Article

Topic maps: templates, topology, and type hierarchies

Hans Holger Rath
STEP Electronic Publishing Solutions GmbH
Technologiepark Würzburg-Rimpar
Pavillon 7
D-97222
Rimpar
Germany
TEL +49.9365.8062.0
FAX +49.9365.8062.66
EMAIL consulting@step.de
WEB http://www.topicmaps.com/

The new ISO standard ISO/IEC 13250 Topic Maps defines a model and architecture for the semantic structuring of link networks. Dubbed the 'GPS of the information universe', topic maps will become the solution for organizing and navigating large and continuously growing information pools, and provide a 'bridge' between the domains of knowledge representation and information management. This paper presents several technical issues of which are of great interest when applying topic maps to real world applications. The main focus of the paper is the introduction of 'topic map templates' — a semi-official term coined by the standards' committee for a concept that the author argues is a necessary but as yet unstandardized addition to the basic model. Furthermore association taxonomies, class hierarchies, and consistency constraints of topic maps are presented and discussed.

Introduction

The ISO (International Organization for Standardization) committee JTC 1/SC 34/WG 3 *Information Technology — Document Description and Processing Languages — Information Association* standardized ISO/IEC 13250 Topic Maps [ISO, 13250:2000] in the autumn of 1999. Formally speaking, the ISO standard defines a model and interchange syntax for Topic Maps. The initial ideas — which date back to the early 1990's — related to the desire to model intelligent electronic indexes in order to be able to merge them automatically. But during several years of gestation, the topic map model has developed into something much more powerful that is no longer restricted to simply modelling indexes.

Markup Languages: Theory & Practice 2.1 (2000): 45–64

A topic map annotates and provides organizing principles for large sets of information resources. It builds a structured semantic link network above those resources. The network allows easy and selective navigation to the requested information. Topic maps are the ' GPS (Global Positioning System) of the information universe'. Searching in a topic map can be compared to searching in knowledge structures. In fact, topic maps are a base technology for knowledge representation and knowledge management.

The basic concepts of the standard are topics, occurrences of topics, and relationships (*associations*) between topics. The section "Topic maps in a nutshell" gives a short overview.

The editors of the standard, together with the other members of ISO JTC1/SC34/WG3 (the author is among those "other members"), defined a well-considered and implementable set of concepts. But first prototypes of practical applications show that there are a number of issues that are not covered by the standard. This was only to be expected since the working group considered it more important to publish a base standard immediately than to delay publication in order to add further refinements. The section "The missing pieces: An overview" discusses some of the concepts that the standard does not cover explicitly and explains why they are important for practical applications.

SGML and XML have DTDs defining classes of instances, but topic maps as currently specified do not have an equivalent construct. The standards working group has recognized this need and coined the term *topic map template* for the 'declarative part' of a map. The section titled "Topic map templates" explains what makes up a template.

Three other additional concepts are also discussed:

- a taxonomy of the basic properties of topic associations ("Association taxonomy"),
- class (or type) hierarchies and how they can be exploited in topic map software ("Class hierarchies"), and
- consistency checking and validity constraints for topic maps ("Validation of consistency").

The final section ("Conclusions") summarizes the paper and gives an outlook on further topic map developments.

Topic maps in a nutshell

The standard defines an interchange representation of topic maps defined in terms of an SGML (Standard Generalized Markup Language) architecture [Megginson, 1998]. A topic map is basically an SGML (or XML (Extensible Markup

Language)) document in which different element types, derived from a basic set of architectural forms, are used to represent topics, occurrences of topics, and relationships (associations) between topics. The key concepts are the *topic* (and *topic type*), the *topic occurrence* (and *occurrence role type*), and the *topic association* (and *association type* as well as *association role type*). Other concepts which extend the expressive power of the topic map model are those of *scope*, *theme*, *public subject* and *facet*.

Note: This short overview about topic map concepts provides the basics only. Application examples can be found in [Rath/Pepper, 1999a], [Pepper, 1999a], [Pepper, 1999b], and [Ksiezyk, 1999].

Topics

A *topic*, in its most generic sense, can be 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. In the words of the standard, the term *topic* refers to the element in the topic map instance (the *topic link*) that represents the subject being referred to. *Examples of topics are*: USA, Pennsylvania, Philadelphia, William Penn.

A topic should have one or more *topic types*. Topic types are a typical class-instance relation and they are themselves defined as topics by the standard. Having topic types as topics the expressive power of topic maps is used to say more about the type. *Examples of topic types are*: country, state, city, person.

Topic characteristics

Every topic has two characteristics (or at least one of them): a *topic name* and an *occurrence*.

The topic name consists of three parts: the *base name*, the *display name*, and the *sort name*. Only the base name is required. *Examples of topic names (base / display / sort) are*: U.S.A. / USA / United States of America.

