ISO/IEC JTC 1/SC34 N0429

ISO/IEC JTC 1/SC34

Information Technology --

Document Description and Processing Languages

TITLE:	Topic Maps Reference Model: Requirements
SOURCE:	Patrick Durusau
PROJECT:	Topic Maps
PROJECT EDITORS:	Michel Biezunski, Martin Bryan, Steven R. Newcomb
STATUS:	Editor's Draft, Revision 1.8
ACTION:	For review and comment
DATE:	3 July 2003
SUMMARY:	
DISTRIBUTION:	SC34 and Liaisons
REFER TO:	
SUPERCEDES:	
REPLY TO:	Dr. James David Mason (ISO/IEC JTC1/SC34 Chairman) Y-12 National Security Complex Information Technology Services Bldg. 9113 M.S. 8208 Oak Ridge, TN 37831-8208 U.S.A. Telephone: +1 865 574-6973 Facsimile: +1 865 574-1896 E-mail: mailto:mxm@y12.doe.gov http://www.y12.doe.gov/sgml/sc34/sc34oldhome.htm
	Ms. Sara Hafele Desautels, ISO/IEC JTC 1/SC 34 Secretariat

American National Standards Institute
25 West 43rd Street
New York, NY 10036
Tel: +1 212 642-4937
Fax: +1 212 840-2298
E-mail: sdesaute@ansi.org

3 July 2003

JTC 1/SC 34 N 0429

Topic Maps Reference Model: Requirements

Version 1.8, 2003/07/03
Go to http://www.isotopicmaps.org/TMMM/RMREQ-latest.html
to view or contribute to the current editors' working revision.

Table of Contents

0   Introduction

In ISO 13250:2002, there is much evidence to suggest that there is an underlying abstraction that is not stated explicitly. Indeed, those who drafted the standard have always insisted that their work was guided by such an abstraction, and they have frequently and openly regretted that, on account of resource constraints and time pressure,

no such abstraction was codified explicitly in the standard. The Topic Maps Reference Model (TMRM) makes the underlying abstraction of ISO 13250:2002 explicit and does not extend or limit ISO 13250:2002.

The TMRM provides a basis for evaluating syntaxes and data models for Topic Maps, including but not limited to those specified by ISO 13250:2002, in terms of their ability to arrive at one proxy for each unique subject. The state of having one proxy for each unique subject, also known as the "Subject Location Uniqueness Objective," is an objective for only some subjects in any given data model or syntax.

The TMRM does not constrain the designs of syntaxes or data models for topic maps. However, *it does provide disclosure mechanisms for such syntaxes and data models*. These disclosure mechanisms are applicable regardless of whether such definitions are formal (i.e., machine processable) or informal (expressed in natural language), or in some mixture of the two. When a syntax or data model is defined it can be objectively evaluated as to its ability to facilitate the achievement of the Subject Location Uniqueness Objective in the topic maps that it governs. The TMRM's disclosure mechanisms provide implementers with the means to assure topic map authors and users that topic maps will be reliably interpreted as their authors intended.

While it is certainly desirable to achieve the Subject Location Uniqueness Objective for all subjects, it is unfortunately impossible for any single data model to accomplish this aim. It is therefore inevitable that multiple data models and syntaxes will be used for topic maps (and for systems that process topic maps) that serve different kinds of purposes. It is vital that each data model's limitations be knowable by anyone who might select it for some particular purpose. Authors cannot assume that all data models are designed to achieve subject location uniqueness for the subjects that are important to them and to the users of their topic maps. Authors must base their choices of data models on reliable information about those data models.

The TMRM provides the means to fully disclose the strategies that will be used to achieve the Subject Location Uniqueness Objective, and the kinds of subjects to which each strategy will be applied. In the case of syntaxes, the disclosure mechanisms make explicit all the provisions for subject addressing.

