

Semantic Resource Exploitation with Topic Maps*

H. Holger Rath

Abstract

Topic Maps are a new ISO standard proving a paradigm for semantic information networks over various kinds of resources. The article gives a brief introduction to the Topic Map paradigm and explains why searching and navigation in Topic Maps has great advantages over other techniques and what the differences to Semantic Networks and RDF are. Knowledge representation is an often cited application domain for Topic Maps. The article presents the essential semantics not defined by the standard: Topic Map Templates, class hierarchies, association properties, inference rules, and consistency constraints. Finally, a first outlook on the Topic Maps Query Language is provided listing the major requirements and design goals of this very new ISO standardization project.

1.1. Introduction

Topic Maps were developed by the same ISO committee which developed SGML, DSSSL, and HyTime—the committee’s name is ISO JTC1 SC34. ISO/IEC13250 (2000) is the official ISO standard published in the beginning of 2000. Its roots go back to 1991 when the Davenport group (initiator of DocBook DTD) wanted to merge the back-of-the-book indices of two UNIX manuals. Over the years Topic Maps became an ISO project and evolved to a powerful but implementable paradigm.

The ISO standard defines the concepts, the data model, and the exchange syntax. Latter is based on SGML and HyTime. XML Topic Maps (XTM) are already under development using XML and XLink as exchange syntax. The XTM 1.0 specification is available at <http://www.topicmaps.org>. The XTM working group *TopicMaps.Org* consists of Topic Map vendors and users and most of the ISO committee members are involved in XTM ensuring the compatibility between the SGML and XML variants. All discussions and results of the group are open to the public and can be followed at the XTM mailing list (<http://groups.yahoo.com/group/xtm-wg>).

The article gives a brief introduction to the Topic Map paradigm explained by a couple of prose and graphical examples but no SGML/XML syntax examples. Syntax examples could be found in various other papers and in the XTM 1.0 specification. This article focuses on the concepts.

* Published in: *Proceedings of the GLDV-Spring Meeting 2001*, Henning Lobin (ed.), Giessen University, March 28th–30th, 2001, pp. 3–15. <http://www.uni-giessen.de/fb09/ascl/gldv2001/>

The representation of knowledge as an application of Topic Maps requires additional semantics which is not defined in the standard. The missing semantics are motivated, introduced, and technical solutions are sketched.

The Topic Map Query Language (TMQL) is the youngest part of the Topic Maps technology. Its ISO standardization project started December 2000 and is in the requirements analysis phase right now (Q1 2001). The major requirements are listed and the implications on Topic Map software are mentioned.

1.2. A Brief Introduction into Topic Maps

A Topic Map defines a meta layer “above” the information resources. The meta layer models all the topics—persons, objects, concepts, thoughts, etc.—which are described “in” the resources and the relations between the topics (see fig. 1.1). The topics and the resources are connected by hyperlinks using HyTime or XLink syntax. A Topic Map could provide different views on the same set of resources (e. g., beginner view and expert view on a technical manual) and the map has a value on its own, even without connected resources.

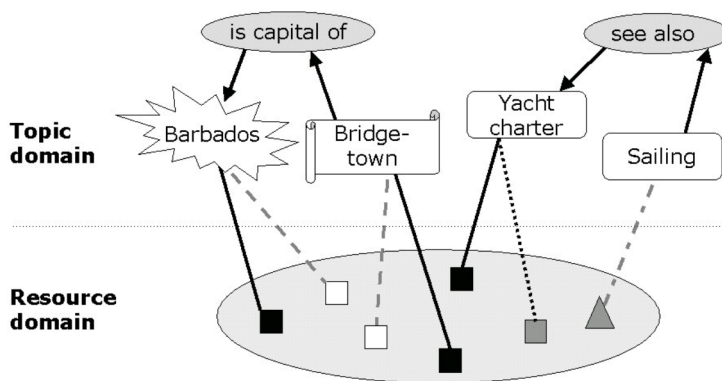


Figure 1.1.: Resource domain and topic domain

1.2.1. Example “back-of-the-book index”

Topic Maps can be easily explained using a back-of-the-book index as an application example. Think about a Caribbean travel guide; its index could look like tab. 1.1.

The index items list the “topics” which can be found in the book. The page numbers point to the “occurrences” of the topics in the book showing where the reader finds the information resources. Different formatting of the topics and the page numbers (occurrences) signals that they are of different *type* (e. g., topic in roman font: island, capital city, site of interest; topic in italic: water sport; occurrence in roman font: description of the topic, occurrence in bold-italic: city map). The “see also” defines a relationship (association) between two topics.