An occurrence is a link to an information resource that is somehow relevant to the topic. The linked resource is typically an information object outside the topic map. *Examples of occurrences are*: chart of the USA, article about Pennsylvania, video about Philadelphia, portrait of William Penn.

Every occurrence belongs to one *occurrence role type*. Occurrence role types are — as topic types — themselves topics. *Examples of occurrence role types are*: chart, article, video, portrait.

Associations

The real power of topic maps results from *associations* between topics.

Examples of associations are: Pennsylvania is in USA, Philadelphia is in Pennsylvania, Philadelphia was founded by William Penn.

Each association has one *association type*.

Examples of association types are: is in, was founded by.

Each topic that participates in an association plays a role. The role is described by an *association role type*.

Examples of association role types are: state / country, city / state, city / person.

Both association types and role types are again topics.

Scopes

The concept of *scope* is important to avoid ambiguities between topics and their characteristics. Any assignment of a characteristic to a topic is considered to be valid within certain limits, which may or may not be specified explicitly. The limit of validity of such an assignment is called its scope. A scope is defined in terms of *themes* and themes are topics.

Examples of scopes are: to distinguish between "Paris" in France and "Paris" in Texas, assign the scopes "France" and "USA" to the two topics.

Identity

Merging of topic maps requires a way of establishing the identity between seemingly disparate topics from different maps. The specification of *identity attributes* on the topic elements that address the same *public subject* is the explicit solution the standard offers. The other solution is implicitly through the *topic naming constraint* which states that any topics that have the same name in the same scope refer to the same subject.

Facets

Facets provide a mechanism for assigning property-value pairs to information resources without modifying them. A facet is a property; its values are called *facet values*.

The missing pieces: An overview

During the years of its gestation the topic map model changed many times — from an extremely high level of generality to much more specific models designed to be used solely for navigation. The final result is — like most standards — a compromise. The working group believes that it offers an optimal balance between extreme power and flexibility on the one hand and sufficiently well-defined semantics on the other.

The members of the working group always had in mind that the standard has to be implementable, and they tended towards a more general model for both implementability and applicability reasons. They knew that first practical applications might uncover concepts which are not explicitly described in the standard, but they felt it was more important to have a base standard approved

and published than to delay publication any longer merely to add further refinements. Adapting the standard to the XPointer (or XPath) addressing format — as soon as it becomes a W3C (Word Wide Web Consortium) recommendation — is already on the agenda of the working group.

The STEP Group[1] started investigating topic map applications in autumn 1998 in the context of reference works (especially encyclopedias and dictionaries). Applying topic maps to encyclopedias is quite natural: Topic maps model knowledge structures and lexicons represent large parts of the 'knowledge' of society. Thus this application field is a perfect candidate for detecting shortcomings and finding improvements.

Separating the declarative part

Topic maps are a well-designed standard for modelling semantic information networks. The topic map specification defines the basic concepts, and almost everything in the map is itself a topic. Even the 'objects' declaring a topic map are topics, namely themes, topic types, occurrence role types, association types, and association role types. Having such recursive declarations makes perfect sense when the goals are to limit the concepts to a sensible minimum and make topic maps self-contained and self-documenting.

But the standard does not provide a name or definition for the list of declarative 'objects' of a map and this can lead to some confusion: Users often mix up 'declarative' 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 user access rights: As long there is no distinction, different rights cannot be assigned to the map.

A separate declarative part could also be used for defining classes of topic maps that share a common set of topics for types with predefined semantics.

The standard therefore stands in need of a formally defined construct that covers the declarative part of a topic map.

Applying theoretical background

The most interesting constructs in topic maps as far as representing knowledge structures is concerned are associations. Because these are in fact relations it makes sense to take a look at mathematics and apply some of the theoretical background of relations. Furthermore the scientific fields of linguistics and philosophy may provide additional taxonomies.

The concepts that we find could lead to predefined basic association types and association properties. Neither of these are covered by the standard today,

1 The STEP Group consists of STEP Electronic Publishing Solutions GmbH (Rimpar, Germany), STEP Infotek AS (Oslo, Norway), STEP Electronic Publishing Kft (Budapest, Hungary), STEP Poland Ltd. (Warsaw, Poland), and STEP-DPSL Ltd. (Swindon, UK).

but they could offer much more precise semantics in the maps. The topic map template will be the ideal place to define them.

Class-instance relation is not enough

All topics, occurrences, and associations can be seen as instances of classes (types). The classes themselves are expressed as topics.[2]

This class-instance relationship is in fact merely a syntactically privileged association type, as the standard makes clear:

> The class-instance relationship ... could alternatively be established by a topic association link whose semantic is the relationship between a class and an instance of that class.