The disclosure mechanisms of the TMRM actually simplify the task of defining data models and syntaxes for topic maps. The TMRM provides an underlying abstraction in terms of which all the key aspects of data models and syntaxes for topic maps can be disclosed. For example, the HyTime syntax of ISO 13250:2002 requires each `<topic>` element to have a unique ID attribute (`id`), and it also provides an optional subject identity attribute (`identity`) (see section 5.2.1 of ISO 13250:2002). Both of these syntactic attributes -- `id` and `identity` -- are designed to facilitate the addressing of the unique subject of each `<topic>` element. However, the differences between these two attributes are critically important, and, in the absence of a model of the structure of topic maps that is more abstract than the structure of the syntax, the semantics of these two attributes are difficult to explain clearly. The TMRM provides a basis for making all such explanations more clearly and consistently than would otherwise be possible.

Every instance of an interchangeable syntax for topic maps must specify, implicitly or explicitly, a single data model that is intended to govern its interpretation. (Otherwise, the meaning of the instance would be indeterminate.) However, multiple different interchange syntaxes can be intended to be governed by a single data model. The design of each different interchange syntax is necessarily driven by assumptions about the usage scenarios in which the syntax is expected to be used, and it is not possible for any single interchange syntax to be *optimal* for *all* usage scenarios. The overwhelming weight of experience in the SGML/XML arena teaches that:

1. in order to be useful, the scope of any syntax (as defined by means of a DTD or using any other formalism) used for information interchange must be carefully and explicitly limited, and

2. syntaxes generally need to evolve in response to changing conditions.

Syntaxes for interchanging topic maps are not exempt from these considerations. The Topic Map standard would defeat its own purpose if, in some future version, it forbade the use of any syntax for topic map interchange other than the ones it already specifies. The TMRM, when adopted, will allow the Topic Map standard to embrace the necessity for users to define their own syntaxes for topic map interchange without sacrificing either the integrity of the paradigm, or the possibility of merging topic maps expressed in

different syntaxes.

The TMRM recognizes two classes of things -- data models and syntaxes -- of which ISO 13250:2002 already contains instances. In order to describe itself without creating undue confusion for those already familiar with the terminology of ISO 13250:2002, the TMRM introduces two new terms, **information model** and **assertion**. An **information model** is a class of things of which the TMRM is itself an instance: a set of notions about abstract information object classes, including abstract information object classes whose instances are relationships between instances of abstract information object classes. Such a set of notions is an idealized model that imposes no predefined data structures on designers of data models. **Data models** are quite different; they represent design choices for implementers. For example, any data model for topic maps would necessarily define an object class for topics (i.e., for proxies for subjects), but it could conceivably define multiple object classes for that purpose. The definition of such a data model could use the disclosure mechanisms of the TMRM. Use of the TMRM **information model** allows designers of conforming **data models** for topic maps to be clear and precise about which kinds of proxies are subject to which kinds of strategies (if any) for the achievement of the Subject Location Uniqueness Objective.

Reflecting its origins in hypertext, ISO 13250:2002 uses the term **association** to mean *an expression of a relationship between two or more subjects*. However, in order to make its syntaxes more intuitive, ISO 13250:2002 uses different terms for a few special kinds of relationships. For example, it uses the term **occurrence** for relationships in which one of the role players is a piece of information relevant to the other role player. Another example is its use of the term **scope** for relationships in which one of the role players is a relationship, and the other role player is a set of subjects that is somehow helpful in understanding the applicability of the relationship (the "scope" of the relationship).

The **information model** of the TMRM, however, regards all relationships as instances of a single uniform structure, the "**assertion**" structure. In order to help those already familiar with the terminology of Topic Maps to understand the TMRM, the TMRM introduces the term **assertion**, meaning *an expression of a relationship between two or more subjects, without exception, regardless of their semantics, and regardless of any syntactic*

conventions that may be used to represent them for interchange. The terminological distinction between ISO 13250:2002's **association** and the TMRM's **assertion** is an essential tool for accomplishing one of the primary goals of the TMRM: distinguishing the existing syntaxes and data models of topic maps from the essential information model of topic maps -- of distinguishing the instances from the class.