This example contains already the fundamental concepts of Topic Maps : topics, topic types, occurrences, occurrence types, associations, and association types.

Barbados	42
Bridgetown	45, 102
<i>Horseback riding</i>	17
Jamaica	55
Kingston	57, 103
Petroglyphs	35
Reggae	48
<i>Sailing</i>	13
— <i>see also</i> Yacht charter	
San Juan	78, 107
Sugar plantations	67
<i>Yacht charter</i>	14

Table 1.1.: Back-of-the-book index as an application example

1.2.2. Topic

A topic, in its most generic sense, represents any “thing” whatsoever—a person, an entity, a concept, really anything—regardless of whether it exists or has any other specific characteristics, about which anything whatsoever may be asserted by any means whatsoever. The topics represent the things—the subjects (e. g., the island Barbados)—which are in the application domain and make them machine processable.

Every topic has an identifier, one or more types, and several characteristics. The mandatory identifier is the unique address of the topic. The types describe of “which kind” the topic is—in other words a “class-instance” relationship (e. g., the topic “Barbados” is an instance of the class “island”). The referenced type is also a topic which allows self-documenting topic maps and the construction of ontologies. A characteristic is the topic name. A topic might have more than one name, e. g., to cover all synonyms of a subject. See fig. 1.2 for examples of topics and topic types.

1.2.3. Occurrence

The second characteristic of a topic is the occurrence. An occurrence is the link to an information resource that is somehow relevant to the topic—it connects the topic domain with the resource domain. Every occurrence plays a role which is expressed by the occurrence role type (e. g., “description”, “city map”)—which is again a topic. A topic can have as many occurrences as necessary. The link addressing is done with either HyTime or HyTXLink/XPointer and therefore as powerful as these standards are. See fig. 1.3 for examples of occurrences and occurrence role types.

1.2.4. Association

The third characteristic of a topic is the association. An association describes a relationship between two or more topics. The kind of relationship is expressed by the association type (e. g., “capital of”, “see also”).

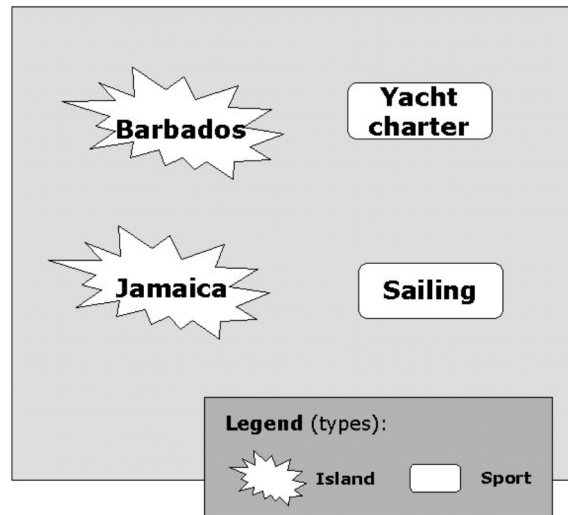


Figure 1.2.: Topics and topic types

Every topic which participates in an association plays a certain role in the relationship. The kinds of roles are expressed by association role types (e. g., “country” and “capital” in the association “capital of”, “term” and “related term” in the association “see also”). Association types and association role types are again topics. See fig. 1.4 for examples of associations, association types, and association role types.

1.2.5. Scope

Any characteristic of a topic (name, occurrence, association) is considered to be valid within a certain context, which may or may not be specified explicitly. The valid context is called “scope” and allows the definition of certain views on the same Topic Map , e. g., “pleasure” and “culture” views on the travel guide or “English” and “German” views on a multilingual Topic Map .

A scope is a set of “themes” which are also topics. The “unconstraint” scope will be assigned implicitly by the Topic Map software when no scope is assigned explicitly. It is up to the application how the different views on the Topic Map are defined (e. g., logical “and” of themes or logical “or” or more complex Boolean operations).

1.2.6. Identity

The difficulty of automatic merging different Topic Maps into one map can be reduced to the question if two topics are about the same subject or not. This sometimes philosophical issue is solved very pragmatically by the Topic Maps standard.

Two topics are the same either if the two topics have the same name in the same scope or if they refer to the same subject indicator. The topics and all their characteristics could be merged if this condition holds. Merging topics means that only one merged topic is the result of the merge process. Merging names and occurrences means that the merged topic has the union of names as well as occurrences of the single topics as characteristics. Merging of associations means that the merged topic plays now the roles in all associations the single topics played the roles before.