This means that the class-instance relation is an association type predefined by the standard. Any topic map software has to support it as a built-in function, e.g. by displaying the name of the referenced topic as the name of the type.

If we are looking at the class-instance relation from an object oriented view, then there is a justifiable demand for a superclass-subclass relationship as well. However, the standard explicitly declares that such a relationship has to be user-defined. Here are the relevant quotes:

> The topic relationships established by the types attribute are not superclass-subclass relationships. They are only class-instance relationships.
>
> Superclass-subclass relationships between topics can be asserted by topic association links that have been user-defined for that purpose.

STEP's experiences made with the encyclopedia applications show that the superclass-subclass relationship is a very powerful mechanism for performing inferencing, i.e. deriving implicit information about the current 'object'. The implicit information can be used when querying the map or when declaring and/or checking consistency constraints. And because these features should be an integral part of a topic map software a user-defined and therefore application-specific solution is too weak.

Questions of consistency

The standard has almost nothing to say on the subject of validation and consistency. The "Conformance" section of the standard focuses on the

2 NB: The recursion "a topic has a type which is a topic which has a type" stops if no type is assigned. This is possible because the type is an optional attribute of the topic, occurrence, and association. If the attribute is not specified, the meaning is that the 'object' has no more specific type (i.e. belongs to no more specific class) than that of the base class to which it belongs ('topic', 'occurrence', or 'association', respectively).

understanding of the defined constructs, the interchange syntax, and import/export of topic maps. But nothing more, as this excerpt from the standard shows:

> This International Standard constrains neither the uses to which topic maps can be put, nor the character of the processing that may be applied by a conforming application.

A topic map author (or authoring team) needs system support when developing a map with 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.

Topic map templates

The ISO working group has already responded to the need to be able to separate the declarative part of a topic map. It coined the term *topic map template* for a topic map that only consists of topics that are declared in order to be used as types in a class of topic maps. At the present time this term is only 'semi-official', since the concept has not yet been refined and added to the standard.

What is a topic map template?

A topic map template consists of all constructs which have a declarative meaning for the map (see figure 1). These are all the topics used as themes and as types for

- other 'regular' topics,
- occurrence roles,
- associations,
- association roles,
- facets, and
- facet values.

As we will see later, the class hierarchy information and consistency constraints will also become part of a topic map template.

The topic map designer should mark the topics in the template to show which kinds of type they could be used for in the 'real' map. This can be done by either grouping the topics (see below "Template modules") or by assigning attribute values. The latter approach provides more flexibility for marking topics for more than one kind of type.

