maps.

Problem 2: Adding occurrences to topics is too labor-intensive.
Solution 2: Allow the occurrences to be both "hard links" and "soft links," which is query-mediated links based on search engines. "Soft links" may be created automatically by default.

Problem 3: Working with XML-coded topic maps is tedious.
Solution 3: Create a visual topic editor that releases users from directly working on the tedious XML-coding.

Problem 4: Structures of topic maps are not easy to display visually.
Solution 4: Enhance the structural hierarchy and create more components to build hierarchies.

These solutions will be further explained in the following sections. The heart of these solutions is a knowledge repository that stores knowledge structures, integrates them with standard protocols, and presents the structures in multiple formats.

3. Knowledge Repository Design

The term “knowledge repository” appears commonly in the literature of knowledge management, especially in association with commercially available knowledge management products. It refers to a system or system architecture that houses and manages a collection of corporate intellectual assets. An idealized example of knowledge repositories is the "Dynamic Knowledge Repositories" proposed by Doug Engelbart and his group (described in Carroll, 2001). Engelbart advocates a bootstrapping process that will “make it easy for everyone to store, organize, access, and analyze the majority of human information online."

Repositories for topic maps will share many ideas with “dynamic knowledge repositories.” Generally, topic map repositories need to maintain syntax and semantics of the topic maps across different types of resources, need to allow easy access to knowledge structures and contents in the repositories, and need to provide users tools for easy editing and updating of the structures and contents. Many of these features can be found in several topic map repositories emerged recently. For example, Mondeca’s topic map repository incorporate existing knowledge structures as its knowledge base for building new topic maps(Delanousse, 2001). Empolois’s k42 knowledge server provides a Web-based topic map editing tool and let users create, retrieve and edit topics and topic associations through the Web browser. An open source project, *Nexist*, is being developed to build topic map interface as a collaborative knowledge repository system (http://sourceforge.net/projects/nexist/).

We design and implement the knowledge repository further along these directions. In particular, we attempt to achieve two objectives: (1) to integrate disparate knowledge structures and facilitate the creation of topic maps; and (2) to promote user interactions with knowledge structures for more effective searching, browsing, and navigation. The objectives guide the design rationales of this knowledge repository:

- The repository should have a scalable and expandable architecture that works well under the current experimental scale as well as for future multiple domain, multiple knowledge structure integration.
- The repository should use XML-based schema language to create a neutral knowledge platform for integrating and displaying multiple knowledge structures. Integrated views should be developed for these knowledge structures.
- The repository should promote interoperability. It should have the capability to import topic maps from other repositories or export topic maps to XTM and other topic maps standards.
- The repository should be user-participatory and user-controllable. It should allow individuals to customize how the interface should look like and how knowledge structures should be modeled. It should provide tools for different groups of users to create and edit knowledge structures, whereas users may be an individual person, a group with a special interest, or an organization with multiple profiles; and
- The system should directly link to the Web or designated Web sites and allow users to search and organize information on the Web through knowledge structures in the repository.

4. Knowledge Repository Implementation

The knowledge repository provides two essential functions, one is to store and integrate various knowledge structures and topic maps, and the other is to accommodate various input/output functions to let users view, edit, and create new topic maps from existing knowledge structures. Figure 1 shows an overview of four major components of the knowledge repository. These components are described below.

