Editor: Peiya Liu
Siemens Corporate Research

XML Topic Maps: Finding Aids for the Web

Michel Biezunski
InfoLoom

Steven R. Newcomb
Coolheads Consulting

Topic maps superimpose an external layer that describes the nature of the knowledge represented in the information resources. There are no limitations on the kinds of information that can be characterized by topic maps. The purpose of the Extensible Markup Language topic maps (XTM) initiative is to apply the topic maps paradigm in the context of the World Wide Web.

Finding information

In a world of infoglut, it's becoming a real challenge to find desired information. Hiding irrelevant information is most effectively and accurately done on the basis of categories, but there's a number of ways to categorize the contents of any corpus, and each system of categorization represents only one particular worldview.

Information users shouldn't be forced to use a single ontology, taxonomy, glossary, namespace, or other implicit worldview. On the Web, we should federate and exploit different worldviews simultaneously, even if those worldviews are cognitively incompatible with each other.

Finding information—metadata that helps information seekers to find other information—is often too valuable to limit its exploitability to a single closed or proprietary environment. Finding information should be application- and vendor-neutral, so that users can freely exploit it in many ways and contexts.

The topic map paradigm provides a solution for interchanging and federating finding information that diverse sources produce and maintain according to different worldviews (see Figure 1).

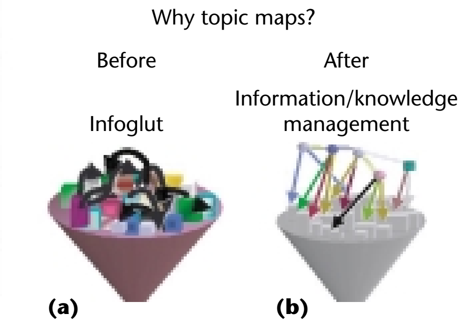

Figure 1. A depiction of a corpus of information (a) without topic map management and (b) with topic map management.

What's a topic map?

A topic map is a representation of information used to describe and navigate information objects. The topic maps paradigm requires topic map authors to think in terms of topics (subjects, topics of conversation, specific notions, ideas, or concepts), and to associate various kinds of information with specific topics. A topic map is an unobtrusive superimposed layer, external to the information objects it makes findable. The findability of a given information object , (that is, the ease with which it can be found) has two aspects:

1. The ease with which a list of information objects that is guaranteed to include the information object can be created by means of some query, and

2. The brevity of that list. The shorter the list, the easier it is for a human being to find the desired information object within the list.

A topic map can act as a kind of glue between disparate information objects, allowing all of the objects relevant to a specific concept to be associated with one another. Topic maps are metadata that need not be inside the information they describe.

Interchangeable versus application-internal topic maps

Topic maps take two forms: interchangeable topic maps that are XML or SGML documents, and directly usable topic map graphs that are the application-internal result of processing interchangeable topic maps. Topic map graphs are abstractly described in terms of nodes and arcs. Topic maps can be formatted as specific kinds of finding aids: indexes, glossaries, thesauri, and so on. We sometimes regard formatted finding aids as a third form of a topic map, but this isn't strictly

1070-986X/01/$10.00 © 2001 IEEE

true, because such finding aids cannot necessarily contain or reflect all of the information present in the topic maps from which they were derived.

Topics

A topic map consists of topics, associations between topics, and very little else. The interchange syntax of topic maps provides an additional element type, called `<mergeMap>`, that refers to another interchangeable topic map with which the containing topic map should be merged at processing time.

A topic is a compound information object—represented and processed by a computer—that serves as a hub to which everything about the subject that the topic represents is connected. In a well-constructed topic map, every topic represents (or has) exactly one subject; a topic can be a computer-processable surrogate for its subject. (The Statue of Liberty can be a subject, but statues can't be processed by computers. A topic whose subject is the Statue of Liberty, however, is processable by a computer. One way to say this is, "A topic reifies a subject.") We chose the word *topic* because it connotes both subject and location (from the Greek *topos*, meaning location).