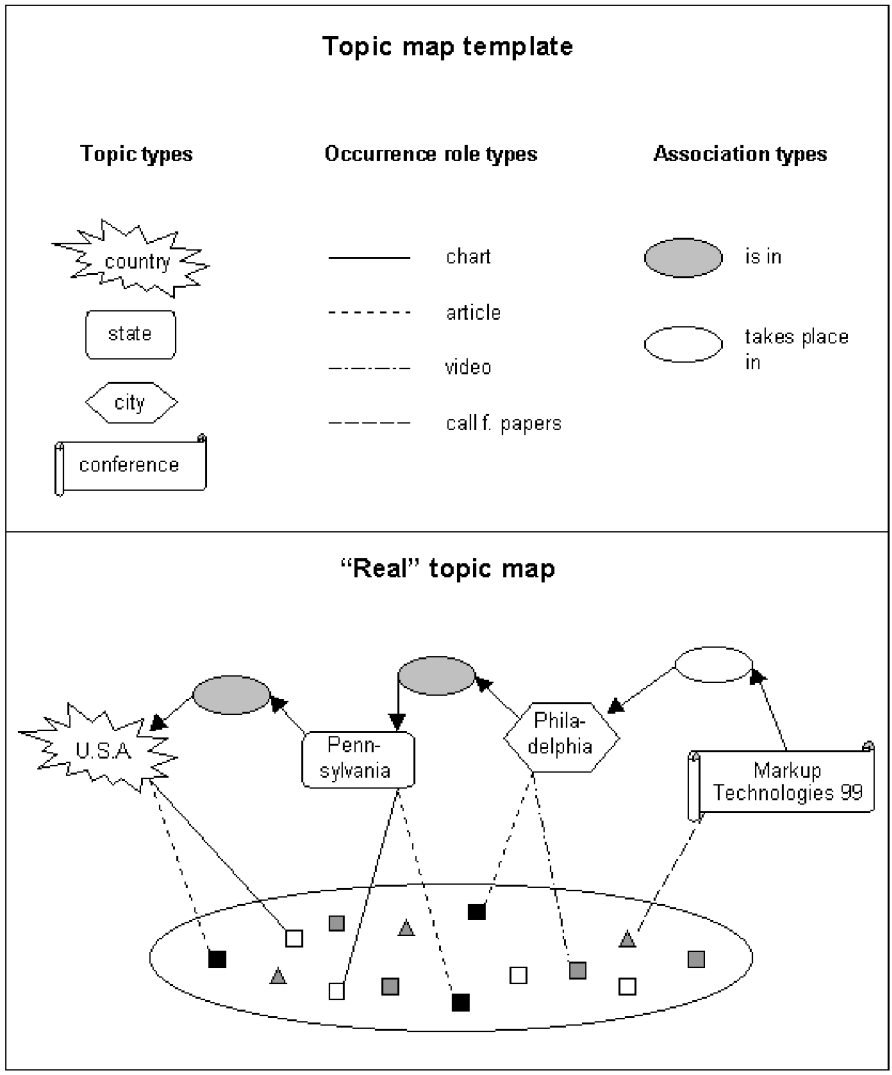

Figure 1 Topic map template

In any case it is clearly important that the topics of the template can be distinguished somehow from the topics of the topic map instance(s) belonging to the class of topic maps defined by the template, and that the template becomes a 'manageable' object with its own (public) identifier, owner, version number, etc.

Using templates in topic maps

The topic map template — which is a topic map — can be copied into or referenced by another topic map.

The copied template acts as a starting point for a new map containing all the themes and types which will be extended during the further development of the map.

The referenced template provides the basic themes and types which are used by the referencing map. A referenced template makes use of the merging features of topic maps defined by the standard. Thus more than one template could be referenced. Though the precondition for merging is the existence of carefully worded subject identities.

Template modules

It might be meaningful for a template to consist of sub-templates to modularize the design. Candidates for template modules are

- clusters of all 'typing' topics for the various 'objects' as listed above, e.g. all topics which are to be used as topic types,
- the class hierarchy information, or
- the consistency constraints.

But this is only one possibility. How the declarations will be clustered in modules depends to a large degree on application-specific requirements. The only important thing is that the template can easily identified and separated from the real map.

Distributing the design and creation tasks

The design and creation of topic maps can now be split up into subtasks because of the availability of templates and template modules. Furthermore, user access rights of user groups as well as roles can be assigned.

The tasks of the designer might be:

- declaration of themes,
- declaration of all topics which are candidates for types,
- marking the topics with the kind(s) of type it is intended for,
- defining the consistency constraints.

The tasks for the editor might be:

- definition of the 'real' topics,
- definition of associations between them,
- establishing the occurrence links to the relevant information objects,
- checking the consistency of the map by applying the consistency constraints (this will be an automatic process).

The assignment of facets can be seen as a completely separate task.

Role of topic map templates for ISO/IEC 13250

The concept of templates offers the ISO working group the possibility of defining various templates which are specific for different application areas. These templates would contain built-in types (i.e. topics) and association types with predefined semantics which could be supported by 'template-conformant' applications.

Such templates could be published as annexes to the standard or as separate standards, as has already been done with SGML DTDs (e.g. ISO 12083).

Association taxonomy

The investigation of the theoretical backgrounds of relations leads us to the domains of mathematics, linguistics, artificial intelligence, and philosophy. All these scientific fields deal with knowledge representation and knowledge structures in one way or another.

We will concentrate on two issues from this broad research area: relations in mathematics (i.e. the abstract properties of associations) and relationship types in artificial intelligence and linguistics (i.e. specific classes of associations).

Association properties

The most important relations — in the mathematical sense — are the *binary relations*.[3]

Definition: A *binary relation between the sets A and B* is: every subset R of $A \times B$ ($R \subseteq A \times B$).

The properties which are of interest for topic maps are only effective for a restricted kind of relations.

Definition: A *binary relation in M* is: a binary relation R with $A = B = M$, thus $R \subseteq M \times M$.

A binary relation is also a binary predicate.

Definition: A *predicate (relation) R is fulfilled (true)* for $x \in A$ and $y \in B$ if and only if $(x, y) \in R$.

$(x, y) \in R$ can be abbreviated as xRy.

Now we can define the properties for relations in M.

Property of R	*Definition*
reflexive	$\forall\ x \in M: xRx$
symmetric	$\forall\ x, y \in M: xRy \Rightarrow yRx$

3 N-ary relations and 'elementary associations' (in which the number of arguments cannot be further reduced) with more than two arguments are not covered in this paper, because they form a more complex class.