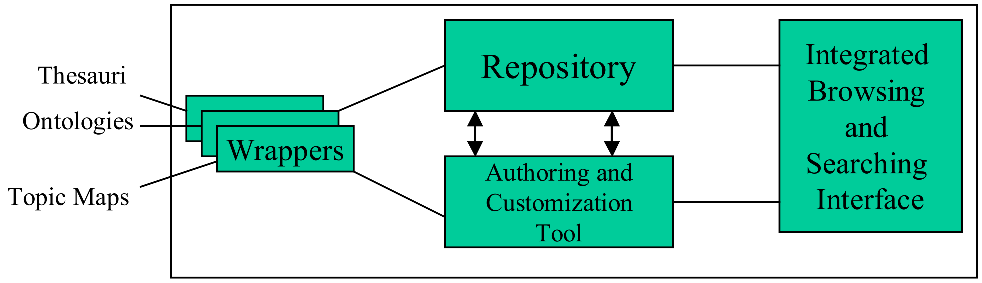

Figure1. Major components of the knowledge repository

4.1 Wrappers

Wrappers are created to convert disperse knowledge structures into an integrated XML schema used in the repository. Initially, we started with XML topic maps (XTM) as the basic XML schema for the repository. As we tried to import several different knowledge structures into the repository, the need to customize XTM schema immediately became clear. For example, traditional subject thesauri emphasize a great deal on hierarchical relationships. ISO 2788 requires to support basic relationships such BT(Broader Term), NT(Narrower Term), TT(Top Term, the term representing the broadest term in the particular thesaurus). This kind of hierarchy can be coded by XTM, but it is often not specific enough. Without additional elements in the schema, it will increase tremendously programming work needed to display or utilize the hierarchy in searching and browsing. Hence, we emphasized more on hierarchies than XTM does. Figure 2 shows a structure of topic maps that have incorporated hierarchical enhancement. The "theme" in Figure 2 defines topical areas, and a new top-level topic, called "branch," represents the broadest concepts within a topic map. The branch in this case functions as equivalent to TT in ISO 2788. Because of its unique functions, "branch" as a topic needs to be treated differently and deserves to be a category on its own. Other elements we added are "order" and "search name". "Order" is needed for changing topic display orders, and "search-name" is defined as a special variance of topic name. This is to recognize the fact that topics may need to be searched by a search name that might be different from topic name or other name variances (more on this feature are discussed in section 4.4).

Using wrappers, we can convert ISO 2788-based thesauri to the repository. We can also convert RDF-based taxonomies, or XML-based ontologies to the repository. For each new standard, only a new wrapper needs to be defined. It needs to emphasize that while we introduce more elements in the internal XML schema, topic maps created through our knowledge repository can be easily export to XTM-based topic maps, and XTM topic maps can also be imported to the repository.

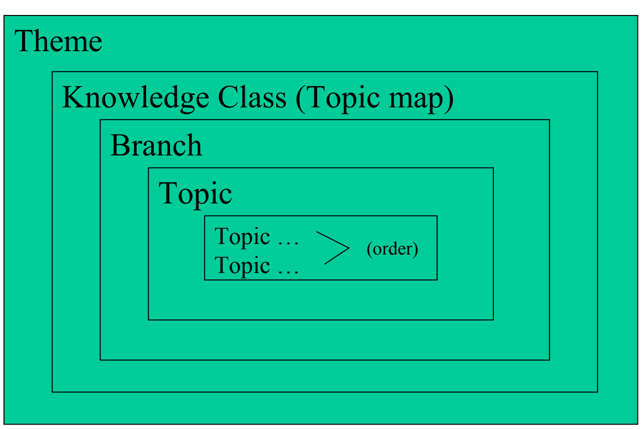

Figure 2
Hierarchically-Enhanced Topic Maps Structure.

4.2 The repository

The first question for an XML-based repository is the choice of databases. One can choose a native XML database, a relational database, or an object-oriented database. Native XML database defines a logical data model for an XML document; it stores and retrieves XML documents directly. While selecting native XML databases seems to be natural and tempting, the option in fact is very limited. Many native XML databases are just on the horizon and have not been tested extensively. Some others are mostly built on top of relational databases with additional functionalities. The most discouraging fact is perhaps the lack of standard XML database API and query languages, although there is work in progress in both the areas (Staken, 2001).