The disengagement of the information model of topic maps from any particular syntax or data model is more than an academic exercise: it is critically important to the usefulness of the ISO 13250:2002 standard, and to its breadth of adoption. For example, the proposed data model (N0396 Topic Maps -- Data Model) explicitly states that the merger of the subject proxies that it calls "locator items" -- the proxies for subjects that are pieces of addressable information -- is not required. While there may be a significant number of usage scenarios in which it is not necessary to achieve the Subject Location Uniqueness Objective for subjects that are addressable pieces of information, it is certainly true that at least some usage scenarios, such as the creation of reverse indexes, absolutely require it. (Indeed, important existing Topic Map applications, including the one used by the U.S. Internal Revenue Service, have this requirement.)

N0396 also specifies that any merging of proxies can be done at any time by anybody for any reason. While this would *allow* the merging of "locator items", it also has the side-effect of leaving the interpretation of any ISO 13250:2002-conforming topic map document entirely in the hands of system implementers, each of which is free to merge, or leave unmerged, the proxies (the topics) of any subjects of any kinds in any topic maps. While the flexibility of N0396 will no doubt be useful to implementers, there is no mechanism provided for implementers to disclose the choices they have made for other such mergers. The disclosure mechanisms of the TMRM provide the ability for implementers to make that disclosure and to communicate it to topic map authors without regard to the data model or syntax in use in a particular application.

The TMRM does not demand that the "locator items" -- or any other objects defined by N0396 or by any other data model -- be merged. Nor does it demand that they be left unmerged. However, in the interests of reliable information interchange, the TMRM does provide the mechanisms that enable, whatever the decision as to the merger or non-merger of a given kind of subject, in any

given data model, disclosure of the design decision to merge or not merge in a data model-neutral and syntax-neutral way.

General requirements for the TMRM are set forth below. After enumeration of those requirements, they are discussed in terms of the advantages that an information model, separate from any syntax or data model, brings to the topic map standard and community. Again: the TMRM is not and should not be construed as a syntax or data model for topic maps. It is an explication of the information model that was obscured by the syntaxes used in the original efforts of to the topic map community to formulate a standard for reliable interchange of topic map information.

1　Requirements

1.1　Provide a definitional framework

The TMRM must provide a syntax and data model independent framework for disclosure of the information objects represented by the syntactic constructs defined in ISO 13250:2002 (or any other definition of a topic map data model or syntax) and the merging rules that govern them. That framework must be be sufficiently and unambiguously defined such that such disclosures can be compared with each other, and matched with user requirements.

This requirement includes the following sub-requirements:

1. Show how data models are independent of syntaxes, and how syntaxes are dependent on data models.

2. Show how data models and syntaxes can be disclosed using the TMRM.

3. Define the uniform structure of relationships.

4. Define the uniform process of merging.

5. Show how relationships can govern merging, and how merging can occur in the absence of relationships.

1.2　Illustrate the disclosure mechanisms by applying them to ISO 13250:2002

The TMRM must provide a definition of the information objects and relationships (explicit or implicit) found in ISO 13250:2002, and the applicable merging rules.

This requirement includes the following sub-requirements:

1. For each syntactic construct, define how it must be interpreted as information objects and relationships between them.

2. Comprehensively define the properties of topics that are implicit in ISO 13250:2002.

3. Comprehensively define the relationship types implicit in ISO 13250:2002.

4. Comprehensively define the merging rules of ISO 13250:2002 in terms of relationship types.

5. Show how users can define their own relationship types, as well as merging rules that depend on those relationship types (both of the syntaxes specified by ISO 13250:2002 allow users to instantiate relationship types that are not specified by 13250:2002).

2 Explanation of Requirements

2.1 Definitional Framework

The information objects implicit in ISO 13250:2002, and the relationships between them, are far from clear, because the syntactic constructs actually obscure them. For example, a `<topname>` element must contain one or more `<basename>` elements (see 5.2.2, Topic Name Architectural Form). What is not made explicit is that each `<basename>` corresponds to an assertion whose significance is the fact that a specific subject (the subject of the `<topic>` that contains the `<basename>`) has a specific name (the content of the `<basename>`). The fact that neither the assertion nor the basename itself is marked up as a `<topic>`, while in fact both are legitimate subjects, is an example of how the structure of the ISO 13250:2002 syntax obscures the structure of the information that, for example, `<topname>` and `<basename>` elements are designed to interchange. (The structure of interchanged information is often

apparently different from the structure of the information that is intended to be interchanged by that structure. This should not be surprising, since interchanged information is always necessarily hierarchical and acyclic, while many kinds of interchanged information, including topic maps, are non-hierarchical and may be cyclic. When they look at a topic map represented in an ISO 13250:2002 syntax, different people apparently intuit different things about how it should be interpreted. Intuition is an insufficient basis for reliable information interchange.)