Property of R	*Definition*
transitive	$\forall\ x, y, z \in M$: $xRy \wedge yRz \Rightarrow xRz$
anti-reflexive	$\forall\ x \in M$: $\neg\ xRx$
anti-symmetric	$\forall\ x, y \in M, x \neq y$: $xRy \Rightarrow \neg\ yRx$
connex	$\forall\ x, y \in M$: $xRy \vee yRx$

Certain combinations of these properties define special classes of relations, of which there are four:

Definitions:

- R is an *equivalence relation*: R is reflexive, symmetric, and transitive.
- R is an *partial ordering relation*: R is reflexive, anti-symmetric, and transitive.
- R is a *total order relation*: R is reflexive, anti-symmetric, transitive, and connex.
- R is a *strong order relation*: R is anti-reflexive, anti-symmetric, and transitive.

Some examples of specific relations will serve to illustrate the various properties and classes of relations ($M = \{0, 1, 2, 3, ...\}$).

Property / class	*is denominator of*	*is less than equal*	*is less than*
reflexive	yes	yes	no
symmetric	no	no	no
transitive	yes	yes	yes
anti-reflexive	no	no	yes
anti-symmetric	yes	yes	yes
connex	no	yes	no
order rel.	yes	yes	no
total order rel.	no	yes	no
strong order rel.	no	no	yes

Why is all the theory relevant for topic maps? Let us analyze the association type “geographical_object *is in* geographical_object”. It is transitive, anti-reflexive, and anti-symmetric; thus it is a strong order relation. Topic map software that was aware of these facts (i.e. the properties of this particular association type) would be capable of automatically deriving implicit knowledge from the map.

An example: From the given associations

- Pennsylvania *is in* USA
- Philadelphia *is in* Pennsylvania
- Pittsburgh *is in* Pennsylvania

the topic map software can derive that

- Philadelphia *is in* USA
- Pittsburgh *is in* USA
- USA *is not in* Pennsylvania
- Philadelphia *is not in* Philadelphia
- etc.

It is obvious that the most informative statements of this example derive from the property of transitivity.

Another example: Let us analyze the association type "street *is parallel to* street". It is reflexive, symmetric, and transitive; thus it is an equivalence relation.

If we have the associations

- Park Avenue *is parallel to* Madison Avenue
- Madison Avenue *is parallel to* Fifth Avenue

then the associations

- Park Avenue *is parallel to* Fifth Avenue
- Fifth Avenue *is parallel to* Madison Avenue
- etc.

can easily be derived. The relevant information comes from the symmetry and again from the transitivity property.

The examples show that a simple set of association properties, i.e. the relation properties introduced above, would give more 'knowledge' from the topic map than explicitly coded in it. This means that the map becomes smaller, that the effort creating a map will be minimized, that possible coding errors will be reduced tremendously, and that the inferencing capabilities of the topic map's query engine will be greatly enhanced. Furthermore the consistency checking can make use of the property information, which again improves the quality of the map.

Basic association types

The previous section introduced the basic association properties. This section considers whether basic association types would also make sense.[4]

A lot of research has been done in the area of knowledge structures[5]. Some of the research work covers relations in the lexicon [Iris et al., 1985]. Others

4 Steve Pepper (STEP Infotek, Norway) provided substantial input to this section.

investigated the linguistic relations in the semantic of English language [Fellbaum, 1998], [Wordnet, n.d.]. The results are a summary of relations that express the basics concepts of knowledge representation.

A large class comprises the *part-whole* or *holonymy/meronymy* relations. [Winston et al., 1987] and [Chaffin et al., 1988] list six and seven subclasses of holonymy respectively:

- component-object (e.g. branch/tree)
- member-collection (e.g. tree/forest)
- portion-mass (e.g. slice/cake)
- stuff-object (e.g. aluminum/airplane)
- feature-activity (e.g. paying/shopping)
- place-area (e.g. Philadelphia/Pennsylvania)
- phase-process (e.g. adolescence/growing up)

Iris et al [Iris et al., 1985] reduce this to four basic subclasses:

- functional-part (← phase-process, feature-activity)
- segmented-part (← component-object, place-area)
- collection-member (← member-collection, stuff-object)
- subset (← portion-mass)

According to [Iris et al., 1985] only *segmented-part* and *subset* exhibit transitivity. Individual *functional-part* or *collection-member* relations could be transitive, but the property does not apply to these classes as a whole.

We can conclude that the *part-whole* class with its subclasses *functional-part*, *segmented-part*, *collection-member*, and *subset* should be predefined association types — declared in a template.

Some other relevant relationship types are

- synonymy (e.g. equals, identical to),
- similarity (e.g. similar to),
- order (e.g. less than, older than, closer to),
- result-agent (e.g. "object" is caused by "agent", "artwork" created by "artist", "painting" painted by "painter"),
- tool-agent (e.g. "tool" is used by "agent", "instrument" is played by "musician"), and
- strict implication[6] (e.g. "activity 1" implies "activity 2", "snoring" implies "sleeping").