Topic occurrences

Information objects that are external to a topic map—but which bear some relevance to the subject of some topic—are called *occurrences* of that topic. A topic occurrence is a resource declared to contain some kind of information about the topic under consideration that's useful under certain conditions (that is, within some "scope"). Occurrences can be typed with user-defined semantics; the types, like everything else in topic maps, are represented as topics. The idea of classifying occurrences by type is comparable to a (much less expressive) convention that causes some page numbers in a book index to appear in boldface.

Topic names

A topic may have zero, one, or several names, in any number of languages, for any set of user contexts and scenarios (such as beginner, expert, casual, official, slang, formal, secret, and so on). Each name, or basename, consists of a string of characters. Basenames appear in namespaces, and each namespace is defined by a scope; thus, you can address a topic by its name in a given namespace (scope). Basenames can also have variants—icons, sound recordings, strings to be used as sort keys, and so on. The variants of a given basename are distinguishable by sets of topics (parameters) whose subjects are the processing contexts in which the variants are intended to be used.

Topics without names may seem useless, but they're actually frequently used. A simple Web link expressed as an HTML `<A>` element can be a nameless topic (with a subject that's unspecified, but which is nonetheless quite real) that has two occurrences—one with an occurrence type of origin, the other with the occurrence type of target. The subject can be the reason why the `<A>` link exists, whatever that reason might be (normally some particular notion that's common to all occurrences). There are other similar strategies for upgrading existing Web knowledge assets, because it's usually the case that when there's such a link between two resources, it's because they're somehow relevant to the same subject. As in nearly all kinds of cross-references, the topic and its subject already exist—they only lack a corresponding formal declaration, and they can then be given any number of names.

Unintentional name clashes must be prevented by means of the scoping facilities of topic maps (see the "Scope" section).

Topic associations

Topics can participate in associations with one another. Topic occurrences and associations may be instances of classes of occurrences and associations. The classes—like everything else in topic maps—are also topics. A class topic and an instance topic play their respective roles in a class-instance association. Classes of topics, classes of associations, and the roles played in associations are all user-definable, and they're all topics (for example, they're the subjects of topics that represent them). Associations are composed of members—each topic participating in an association plays a specific member role in the association. Topic maps neither interfere with nor limit the use of existing categorization schemes in knowledge bases, ontologies, taxonomies, vocabularies, indexes, and so on. All are welcome and supportable, and all can be federated to meet the needs of users.

Instances of associations can be subjected to validation via an association templating facility.

Scope

Topic characteristics (their names, occurrences, and participations in associations) are all scoped. Scopes—which are themselves sets of topics—define the extent of validity within which topics have each of their characteristics. Topics them-

Building Standards for Finding Aids

Developers initiated work on topic maps in 1991 when Unix system vendors (and others, including the publisher O'Reilly and Associates) founded the Davenport Group. The vendors were under customer pressure to improve consistency in their printed documentation. Users were concerned about the inconsistent use of terms in the documentation of systems and in published books on the same subjects. Vendors wanted to include independently and seamlessly created documentation under license in their system manuals. One major problem was providing master indexes for independently maintained, constantly changing technical documentation aggregated into system manual sets by the vendors of such systems. The first attempt at a solution to the problem was humorously called SOFABED (Standard Open Formal Architecture for Browsable Electronic Documents).