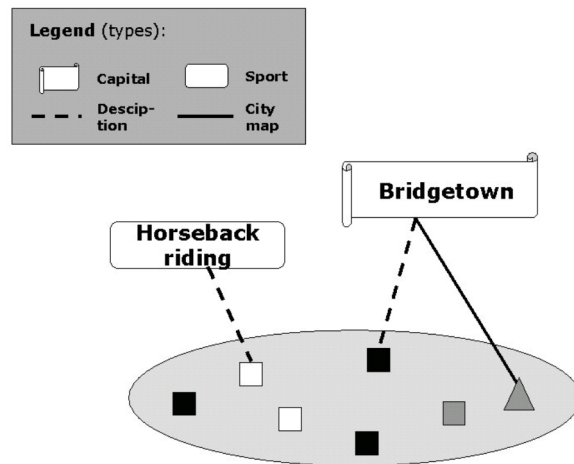


Figure 1.3.: Topics and their occurrences with different roles

It is the goal of the standards committees (ISO and TopicMaps.Org) to provide lists of “Published Subject Indicators” (PSIs) which could be used in various Topic Maps. The PSIs ensure two issues: (i) merging of Topic Maps using them becomes easy and straight forward and (ii) Topic Maps software could support the associated semantics. Therefore, sets of PSIs could be seen as application profiles.

1.2.7. Searching and Navigation in Topic Maps

The description of the Topic Map paradigm showed that resource domain and topic domain are two distinct layers. The resource domain contains all information resources—documents, graphics, images, database records, audio/video clips, etc.—and the topic domain contains all topics, their characteristics, and type information. Topics represent the subjects which are expressed by the resources. Associations model the implicitly existing knowledge—which is in the resources—by bringing the topics in meaningful relationships. The two domains (layers) are connected by occurrence links pointing from topics to or into information resources.

The user can use both layers for searching and navigation. Full text search or XML-based structure search are known search functions which can be applied to the resource domain. Their results are not very precise because result sets might consist of a large number of resources and the requested information has to be found in each resource by reading through it. Searching in the subject level—the topic domain—leads to precise results. The result is a list of topics. The size of the list can be easily decreased if the query was under-specified or can be increased if the query was over-specified. Getting from the topics to the needed corresponding resources or resource fragments is only one click away—by following an occurrence link with clearly defined semantic of the resource for this topic—done by the occurrence role type.

Searching and navigation in the topic domain are very similar issues. Navigation is done by following explicitly existing links—e. g., getting from the typing topic “sport” to the topic “Sailing” and following the association of type “see also” to get to the topic “Yacht charter” and its description (= occurrence) on “page 14”. Thus Topic Map navigation can be seen as a path through the Topic Map consisting of concrete values. Searching could be modeled as a navigation

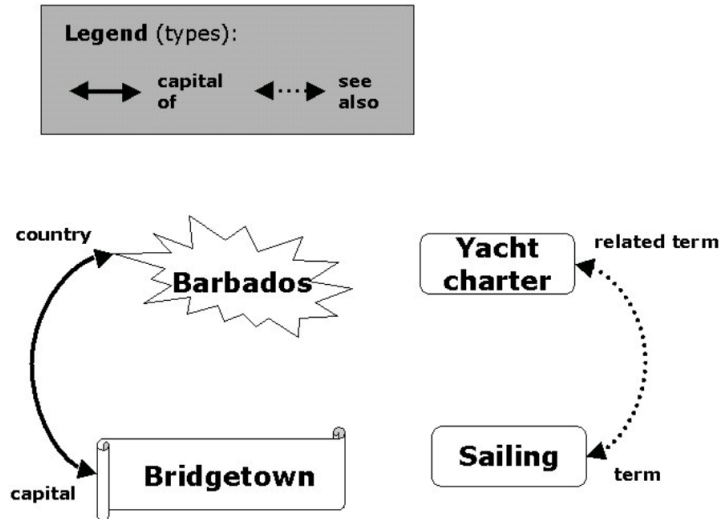


Figure 1.4.: Topic associations of different types

path with concrete values and un-instantiated variables. The goal of searching is to find possible instantiations of the variables which result in a meaningful navigation path—e. g., find all topics X which are associated with topic “Barbados” and have occurrences of role “city map”.