---

5 See [Ringland/Duce, 1988] for an introduction and extensive bibliography.

6 Definition of *strict implication*: A proposition P entails a proposition Q ($P \Rightarrow Q$) if and only if there is no conceivable state of affairs that could make P true and Q false.

The *synonymy*, *order*, and *strict implication* are transitive relations. *Synonymy* and *similarity* are symmetric. For every *result-agent* and *tool-agent* relation exists an inverse one ("agent" causes "object", "agent" uses "tool"). Strict implication is non-symmetrical: you can sleep without snoring, but you cannot snore without sleeping! All these relations are candidates to be predefined association types that are declared in a template.

The contributions from linguistics introduce further subclasses for *synonymy* relations (thesauri: [Aitchison et al., 1997]) and build a class hierarchies with the *hyponymy* for nouns and the *troponymy* for verbs (dictionaries: [Fellbaum, 1998], [WordNet, n.d.]). Both *hyponymy* and *troponymy* represent the "is a" or "is a kind of" relation, which is already covered by the topic type construct. The *synonymy* subclasses seemed to be quite specific, thus there is no need to have them as predefined association types. They are in any case more appropriately handled through the use of multiple topic names.

Class hierarchies

The realization of the need for class hierarchies stems from STEP's encyclopedia projects. A topic map for a lexicon contains a very large number of topics (typical orders of magnitude are hundreds of thousands or millions) and associations (even more). But most of the topic, association, and occurrence role types can be reduced to a small number of 'super-types' — as we have already seen in the previous section.

Superclass-subclass

The superclass-subclass relationship of topic types, association types, and occurrence role types go hand in hand, as following examples shows:

- Topic types: (person) → (artist, ...) → (painter, sculptor, writer, poet, composer, ...); (object) → (artwork, ...) → (painting, sculpture, novel, poem, opera, ...)
- Association types and occurrence role types: (object "was caused by" person) → (artwork "was created by" artist) → (opera "was composed by" composer)

Class hierarchy and association type properties

The class hierarchies become even more important when the end-user navigates or queries the map. If someone would like to know "Which pieces of music were composed by Germans that were influenced by W.A. Mozart?", it is very likely that this information is not exactly part of the map. But with just a few topics, transitive associations, and a class hierarchy the answer can be found very easily.

The facts of the map:

- *The topic type (class) hierarchies:* person → composer; piece of music → opera; geographical object → country; geographical object → city.
- *The transitive association type:* "geographical object" is in "geographical object".
- *Other association types:* "composer" has composed "piece of music"; "person" was influenced by "person"; "person" was born in "geographical object".
- *The topics:* W.A. Mozart (composer); R. Wagner (composer); L. van Beethoven (composer); Bonn (city); Leipzig (city); Germany (country); Lohengrin (opera).
- *The associations:* Bonn is in Germany; Leipzig is in Germany; L. van Beethoven was born in Bonn; R. Wagner was born in Leipzig; Lohengrin was composed by R. Wagner; R. Wagner was influenced by W.A. Mozart.

The algorithm how the topic map software would find the solution with these facts could work as follows:

- Extraction of the known topics from the query: Germany, W.A. Mozart.
- Extraction of the types of the unknown topics: person (X), piece of music (Y).
- Extraction of the association types: born in, influenced by, composed by.
- Finding the missing topics using the associations and class hierarchies:

  X is born in Germany (country is also a geographical object) ⇒ X is born in Bonn or Leipzig (both cities are in Germany) ⇒ X is L. van Beethoven or R. Wagner (both composers are also persons);

  X was influenced by W.A. Mozart (composer is also a person) ⇒ R. Wagner was influenced by W.A. Mozart (both composers are also persons) ⇒ X is R. Wagner;

  Y was composed by X ⇒ Y was composed by R. Wagner ⇒ Lohengrin was composed by R. Wagner (opera is also piece of music) ⇒ Y is Lohengrin.

This very simple example shows the power of combining class hierarchies with properties of association types (here transitivity). As already stated above, both class hierarchies and association type properties are the basis for compact topic maps, minimized creation and maintenance efforts, and a reduction of coding errors.

This supports our contention that the concept of class hierarchies should be a predefined association type of topic map template ensuring the correct built-in interpretation by the topic map software.

Validation of consistency

All the previously introduced concepts extend topic maps in ways that increase their expressive power and ease creation and maintenance efforts. In addition to this, the topic map developer wants to have something at hand to help ensure the quality of the map. The information provided by a topic map based on the standard architecture is not enough — the developer asks for validation concepts.

Real life topic maps will consist of millions of topics and associations. Checking a map of such a size manually is clearly impossible, and yet checking is absolutely necessary for both proof-reading and quality assurance. It is obvious that both the designer and the editor need access to an automatic process that can validate a topic map against a set of consistency rules.