The problem of providing living master indexes was so fascinating that a new group was created in 1993, called the Conventions for the Application of HyTime (CApH). The group applied the sophisticated hypertext facilities of the ISO 10744 Hypermedia/Time-based Structuring Language (HyTime) standard. HyTime was published in 1992 to provide Standard Generalized Markup Language (SGML, ISO 8879:1986), the standard on which XML is based, with multimedia and hyperlinking features. The Graphic Communications Association Research Institute (GCARI, now called IDEAlliance) hosted the CApH activity. After the CApH group reviewed the possibilities of extended hyperlink navigation, it elaborated the SOFABED model, renaming it *topic maps*. By 1995, the model was mature enough that the ISO/JTC1/SC18/WG8 working group accepted it as a new work item—a basis for a new international standard. The topic maps standard was ultimately published as ISO/IEC 13250:2000 (http://www.y12.doe.gov/sgml/sc34/document/0129.pdf).

During the initial phase, the ISO/IEC 13250 model consisted of two constructs: topics and relationships between topics (later called associations). As the project developed, the need for a supplementary construct that could handle filtering based on domain, language, security, and versioning emerged. A vestigial form of the first filtering mechanism, called *facet*, persists in the ISO standard, but it is not found in the XTM 1.0. Scope-based filtering is far more powerful than facet-based filtering, and since implementations of topic maps must support scooping anyway, the facet facility is redundant.

The ISO 13250 standard was finalized in 1999 and published in January 2000. The syntax of ISO Topic Maps is simultaneously open and constrained, expressed as a set of architectural forms. (Architectural forms are structured element templates; this templating facility is the subject of ISO/IEC 10744:1997 Annex A.3, http://www.ornl.gov/sgml/wg8/document/n1920/html/clause-A.3.html.) Applications of ISO 13250 can freely subclass the element types provided by the element type definitions in the standard syntax, and freely rename the element type names, attribute names, and so on. Thus, ISO 13250 meets the requirements of publishers and other power users for the management of their source codes for finding information assets.

However, the advent of XML—and XML's acceptance as the Web's lingua franca for communication between document and database-driven information systems—created a need for a less flexible, less daunting syntax for Web-centric applications and users. Such a syntax is achievable without losing any of the expressive or federating power that the topic maps paradigm provides to topic map authors and users; the XTM specification provides such a syntax.

The XTM initiative began as soon as the ISO 13250 topic maps standard was published. Working with IDEAlliance and others, the authors founded an independent organization called TopicMaps.Org (http://www.topicmaps.org) for the purpose of creating and publishing an XTM 1.0 specification as quickly as possible. In less than one year, TopicMaps.Org was chartered and it delivered the core of the XTM 1.0 specification at the XML 2000 conference in Washington, D.C. on 4 Dec. 2000.

The authors were the founding coeditors of the core deliverables portion of the XTM specification, as well as of the remaining portions of the authoring group review version of the specification. In January 2001, Graham Moore (Empolis) and Steve Pepper (Ontopia) became the new coeditors, and Eric Freese (DataChannel) was appointed the chair of TopicMaps.Org.

Another version of the XTM 1.0 specification was released on 17 Feb. 2001. It contains a corrected version of the XTM 1.0 document type definition, and its Annex F proposes certain constraints on the processing of XTM topic maps. Meanwhile, the authors are independently publishing their ongoing work relevant to topic maps, RDF, the Semantic Web, and so on at http://www.topicmaps.net.

selves don't have scope—only their characteristics have scope, and each characteristic has its own scope. The set of topics that defines a given topic characteristic's scope is entirely determined by the author of the topic map, who also determines the topics in the topic map. Any topic can be a member of the set of topics that defines the scope of any topic characteristic, but there's no requirement that any topic must participate in the definition of any scope.

Subject-based topic merging

It's possible for two or more `<topic>` elements to have the same subject. In such a case, applications merge them into a single abstract topic node that exhibits the union of their characteristics (the subject-based merging rule). A similar rule—the name-based merging rule—applies when two `<topic>` elements have the same basename string within the same scope.

For purposes of subject-based merging, we assume that two topics have the same subject if they specify that the same resource serves as their subject indicator. A subject indicator is a resource that the topic map author believes will precisely indicate to users the subject of the topic. Similarly, two topics have the same subject if they specify that one and the same resource actually *is,* rather than *indicates*, the subject.