Navigation in the resources is mainly about following the hyperlinks which connect the different resources—the well-known link maze problem arises and it does not matter for the user if the links are stored inline in the documents or if they are stored separately in a link base. But if the resources are linked as occurrences to topics then there is an “up-link” to the much richer conceptual level. And if the server software (e. g., Web server) delivering the resources to the client software (e. g., Web browser) is aware of all the up-links then the related topics or other parts of the Topic Map can be displayed together with the resources. This means that the user always sees the subjects in which the current resource is embedded and that s/he can always switch back and forth between resource domain and topic domain.

The topics which provide the user with the hint where s/he is when looking at a resource behave like satellites of the GPS (Global Positioning System). The satellites send a signal down to earth that is used by GPS receivers which informs you where you are on the globe and help you not to get lost. The topics “send” the occurrence “signal” down to the resource domain that is used by Topic Maps software which informs you where you are in the information universe and helps you not to get lost in hyperspace. In other words: “Topic Maps are the GPS of the information universe”.

1.2.8. Topic Maps , Semantic Networks, and RDF

The Topic Map paradigm is very similar to the Semantic Network paradigm which is an AI (artificial intelligence) concept developed for knowledge representation. Topic Maps and Semantic Networks have a lot in common, but there are some minor differences which make Topic Maps a little bit more powerful. The differences are:

- The predefined scope concept supporting multiple views on the same Topic Map .
- The occurrence link connecting topic domain and resource domain and thus “bridging information management and knowledge management”.
- The precise definition how to merge different Topic Maps .

The concrete relation between the W3C recommendation Resource Description Framework (RDF, <http://www.w3.org/RDF/>) and Topic Maps is under investigation. But it looks like that RDF is more general than Topic Maps are. Which implies on the one hand that Topic Maps could be expressed by the means of RDF (e. g., as an RDF schema) and on the other hand that Topic Maps software provides more functionality out-of-the-box because of the predefined Topic Map semantics. Both the RDF and Topic Map committees are working together to develop a harmonized solution which could be one of the base layers for the Semantic Web.

1.3. Knowledge Representation with Topic Maps

One obvious application domain of Topic Maps is knowledge representation. The similarity to Semantic Networks implied this already. But the Topic Map standard only defines the basic techniques but no special semantics required for efficient knowledge representation. The missing predefined semantics are:

- Distinction of the ontology/taxonomy part from the remaining “regular” Topic Map .
- Class hierarchies to build rich ontologies/taxonomies.
- Association properties and inference rules to deduce implicit knowledge.
- Consistency constraints for validation purposes.

The sum of all listed semantics results in a “Topic Map Template” (see below). The template itself can be expressed as a Topic Map . Thus, the “template” map controls the “real” map and defines the necessary semantic needed by Topic Map tools.

1.3.1. Topic Map Templates

Most of the “objects” declaring a Topic Map ontology are topics; namely themes and types of topics, occurrence roles, associations, and association roles. But the standard does not provide a name or definition for the list of ontology “objects” of a map and this can lead to some confusion: Users often mix up “ontology” topics and “regular” topics during discussions. In addition to that, the different tasks of topic map design, creation, and maintenance are hard to distinguish and to separate.

The same is true for the control of user access rights: As long there is no distinction, different rights cannot be assigned to the different parts of the map. A separate ontology part could also be used for defining categories of Topic Maps that share a common set of classes with predefined semantics.

The ISO working group has already responded to the need to be able to identify the ontology part of a Topic Map . It coined the term “Topic Map Template” for all ontology topics of a map.

Topic types	Island, Sport, Capital
Occurrence role types	Description, City map
Association types	capital of, see also
Association role types	capital/country, term/related term

Table 1.2.: Examples of a Topic Map Template (see also fig. 1.5)

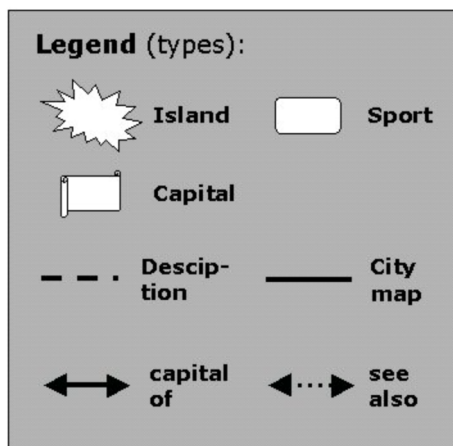


Figure 1.5.: Essential parts of a Topic Map Template

At the present time this term is only “semi-official”, since the concept has not yet been refined and added to the standard.