The validation is the task of the topic map development environment (e.g. an editorial system). It should be performed continuously or on demand — like structure validation against the DTD in an SGML/XML editor.

The standard has almost nothing to say on the subject of validation and consistency. The "Conformance" section of the standard focuses on the understanding of the defined constructs, the interchange syntax, and import/export of topic maps. But nothing more, as this excerpt from the standard shows:

> This International Standard constrains neither the uses to which topic maps can be put, nor the character of the processing that may be applied by a conforming application.

This shows that we have to develop a schema language for the definition of the consistency constraints.

Consistency constraints

The topic map standard provides the architectural element types which can be used in a derived DTD (Document Type Definition). However, the degree to which semantics can be modelled in a DTD and through content models is rather limited. A topic map will consist of a large number of 'independent' elements which are connected by links and not by element structures.

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 constraints are a set of predefined association types declared in the template.

What should be validated?

Constraints may be assigned to three potential layers:

- topic map modeling,
- user interface for topic maps, and
- operations on the map.

Here, we focus on the topic map modeling layer.

Associations: The most important candidates for validation are the associations. This is obvious because they are the key concept and carry a large number of parameters which might be 'misused'.

The starting point is the association type. This controls which association role types can be combined. Beside the possible combination(s) the number of the various roles within these combinations might be of interest.

The association role type in turn governs the set of topic types which may be referenced.

It is necessary that the constraint schema brings the association type, the role type, and the topic type into a meaningful combination.

An example:

Association type	*is in* (geographical containment)
Valid association role types	one *containee:* one *container*
Valid topic type combinations	*city: (country | state | county)* *county: (state | country)* *state: (country)*

Occurrences: The assignment of the proper information resource types — if type information is provided by the editorial system — to the occurrence role types is also of interest as well as the meaningful combination of topic types and occurrence role types.

An example:

Topic type:	*person*
Valid occurrence role types:	*biography*, *portrait*
Valid resource types for *biography:*	SGML/XML instance with public identifier "-//STEP//DTD biography//EN"
Valid resource types for *portrait:*	object types TIFF, GIF, JPEG

Scopes: Furthermore the correct use of scopes and especially the combination of different scopes might be checked.

The topic type could restrict the possible scopes for the topics, their topic names, base name, display name, sort name, and their occurrences.[7]

The association types might restrict the meaningful scopes for the associations also. The combination of the meaningful scopes of the association and the referenced topics should be checked also because the association type is closely related to the possible types of the referenced topics.

An example:

Themes:	*before Einstein's theory of relativity*, *after Einstein's theory of relativity*
Topic types:	*physical law*, *mathematical axiom*
Occurrence role types:	*definition*
Constraints:	The scope *before Einstein's theory of relativity* might be used for occurrences with role *definition* for topics of type *physical law*; but it must not be used for *definitions* of *mathematical axioms*.

Topic names: For reasons of completeness checking of the topic names should also be possible. Topic names might be checked against text patterns or against database entries. The constraints will be governed by the topic type in question.

An example:

Topic types:	*component in assembly group*, *chemical substance*
Constraints:	Check base name of topic of type *component* against pattern (regular expression) "`P[0-9]+[A-D][E-G][0-5]`"; check sort name of *chemical substance* against table "substance names" in chemical database.

All type combination constraints might limit the number of superclasses and/or subclasses of the affected types.

7 Because assigning scopes to the topic or the topic name are just shortcuts for assignments to every name or occurrence, the set of scopes of the topic must be a superset of the scopes for the names and occurrences, and the set of scopes of the topic name must be a superset of the scopes for the individual names.

Conclusions

The new topic map standard ISO/IEC 13250 defines a model and architecture for the semantic structuring of link networks. It can be seen as a base technology for modeling knowledge structures. The standards working group defined topic maps in such a way that a limited but implementable set of core concepts express the necessary semantics.

The STEP Group has investigated how topic maps can be applied to reference works and uncovered some concepts which are not made explicit in the standard:

- ability to separate the declarative part from the 'real' map,
- predefined association types and association type properties,
- class hierarchies for types, and
- consistency constraints as input to map validation.

The paper has explained these concepts and presented meaningful solutions.

First experiences have shown that the part of a topic map made up by all topics used as themes and types by other 'objects' in the map should be clustered somehow. For this purpose the term *topic map template* was coined by the ISO working group. Templates can be used as starting points for new maps or can be used by reference in order to provide all the themes and types the map needs. Standardizing topic map templates will offer base topic maps for specific application areas and could form the basis of semantic application profiles.

We looked at related academic fields like mathematics, linguistics, and philosophy to get some substantial input about relations. The results are a list of association type properties which give important hints to the topic map software and a list of basic association types which could act as built-in superclasses.