Theoretically and ideally, after completing all topic map processing, every topic has exactly one subject, and no subject is represented by more than one topic—there's a one-to-one correspondence between topics and subjects. Merging multiple independently produced topic maps can pose serious challenges. Topic map authors can greatly facilitate the federation of their work with the work of others by taking the trouble to refer to widely used published subject indicators. All topics that have a subject indicator in common readily merge into a single topic. The published subject indicator notion is expected to spawn businesses around lists, indexes, ontologies, vocabularies, and so forth.

Name-based topic merging

Several topics can have the same name (for example, they can have a homonymous one), and yet not be intended to represent the same subject. For example, New York (the state) isn't the same subject as New York (the city). To make the two topics addressable by their names, it's necessary that the two topics have these names within different namespaces. (The term namespace is used here in the generic sense. A namespace is an abstract place where, when one utters a name, one either gets the one and only thing [in that place—that namespace] that has the name one uttered, or an error message saying, "There's nothing here that has the name you uttered.") In topic maps, the namespace within which a topic has a name is defined by the scope (that is, the set of topics that defines the scope) within which it has that name. In the New York example, the topic about the subject of New York City might have the basename New York within a scope that includes a topic whose subject is the notion of "cityness," while the scope within which the topic whose subject is New York State has the name New York wouldn't include the "cityness" topic.

The name-based merging rule requires that whenever two or more topics have the same name in the same scope, they're assumed to have the same subject, and they're automatically merged into a single topic node. This rule can be exploited to facilitate topic map maintenance and the merging of independently maintained topic maps. It's up to the author or maintainer of a topic map to decide whether particular topics should be merged, and name-based merging is just one of the knowledge-federation aspects of the paradigm. When topics with different subjects are merged, the result is always incorrect, and such incorrect merging has undesirable impacts on the usefulness of the resulting merged topic map. To protect against all unintended name-based topic merging when whole topic maps are merged—as well as for other reasons—the `<mergeMap>` element can add topics to all the scopes of all the topic characteristics declared by the `<topicMap>` being merged.

Formalisms

The XTM 1.0 specification expresses the high-level concepts of the topic maps paradigm, and the relationships between the concepts, by means of Unified Modeling Language (UML) diagrams.

The XTM interchange syntax, like the ISO 13250 interchange syntax, is expressed as a document type definition (DTD). Although the XML interchange syntax is currently expressed as a DTD, it could also be expressed as an XML schema or in other ways, such as Relax, Schematron, TREX, and so forth. The DTD formalism was chosen because it's by far the most widely supported and most mature formalism for XML syntaxes, and it's completely adequate for the purpose.

Work on rigorous processing models for topic maps is in progress. A rigorous illustration of the authors' deepest understanding of the meaning of topic map information is being maintained by them at http://www.topicmaps.net/pmtm4.htm. The graph representation used to describe at least one kind of processing model for topic maps has recently led to the creation of a visual vocabulary for diagramming topic maps, their various constructs, and the decomposition of those constructs into their underlying concepts. An alternative mathematical expression language for this visual vocabulary is also under development. Because the graph-based language enables the representation of the properties of topic map constructs at their most elementary levels, it's expected to be helpful in showing how the Topic Maps paradigm intersects with (and, ultimately, can interoperate with) other metadata representation and interchange standards.

Related Resources

Articles

We recommend the following articles as further resources.

L. Alschuler, "Going to Extremes," 13 Sept. 2000, http://www.xml.com/pub/a/2000/09/13/extremes.html

L. Alschuler, "Topic Maps," 16 June 2000, http://www.xml.com/pub/a/2000/06/xmleurope/maps.html

T. Berners-Lee, "Semantic Web," presentation at Extensible Markup Language 2000 Conf. (XML 2000), 6 Dec. 2000, http://www.w3.org/2000/Talks/1206-xml2k-tbl.