A Topic Map template consists of all constructs which have a declarative meaning for the map (see fig. 1.5). These are all the topics used as themes and as types for other “regular” topics, occurrence roles, associations, and association roles. They can be identified as template topics by assigning predefined Public Subject Indicators (PSIs, cf. tab. 1.2).

1.3.2. Class Hierarchies

All topics, occurrences, and associations are instances of classes (types). The classes themselves are expressed as topics. This class-instance relationship is in fact merely a syntactically privileged association type defined in the standard. If we are looking at the class-instance relationship from an ontology/taxonomy view, then there is a justifiable demand for a superclass-subclass relationship as well. Superclass-subclass relationships are essential to build taxonomies and more complex ontologies.

Topic types	Person, Artist, Painter, Composer
Superclass-subclass association	superclass: Person subclass: Artist
Superclass-subclass association	superclass: Artist subclass: Painter, Composer

Table 1.3.: Example of a superclass-subclass association

Transitive association type	geographical containment (“is in”)
Association role types	Container, Containee
Association instances	“Giessen” is-in “Hessen” “Hessen” is-in “Germany”
Inferred knowledge	“Giessen” is-in “Germany”

Table 1.4.: Example of a transitive association

The superclass-subclass relationship is realized by an association with predefined association type “superclass-subclass” and role types “superclass” and “subclass”. All three types are again identified by PSIs.

The examples in tab. 1.3 express that every Composer is also an Artist and a Person. This information could be used when searching in the map for topics of type Person: The Topic Map software could automatically return also all topics typed as Artists, Painters, and Composers grouped by the different types. If the users searches for Painters with specific properties and no Painter with these properties is in the map, the software could automatically return Artists or subclasses of Artists which have the properties.

1.3.3. Association Properties

Mathematics define the properties reflexive, symmetric, transitive, anti-reflexive, and anti-symmetric for binary relationships. Because associations can be seen as relationships the properties could be applied to associations which connect two topics. Taking a closer look on the benefits of each property shows that only transitivity is of real value (cf. tab. 1.4).

The transitivity property can be assigned to the association type. Having this information about the association a Topic Map software is able to derive implicit—means not explicitly coded—knowledge from the map.

Topics (and their types)	“Jimi Hendrix” is-instance-of-type “Rock star” “Guitar” is-instance-of-type “Instrument”
Association (and its type and role types)	“Jimi Hendrix” (“Musician”) played “Guitar” (“Instrument”)
Inference rule	If TOPIC1 (“Musician”) played “Guitar” (“Instrument”) and TOPIC1 is-instance-of-type “Rock star” then TOPIC1 is-instance-of-type “Famous rock guitarist”
Inferred knowledge	“Jimi Hendrix” is-instance-of-type “Famous rock guitarist”

Table 1.5.: Example of an inference rule

1.3.4. Inference Rules

The definition of superclass-subclass associations between classes and of transitivity properties for associations already allow powerful inferencing of knowledge. But a Topic Map may contain further knowledge which could be inferred if inference rules are specified.

The inference rules consist of conditions with un-instantiated variables and statements expressing the inferred knowledge. The Topic Map software tries to match the conditions by instantiating the variables and creates the statements with the instantiated variables as new part of the Topic Map. If the new part becomes a permanent part of the map or is just a temporary one—e. g., to answer a query—depends on the storage strategy of the software (see tab. 1.5).

1.3.5. Consistency Constraints

All the previously introduced concepts extend Topic Maps in ways that increase their expressive power and ease creation and maintenance efforts. Both the designer and the editor of Topic Maps expect system support when designing and creating a map which will consist of millions of topics and associations. The question of the consistency of the map becomes a key issue, because it is nearly impossible to check a map of that size manually. For that reason we need concepts to declare consistency constraints and to validate that those constraints have been obeyed.

Consequently a separate schema is needed which contains all the information necessary for the validation process. We call this construct “Consistency Constraints” or just “Constraints”. The validation is the task of the Topic Map development environment (e. g., a editor or an editorial system). It should be performed permanently or on demand—like structure validation in an SGML/XML editor or by an SGML/XML stand-alone parser.

Constraints may be assigned to three potential layers: (i) Topic Map modeling, (ii) user interface for Topic Maps, and (iii) operations on the map. Here, we focus on the Topic Map modeling layer. Constraints are assigned to the types of topics and associations controlling how the instances of these types are allowed to “look like”.