Currently, we implemented our repository on MySQL database. MySQL database is a popular open source database. It is robust and easy to manage; it is easy to connect to any application server via JDBC; and its table architecture is particularly suitable for both document-centered and data-centered approaches for XML data storage (i.e., whether storing XML data as documents or as data in the database). We create a database for each theme, and convert data in the database to XML documents using DOM-based servlets. Within each theme, there are tables for topics, associations and occurrences. There are also tables for topic maps, branches, and facets, etc. These structures can be easily replicated when a new theme is added to the repository.

4.3 The authoring tool

The authoring tool is a web-based interface supported by Java servlets. Through the interface, a registered user can create topic maps without having to know syntaxes of topic maps and XML. The user can add or modify topics, occurrences, associations, and other related information directly on the web. The user can also upload an XTM file for topic maps into the repository.

In addition to general authoring tasks, we particularly address two problems with the authoring tool. The first is to make it easy to create topic map structures. The second is to make it easy to add occurrences and other information resources to topic maps. For each term in the topic map hierarchy, users can edit it, adding synonyms and other related terms to it. Users can also associate the term with specific information resources through a URL (hard links) or through a query-mediated link (soft links). Query-mediate links are links through search engines with pre-defined search queries. Query-mediated links can be as effective as having a search engine for the user to type in specific queries, but they are much easy to use (Golovchinsky, 1997). To make it easier to define search queries, we create several search strategies for users to choose from. Figure 4 show an example term editing screen. With a click on one of the option boxes, the same term can be linked to a search engine in different search strategies such as:

- Search by the term alone
- Search with the term and a scope term
- Search with the term and all its synonyms
- Search with the term and all the terms in its upper hierarchy
- Search with the term and all the terms in its lower hierarchy
- Search with the term and all the sibling terms

