

# Topic Maps go XML

**Steven R. Newcomb**

TechnoTeacher, Inc., Allen, USA

srn@techno.com

<http://www.techno.com>

**Michel Biezunski**

Infoloom, Paris, France

mb@infoloom.com

<http://www.infoloom.com>

***Abstract:***

*No abstract was provided for this paper.*

## Introduction

A topic map is a document conforming to a model used to improve information retrieval and navigation using topics as hubs in an information network. Topic maps are created and used to help people find the information they need quickly and easily. Topic maps can be formatted as a wide variety of finding aids, including printed indexes, glossaries, and many kinds of high-performance online finding aids.

Public topics, shared by potentially large communities of users, are expected to dramatically increase performance and reliability of search engines on the Web and elsewhere.

The Topic Maps model has given rise to an international standard (ISO/IEC 13250:2000), which was developed as an application of HyTime. "HyTime" (ISO/IEC 10744:1997) comprises a powerful set of standard hypermedia extensions to the SGML Standard Generalized Markup Language ("SGML", ISO 8879:1986), of which XML (the eXtensible Markup Language) is a W3C-recommended profile. This paper outlines the current state of the expressibility of Topic Maps in XML, the growing usefulness of Topic Maps in Web-oriented contexts, and the ongoing XTM Specification effort of TopicMaps.Org to increase the exploitability of the Topic Maps paradigm in mass markets.

## Topic Maps: global positioning systems for topic spaces (e.g., the Web)

Topic Maps are arousing much interest in the Internet community, for reasons that include the following:

### Extended linking and infoglut

In the context of the coming "extended linking" paradigm, in which the anchors of links can be elsewhere than the links themselves, a single web-based information object may become a traversal initiation anchor for millions of links. This can cause a classic infoglut ("needle in a haystack") situation, in which so much information is accessible that the desired information cannot be found. In the "extending linking" paradigm, the glut problem is somewhat reduced by its provision of a way to distinguish between types of relationships. However, when many things bear the same type of relationship to an anchor, there can still be an infoglut problem. In the "extended linking" paradigm (as it is described in both the HyTime standard and in the XLink draft W3C Recommendation), there is little standard provision for distinguishing between potential traversal targets that all bear the same relationship to the traversal initiation anchor. The Topic Maps paradigm provides a powerful way to make it possible for people to use computers to help them sort through otherwise unmanageably large sets of possible traversals by allowing search criteria to be focused

using a sophisticated array of topic-combinatorial principles. Together, these principles can provide, in effect, a "GPS for the Web", as Charles Goldfarb has put it.

### **Topics are better finding aids than strings**

Topics are described in the Topic Maps paradigm as first-class objects which contain a set of names and pointers to an unlimited number of occurrences. Furthermore, topics can be related to others by means of "topic associations".

### **Topic maps will promote a new generation of search engines**

Even without extended linking, existing Web search engines provide suboptimal access to the contents of the Web. Many important resources are missed because search engines are not topic-oriented. Instead, search engines typically leverage string-matching and variants of string- and phrase-matching algorithms, they observe the search behaviors of human beings and make inferences therefrom, etc. But people are not interested solely in hits on particular terms used to identify topics, nor solely in statistics about the past behaviors of people who appeared (to a computer) to be interested in similar topics, nor solely even in subsets of worldwide information that have already been indexed by particular information vendors or portal operators. The Topic Maps paradigm provides an astonishingly powerful way for web users to gain the benefit of combining the finding and focusing power of many indexing strategies, search engines, and organizations. The Topic Maps paradigm permits the results of independent indexing efforts to be ongoingly merged by unrelated individuals and organizations, for the benefit of all concerned – and especially for the benefit of people who are looking for information.

### **How are Topic Maps relevant to XML?**

#### **XML is making semantic markup the lingua franca of the Web**

HTML has been the first choice language for the Web as far as presentation and one-way hypertext links are concerned, but XML is now meeting requirements that HTML could never meet. Unlike HTML, XML provides a way to use markup to associate arbitrary semantics with arbitrary chunks of information. By using appropriate algorithms and a modicum of good sense, creators of topic maps can significantly improve their productivity by leveraging such semantic markup. They can more readily incorporate additional resources into the base of resources for which a topic map provides the basis of a set of finding aids, and they can more easily identify topics, topic types, topic relationships, topic relationship types, occurrences, occurrence types, names and scopes that should be added to their topic maps.