Topic type	Person
Name	Exactly one name without scope; maximum one name with scope “Nickname”
Occurrence	Exactly one of role type “Portrait”; unrestricted number of role type “mention”
Enforced association and association role to play	“born in” playing role “Person” (constraint enforces that “every person is born somewhere”)

Table 1.6.: Example of topic type constraint

Association type	Geographical containment
Scope	“Geography”
Roles and their topic types	Exactly one “Container” (“Country”) Minimum one “Containee” (“State”, “City”)
Roles and their topic types	Exactly one “Container” (“State”) Minimum one “Containee” (“City”)

Table 1.7.: Example of association type constraint

The topic type constraint controls the names (their minimum and maximum number, possible and necessary scopes) and the occurrences (their role type, minimum and maximum number, possible and necessary scopes).

The association type constraint controls the scope (possible and necessary themes), association roles (their type, minimum and maximum number), and the topic types which are allowed to play the roles. Furthermore, it makes sense to ensure that every topic of a certain type plays a certain role in a certain association.

1.4. Topic Map Query Language

The concepts (data model and processing model) and exchange syntax of Topic Maps are well-defined by the ISO and XTM standards. The missing piece is a standardized way to query and to manipulate a Topic Map .

December 2000, the ISO Topic Maps committee proposed a new work item called “Topic Map Query Language” (TMQL). TMQL will define search and manipulation methods.¹

¹ <http://groups.yahoo.com/group/tmql-wg>

The major requirements for TMQL are:

- Information retrieval from one or more persistent topic maps
- Adding and removing information within a persistent topic map
- Support for ontology access
- Support for inference rules
- Support for constraints
- Support for different access points (e. g., command line, built-in into programming language, API, messages within appropriate protocol)
- SQL-like functions (e. g., ordered-by, dealing with large result sets, transaction protocol, views, query optimization, report generation)
- Distinction between high-level operations (result is always a consistent map) and low-level operations (result may be an inconsistent map)
- Reference abstract data model, containing all and only the information required to ensure interoperation between conforming applications
- Establishment of a persistent topic map
- Integrity constraints on a persistent topic map
- Support for modern distributed system architectures
- Generic platform issues, e. g., support for making use of very large topic maps from a very small workstation

The list gives a first impression what status TMQL will have for Topic Map applications—the status will be similar to the one SQL has for relational database systems. TMQL will be *the* interface to Topic Maps .

Some early implementations of the retrieval part TMQL show already the power of the language. Having the retrieval part simplifies the realization of general features such as inferencing, profiling, and constraints—all are somehow based on retrieval techniques.

1.5. Conclusions

The article provided an introduction to the new standard Topic Maps . With its clearly defined concepts the Topic Maps paradigm provides a sophisticated meta model which is appropriate for a wide range of applications. The general idea is the creation of an easy-to-explore semantic summary of the content—the subjects—of the underlying resources. Topic Maps are a kind of super index.

The explicit distinction between resource domain and topic domain as well as the powerful resource addressing inherited from HyTime or XLink imply some major advantages: Topic

Maps can be defined over resources of any format, the same resources could be assigned to different maps, and a Topic Map could be created and exchanged without the occurrence links and connected to the resources later. Topic Maps can grow together with the continuously growing amounts of resources.

The concept of scope offers a solid way to define different views on information networks or knowledge representations. Associations bring the various topics in meaningful contexts implementing Vannevar Bush's "As We may Think".

Knowledge representation is an often cited application domain for Topic Maps . But essential semantics are not defined by the standard. This article presented them: Topic Map Templates allow the distinct creation of ontologies, class hierarchies support the definition of taxonomies and complex ontologies, association properties as well as inference rules provide us with the ability to derive not explicitly coded knowledge from the map, and constraints ensure the consistency of Topic Maps . Furthermore, class hierarchies, association properties, and inference rules minimize the coding effort resulting in compact maps with reduced coding errors.

Finally, the section about TMQL—the Topic Map Query Language—gave an outlook into the future. TMQL will become a further international standard defining the query and manipulation interface to Topic Maps . The presented requirements list showed that TMQL will cover all necessary features for a modern and future-save interface language.²

Bibliography

ISO/IEC13250 (2000): "Information Technology – Document Description and Processing Languages – Topic Maps". International Standard, International Organization for Standardization, Geneva. Available online at <http://www.y12.doe.gov/sgml/sc34/document/129.pdf>.

² Further interesting articles about Topic Map technology can be found at <http://www.topicmaps.com>, <http://www.gca.org/papers/xmlleurope2000/re1/sess10.html>, and <http://www.gca.org/papers/xmlleurope2000/re1/sess25.html>.