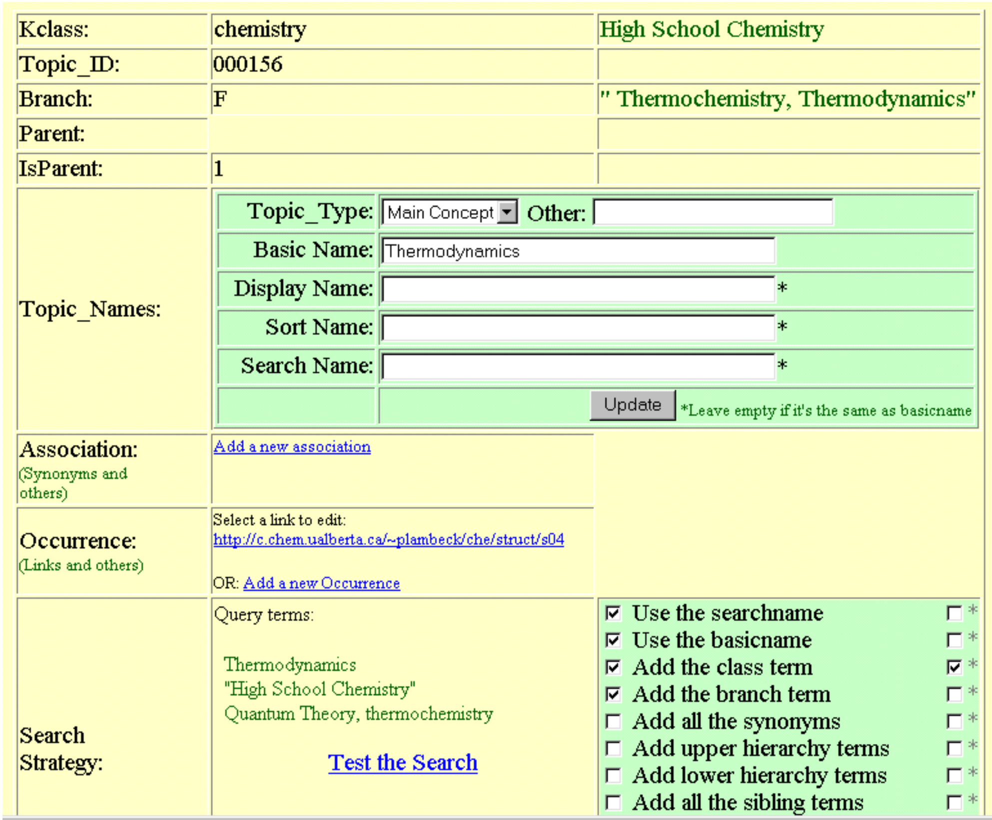

Figure 3. A topic map editor interface

4.4 The Browsing and Searching Interface

The last component of the knowledge repository is the browsing and searching interface. The user can perform a series of tasks with the interface -- browsing through a list of available topic maps, selecting a topic map and displaying it hierarchically or alphabetically (Figure 4). The hierarchical branches are collapsible and expandable. The user can also display just the list of occurrences – a kind of bookmark list. While not every topic has associated occurrences, every topic is associated with a query-mediated link, either by a default search strategy, or by a user specified search strategy as described in the previous subsection. Thus, when the user clicks on any topic in any of the lists, the query-mediated link will be activated, opening up with a new window showing the result of the query-mediated link from a selected search engine. The result can also be changed to other search engine's results with a single click on any of the search engine's names provided (Currently, the search engines listed include AltaVista, Excite, Google, Goto.com, Hot Bot, Northern Light, and Yahoo!). Being able to associate information resources directly with knowledge structures is a main advantage of topic maps. This interface allows the user to navigate knowledge structures and information resources, as well as in-between them.

Fig. 4. The Searching & Browsing Interface

5. Conclusion

Topic maps as a new type of knowledge structures for content representation and navigation have great potentials. To realize the potentials, a comprehensive knowledge repository is indispensable. This paper reports some preliminary work in developing a topic map repository. The repository helps integrate existing knowledge structures and let users create topic maps based on existing knowledge structures. The repository also maintains visual interfaces to let users create, edit, and use topic maps in the Web environment. As the system is being continuously refined, we plan to conduct tests and evaluations of the repository with real users' participation. We expect that the user evaluation will provide valuable input for how this knowledge repository should be improved further and how the repository can be deployed in the real world.

6. References

Carroll, N. (20010. Retrieving information from dynamic knowledge repositories: overview. http://www.hastingsresearch.com/net/02-dkr-ir-intro.shtml

Delanousse, J. (2001). Email communication on the topicmap-comment listserv. http://lists.oasis-open.org/archives/topicmaps-comment/200111/msg00030.html

ISO-2788. International Standard ISO-2788, Documentation -- Guidelines for the development of monolingual thesauri, Second edition -- 1986-11-15.

ISO/IEC 13250:2000 Topic Maps: Information Technology -- Document Description and Markup Languages, Michel Biezunski, Martin Bryan, Steven R. Newcomb, ed., 3 Dec 1999. http://www.y12.doe.gov/sgml/sc34/document/0129.pdf

Lin, X.; & Chan, M. L. (1999). Personalized knowledge organization and access for the web. Library and Information Science Research, 21(2), 153-172.

Ogievetsky, N. (2001). Harvesting Topic Maps with XSLT. In: Proceedings of the Knowledge Technologies Conference 2001, http://www.cogx.com/kt2001/.

Pepper, S. (2000). The TAO of topic maps. http://www.gca.org/papers/xmleurope2000/papers/s11-01.html

Staken, K. (2001). Introduction to Native XML Databases. In: XML.com, http://www.xml.com/pub/a/2001/10/31/nativexmldb.html.

TopicMap.Org. (2001). XML Topic Maps (XTM) 1.0: TopicMaps.Org Specification. http://www.topicmaps.org/xtm/index.html

Trippe, B. (2001). Taxonomies and topic maps: categorization step forward. EContent Magazine, http://www.econtentmag.net/Magazine/Features/trippe8_01c.html