2.2   Define the uniform structure of relationships.

ISO 13250:2002 does not define models for any of the information objects or their relationships. For example, it is commonly recognized in the topic maps community that occurrences and scope are actually forms of what is referred to as **associations** in topic map discussions. In part, that late realization of the common underlying structure was due to the lack of an explicit models for **associations** (as traditionally understood) and occurrences and scope. Had models for these relationships been available, describing the relationships between the various information objects in these relationships, the commonality of those models would have been immediately obvious.

Beyond simply demonstrating the underlying structure of relationships, the TMRM will provide models for the information required to support the merging of topics. No syntax or data model is compelled to follow these models, but their existence will enable the evaluation of such syntaxes or data models for their ability to follow the model of merging set forth in the TMRM. The reliable merging of topics, based upon their subject identity, is the characteristic that distinguishes the topic maps paradigm from other information technologies. The result of merger, in an idealized model, results in all information about a particular topic being discoverable from that topic.

2.3   Define the uniform process of merging.

As already noted, the principal goal of topic maps is to facilitate the achievement of a state in which the proxies of at least some subjects are unique to their subjects. This requires the merger of proxies whenever they are proxies for the same subject. Merger (or

non-merger) is controlled by the data model; the TMRM takes no position on what mergers are or are not proper for a particular topic map instance, except to say that, whatever merging its governing model demands should be done, and whatever merging its governing model does not demand should not be done. Such deterministic merging is essential to the interchange of topic maps.

Without a generalized model for merging, it is not possible to meaningfully describe or discuss the merging rules of any data model or topic map application. A generalized model of merging allows the description and disclosure of the merging rules that the author of a topic map instance intended to be applied to it.

For purposes of illustration, the TMRM will provide a model composed of information objects that are treated as topic information objects for the purposes of merger. Since topic information objects are the only objects within the model that are subject to merger, this will allow users of the model to choose less-complete models of merger for both syntaxes and data models, with full knowledge of the impact that such choices have on the resulting topic map instances.

2.4 Illustration of Disclosure

To illustrate the application of the TMRM and the utility of disclosure for implementers and topic map authors, the disclosure mechanisms of the TMRM will be applied to ISO 13250:2002. The results of applying the TMRM to ISO 13250:2002 will be produced as a non-normative appendix to serve as a guide to use of the TMRM.

3 Conclusion

The specific requirements of the TMRM can be summarized as outlining how to achieve two fundamental objectives:

1. to completely describe topic maps and their components, and

2. to provide a means of disclosing the rules for topic maps and their components in any particular instance.

The first objective is a necessary step forward to allow meaningful discussion of topic maps, their data models and applications. Without models of the various components of topic maps and their relationships, varying interpretations of syntax, data models and applications will continue to be the rule of discussions, rather than the exceptions. This problem will only be aggravated as topic maps move into the mainstream of information technology and developers or information architects outside the present topic map community begin to develop topic maps. Such developers or information architects will not share the common understandings or avoidance of dead ends that are common knowledge among the present topic maps community.

Disclosure, the second goal of the TMRM, is at least as important as the goal of describing topic maps and their components. Disclosure is the means by which topic maps and their data models or applications can be judged against particular user requirements for achievement of the merger of topics. The goal of most users (and, one suspects, topic map authors as well) is the achievement of the **subject location uniqueness objective** for topics in which they are interested and not necessarily for others. Disclosure of the rules followed by particular topic maps allow the preservation of those choices as well as the making of new choices, where topic map authors desire to merge topic maps or other information resources that have followed different choices for the merging of topics.

The combination of description and disclosure contemplated by the TMRM will support the development and selection of syntaxes, data models and applications for topic maps based upon meaningful choices by users and topic map authors. Those choices may not always be the same, but the same description and disclosure will support additional choices by other users and authors to extend those made by others.