#### **Topic Maps extend the power of XML to make information self-describing**

Even in the coming XML-dominated Web, topic maps are important because persons other than the authors or owners of arbitrary chunks of XML information often have need to associate their own arbitrary semantics with such chunks, even though they do not have the authority to write on them or to contribute to their inherent semantic markup. For a variety of technical, economic and political reasons, standard topic map documents are an ideal way to allow anyone to contribute what, in effect, amounts to added or alternative semantic markup to any component(s) of any set of information resources, without having to change those resources in any way. With pure XML, we can have only a resource author's views (and only the author's purely hierarchical views) as to the semantics of each information component. With additional topic maps, we can take advantage of different perspectives on the same information. The topic maps paradigm dramatizes the fact that the distinction between data and metadata (data about the data) is purely a matter of

perspective. Topic map documents reflect the perspectives of their authors on all the information components they address; they define semantically customized views.

### **Topic maps are XML documents**

Topic map documents can already be expressed as XML documents, in complete conformance with the W3C XML 1.0 Recommendation and with the syntax specified by the ISO Topic Maps standard. To express a topic map document in XML, one need only respect the syntactic constraints imposed by XML, in addition to those imposed by SGML. (For example, unlike XML, SGML permits the omission of redundant markup, such as implicitly unnecessary end-tags for elements.) None of the additional syntactic constraints imposed by the syntax of XML interferes with the expressibility of topic maps; it is straightforward to express topic maps in XML.

### **How can Topic Maps be exploited in mass markets?**

The ISO Topic Maps standard, like the other members of the ISO SGML family of standards, is designed for extreme generality and flexibility, and for comprehensive utility in all systems contexts. This level of generality is highly desirable in an international standard for system-independent, vendor-independent, and application-independent information interchange, but it introduces complexities that may be irrelevant to the majority of applications.

In recent years, considerable marketing effort has been directed at making SGML exploitable in mass markets. These marketing efforts have given SGML a new name ("XML"), a simplified syntax, and an expectation that XML software and systems will cost much less (while being almost as powerful) than comparable SGML software and systems. Similarly, the draft W3C XLink Recommendation embodies a subset of the concepts and syntax of HyTime's linking and addressing facilities. XML and XLink can afford to be much more limited in their scope and flexibility partly because they have been designed for exactly one information delivery system: the World Wide Web. It seems reasonable to assume that, if XML and XLink are viable mass-market standards, then some similar syntactic and/or functional subset of the Topic Maps standard should also be a viable mass-market standard. Only the worldwide information management industry can determine exactly how to do that, by means of some consensus-seeking process.

Such a consensus-seeking process has been inaugurated by IDEAlliance (formerly known as the GCA Research Institute, the body that sponsored the development of the Topic Maps paradigm beginning in 1993) under the moniker "TopicMaps.Org". The process is intended to culminate in the publication of a standard called the "XTM Specification." The general mission of TopicMaps.Org is to "engage in technical, educational and marketing activities to facilitate the use of topic maps based on XML, including but not limited to application on the Web." The XTM Specification will be "an application of the ISO/IEC 13250:2000 Topic Maps standard that trades some of the generality of the international standard in exchange for a simpler, more predefined and more immediate way to exploit topic maps in mass market applications. One aspect of [ ] task is to develop industry-wide consensus about how topic maps should be expressed in XML, and how the Topic Maps paradigm can be most readily exploited in Web-based applications. [ ] It is not a goal to preserve the full generality and applicability of the ISO Topic Maps standard. ISO/IEC 13250:2000 already does that; it already works equally well with both XML and SGML, so there is no good reason to make an XML version of pure 13250. [ ] It is a goal to preserve the ability of applications that understand ISO-conforming topic maps in their full generality, to also understand XTM-conforming topic maps, in accordance with the relevant international standards, as just one conventional way of expressing ISO-conforming topic maps. In other words, XTM topic maps should retain the ability to be merged with

other kinds of topic maps, in accordance with ISO/IEC 13250:2000. [ ] develop a set of use cases for the particular communities we seek to serve."

### **The XTM specification: some known technical issues**

Assuming that the XTM Specification is a subset of the syntax prescribed in ISO 13250, the issues to be resolved in the XTM Specification include, but are not limited to, the following:

#### **Should the linking syntax be restricted to the W3C "extended" XLink syntax?**

ISO 13250's linking syntax is presently the ISO/IEC 10744:1997 "HyTime varlink" (variable link) syntax. HyTime varlink syntax is not significantly different from that of "extended" XLink, so this restriction poses no great challenge or difficulty. It seems advisable to conform to the "extended" XLink architectural form described in the corresponding draft W3C Recommendation.

The Topic Maps paradigm is not supportable using "Simple" XLink.

#### **How should addressing be accomplished?**

Should the notations used to address both topic occurrences and topics themselves be restricted? For example, the only supported addressing notation could be the W3C XPath notation. This would be adequate for addressing whole resources expressed in any notation, and components of resources expressed in XML, including the topic links used to express topics.