N. Ogievetsky (Cogitech), "XSLT Transformations between ISO Topic Maps and XTM," http://www.cogx.com.

S. Saint-Laurent, "Getting Topical," *xml.com*, 20 Dec. 2000, http://www.xml.com/pub/a/2000/12/20/topicmaps.html.

Web Sites

The following Web sites deal with topic maps or expand upon related topics discussed in this article.

M. Biezunski and D. Kennedy, http://www.infoloom.com

Empolis (Bertelsmann), http://www.topicmaps.com

Enabling technology for XML and SGML architectural forms is freely available at http://www.hytime.org/SPt

Robin Cover's extensive bibliography on topic maps and current related events is available from OASIS at http://www.oasis-open.org/cover/topicMaps.html

Information about work in progress on the graphic representation of processing models, including the one being elaborated for XTM and ISO 13250. Page maintained by Sam Hunting, http://www.etopicality.com

Mondeca, topic map products, http://www.mondeca.com

Ontopia, products, http://www.ontopia.net

TM4J, open source software (Kal Ahmend), http://www.techquila.com

Xml.com Resource Guide, http://www.xml.com/pub/rg/Topic_Maps

XTM official Web site, http://www.topicmaps.org

Topic maps and the semantic Web

There's some overlap between topic maps and the Resource Description Framework (RDF, http://www.w3.org/rdf) specification. Both standards aim to represent connections between information objects and can encode metadata, among other things. Since the Extreme Markup Conference in August 2000 (http://www.extrememarkup.org), where a memorable boxing match between RDF and Topic Maps was held, discussions have been ongoing between those responsible for the two standards. Later, at the Graphic Communcations Association (GCA) XML 2000 Conference where the publication of the XTM 1.0 Core Deliverables was announced, Tim Berners-Lee, the Director of the World Wide Web Consortium (W3C) proposed in his presentation on the semantic Web (http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide15-1.html) that there should be a convergence between RDF and topic maps. The implications of such a convergence will include benefits for DARPA's Agent Markup Language (DAML) initiative, as well as with the ontology interface layer (OIL).

The central notions of topic maps will play increasingly significant roles in future generations of Web technology, because the severity of the infoglut problem is only going to increase. The Topic Maps paradigm is designed not only to accommodate diversity; it preserves, cherishes, and leverages diversity in the conquest of infoglut. Whenever a new vocabulary, ontology, and so on appears, it need not be regarded as evidence suggesting that the dream of global knowledge interchange can't be realized. On the contrary, it's cause for hope, because of the knowledge-federating, diversity-leveraging power of topic maps. No great difficulty is posed by the need to welcome yet another community of interest into the global community of communities of interest. Communities of interest are defined by their worldviews, and whenever a community of interest rigorously exposes its worldview in a fashion that permits its knowledge to be federated with the worldviews and knowledge of other communities, the whole human family is enriched. **MM**

Michel Biezunski and Steven R. Newcomb have been working together since 1992. They cofounded the CapH Group (Conventions for the Applications of HyTime), hosted by the Graphic Communications Association Research Institute (now called IDEAlliance). With Martin Bryan, they coedited the ISO/IEC 13250:2000 "Topic Maps" standard. They cofounded TopicMaps.Org, and they were the editors of the first release of XTM 1.0, an interchange syntax for topic maps in XML.

*Michel Biezunski is a consultant at InfoLoom (*www.infoloom.com*). Among other things, he designs and implements topic maps for commercial and government clients. Readers may email Biezunski at mb@infoloom.com*

Steven R. Newcomb is an independent consultant in information management, with clients in industry and government. He is a coeditor of the HyTime international standard (ISO/IEC 10744:1997, Hypermedia/Time-based Structuring Language). Readers may email Newcomb at srn@coolheads.com.

Contact Standards editor Peiya Liu, Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540, email pliu@scr.siemens.com.