The introduction of the superclass-subclass relationship was the logical consequence.

Another technical issue covered by the paper is the validation problem. Topic maps might become rather big with millions of topics, occurrences, and associations. Manual consistency checking will be impossible. All the previously defined concepts open the possibility for sophisticated rule-based validation of topic maps. The proposed consistency constraints are those rules which declare the semantics not expressible with DTDs and which control the validation process.

A couple of examples proved that standardizing the missing concepts as predefined topic map templates will help both the topic map developer and the topic map user. The improvements were presented on a level that they can be used as input to the ISO working group for further discussions.

Received 17 December 1999
Accepted 15 February 2000

References

[Aitchison et al. 1997] Aitchison, J., A. Gilchrist, and D. Bawden. *Thesaurus construction and use — a practical manual*. 3rd edition, London: Aslib, 1997.

[Chaffin et al. 1988] Chaffin, R., D. J. Hermann, and M. Winston. "An empirical taxonomy of part-whole relations: Effects of part-whole relation type on relation identification". *Language and Cognitive Process* 3 (1988).

[Fellbaum 1998] Fellbaum, C., ed. *WordNet — An Electronic Lexical Database*. Cambridge: MIT Press, 1998.

[ISO 13250:2000] International Organization for Standardization. *ISO/IEC 13250, Information technology — SGML Applications — Topic Maps*. Geneva: ISO, 2000.

[ISO 10744:1999] International Organization for Standardization. *ISO/IEC 10744:1999 Information technology — Hypermedia/Time-based Structuring Language (HyTime)*. Geneva: ISO, 1997.

[ISO 2788:1986] International Organization for Standardization. *ISO 2788:1986. Guidelines for the establishment and development of monolingual thesauri*. Geneva: ISO, 1986.

[ISO 5964:1985] International Organization for Standardization. *ISO 5964:1985. Guidelines for the establishment and development of multilingual thesauri* Geneva: ISO, 1985.

[Iris et al. 1985] Iris, M., B. Litowitz, and Evens, M. "Problems of the part-whole relation" *Relational models of the lexicon*, ed. M. Evens. Cambridge, 1988.

[Ksiezyk 1999] Ksiezyk, R. "Trying not to get lost with a Topic Map". *Proceedings of XML Europe 99 Conference*. Alexandria, VA: GCA, 1999.

[Megginson 1998] Megginson, D. *Structuring XML Documents*. Prentice Hall, 1998.

[Pepper 1999a] Pepper, S. "Euler, Topic Maps, and Revolution". *Proceedings of XML Europe 99 Conference*. Alexandria, VA: GCA, 1999.

[Pepper 1999b] Pepper, S. "Navigating Haystacks, Discovering Needles". *Markup Languages* 1.4 (1999).

[Ranganathan 1967] Ranganathan, S.R. *Prolegomena to Library Classification*. Bombay: Asia Publishing House, 1967.

[Rath/Pepper 1999a] Rath, H.H., and S. Pepper. "Topic maps: Knowledge navigation aids". *XML Handbook*, ed. C. F. Goldfarb and P. Prescod. 2nd edition. Prentice Hall, 1999.

[Rath/Pepper 1999b] Rath, H.H., Pepper, S. "Topic Maps: Introduction and Allegro". *Proceedings of Markup Technologies 99 Conference*. Alexandria, VA: GCA, 1999.

[Rath 1999] Rath, H.H. "Technical Issues on Topic Maps". *Proceedings of Metastructures 99 Conference*. Alexandria, VA: GCA, 1999.

[Ringland/Duce 1988] Ringland, G.A., and D. A. Duce. *Approaches to Knowledge Representation: An Introduction*. Research Studies Press/John Wiley, 1988.

[Ruggles 1997] Ruggles, R.L., ed. *Knowledge management tools*, Boston: Butterworth-Heinemann, 1997.

[Streich 1999] Streich, R. "Techniques for managing collections of interrelated text modules". *Markup Languages* 1.2 (1999).

[Vickery 1960] Vickery, B.C. *Faceted classification: a guide to construction and use of special schemes*, London: Aslib, 1960.

[Vickery 1966] Vickery, B.C. *Faceted classification schemes*. New Brunswick: Rutgers, 1966.

[Winston et al. 1987] Winston, M.E., R. Chaffin, and D. Hermann. "A taxonomy of part-whole relations". *Cognitive Science* 11 (1987).

[Wordnet n.d.] WordNet: "A Lexical Database for English", Cognitive Science Laboratory, Princeton University, Princeton, NJ, http://www.cogsci.princeton.edu/~wn/.

[ANSI/NISO 1993] ANSI/NISO: *Z39.19. Guidelines for the construction, format and management of monolingual thesauri*. Bethesda: ANSI/NISO 1993.