XPath expressions cannot be used to address multiple targets. If XPath is chosen as the only addressing notation, this limitation may necessitate very specific constraints on the way in which, for example, topic links are expressed as XLinks.

#### **Should a set of predefined topics, such as topics needed for topic map housekeeping, be provided and required in all XTM-conforming topic maps?**

If so, such topics can become public topics that are defined as a work product of TopicMaps.Org.

#### **Should Topic Map Templates, an implicit feature of ISO 13250 topic maps, be explained explicitly in the XTM Specification?**

Topic Map Templates are a convention for applying the Topic Maps paradigm in a way that is expected to be essential in providing "easy on-ramps" and other benefits to topic maps creators and users.

#### **Should the names of attributes, etc. defined in ISO 13250 be changed in the XTM Specification?**

Daniel Rivers-Moore urges that some attributes (especially "types" and "type") be renamed, because their present names are overloaded and confusing.

The architectural forms paradigm allows attributes and element types to be renamed without sacrificing conformance to ISO/IEC 13250:2000. The real issue here is what names would offer advantages over the ISO standard names.

#### **What about the transitivity of relationships and other inference rules?**

Steve Pepper, Holger Rath and others have urged that some associations, including class-subclass associations, should be specially recognized for their "transitivity". (For example, if A is a subclass of B, and B is a subclass of C, then it can be inferred that A is a subclass of C.) Transitivity inferences are a subset of all of the kinds of inferences that can be made. (An example of another kind of inference is that if A is a daughter of B, and C is a sister of B, then C and A enjoy an aunt/niece relationship with each other.) The whole question of inference rules is not addressed in ISO 13250; inferencing is regarded as a property of systems and not of documents, and therefore inference rules are regarded as outside the scope of the standard. Of course, ISO 13250 topic map documents can contain the results of inferencing. For example, a topic map can contain an explicit aunt/niece relationship. Indeed, topic maps can contain the results of any other kinds of processing and methodologies. Also, ISO 13250 standardizes neither specific association types nor the specific properties of certain kinds of association types. TopicMaps.Org will consider whether to provide additional syntax for inference rules, thus extending the scope of the XTM Specification beyond the scope of ISO 13250.

### **How should public topics be addressed?**

Should the XTM Specification require that public topics be addressable using XPath? If so, the overwhelming bulk of the world's predefined public topics (including, for example, the Subject Headers of the Library of Congress) will not be usable by XTM-conforming applications (unless and until they are maintained by their owners as publicly available XML documents). If not, the whole question of the addressing capabilities of XTM applications is not going to be resolved by limiting all addressing to XPath.

Etc.

## Authors

### Steven R. Newcomb

TechnoTeacher, Inc.  
Postal Address:  
405 Flagler Court  
75013 Allen  
Texas  
USA  
Telephone: +1 972 517 7954  
Fax: +1 972 517 4571  
E-mail: [srn@techno.com](mailto:srn@techno.com)  
Web: [www.techno.com](http://www.techno.com)

**Steven R. Newcomb** - Principal in TechnoTeacher Inc., a software developer and consultancy that is the source of the GroveMinder technology, with licensees in telecommunications, computers, defense, education, energy, publishing, government, and aerospace. Co-editor of ISO/IEC 10743:1996 Standard Music Description Language (SMDL). Co-editor of ISO/IEC 10744:1992 (and 10744:1997) Hypermedia Time-based Structuring Language (HyTime). Founding Conference Chair, International HyTime Conference, 1994-1997. Conference Co-chair (with Carla Corkern) of the successor Metastructures conference, 1998-1999. Founding co-chair, Extreme Markup Languages Conference, 2000. Founding Chairman, Conventions for the Application of HyTime (CApH) activity of the Graphic Communications Association Research Institute (now IDEAlliance), the original developer of the Topic Map paradigm, and co-editor of ISO/IEC 13250:2000, the Topic Maps information architecture. Co-chair, TopicMaps.Org. Consultant to ISOGEN International Corp. (now a DataChannel company), 1997-present.

### Michel Biezunski

Infoloom  
Postal Address:  
1 boulevard du Temple  
75003 Paris  
France  
Telephone: +33 1 44 59 84 29  
Fax: +33 1 47 50 90 33  
E-mail: [mb@infoloom.com](mailto:mb@infoloom.com)  
Web: [www.infoloom.com](http://www.infoloom.com)

**Michel Biezunski** - Michel Biezunski, Ph.D., is working as an independent consultant. He is actively involved in the creation of an information industry based on XML-related technologies. He is co-editor of the ISO/IEC 13250 Topic Maps standard and is now working on topic map implementations, including software tools to create and maintain topic maps. Michel is co-chair of the TopicMaps.Org XML Topic Maps activity.