Technical Issues on Topic Maps

Hans Holger Rath

Director Consulting
STEP Electronic Publishing Solutions GmbH
Technologiepark Würzburg-Rimpar, Pavillon 7
D-97222 Rimpar, Germany

Phone: +49.9365.8062.0, Fax: +49.9365.8062.66
E-mail: consulting@step.de, Web: http://www.step.de/

Abstract: A topic map structures link networks as SGML/XML structures data. Applying SGML/XML markup to raw data creates information. Applying a topic map to an information pool creates knowledge structures. The paper will cover three technical key issues about topic maps: collecting the declarative part of a map in the "topic map template", checking the consistency of a map using constraints, and automatic generation of a topic map from a given set of structured information resources using generation rules.

1 Introduction

The new standard ISO/IEC 13250:1999 Topic Maps describes a model and declares an interchange format for topic maps. The initial ideas – which date back to the early 1990's – wanted to model intelligent electronic indexes and support for merging them. But during the years topic maps became a much more powerful model which is not restricted to indexes anymore.

A topic map organizes large sets of information resources. It builds a structured semantic link network over the resources. The network allows easy and selective navigation to the requested information. Topic maps are the "GPS for 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.

The paper is not an introduction into topic map concepts but it summarizes the base model and focuses on selected technical issues.

1.1 The topic map model

The standard defines an interchange representation of topic maps defined in terms of an SGML architecture. A topic map is basically an SGML (or XML) document in which different element types, derived from the basic set of architectural forms, are used to represent topics, occurrences of topics, and relationships between topics.

The key concepts are:

- *Topic* (and *topic type*): 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. Every topic shall have one or more topic types. The topic types are a typical class-instance relation and they are themselves defined as topics which allows self-documenting topic maps.

  Examples (topics): Canada, Québec, Montréal, MetaStructures 99.

  Examples (topic types): country, province, city, conference.

- Topic *name*: The topic name consists of three parts: the base name, the display name, and the sort name. Only the base name is required.

  Examples (base / display / sort name): Quebec / Québec / quebec.

- Topic *occurrence* (and *occurrence role type*): An occurrence is a link to an information resource that is somehow relevant to the topic. Every occurrence plays a role which is expressed by the occurrence role type and the occurrence's *scope*. Occurrence role types are themselves topics.

  Examples (occurrence role types): chart, article, video, call for participation.

- Topic *association* (and *association type* as well as *association role type*): An association describes a relationship between two or more topics. Every

association belongs to a type. Each associated topic plays a role in the association. Both association type and association role type are again topics.

Examples (associations): Québec is in Canada, Montréal is in Québec, MetaStructures 99 takes place in Montréal.

Examples (association types): is in, takes place in.

Examples (association role types): province / country, city / province, conference / city.

- *Scope* (and *theme*): Any assignment to a topic (name, occurrence, association) 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 which are also topics.

  Examples (scopes): to distinguish between “Paris” in France, “Paris” in Texas, and “Paris” the Greek hero, assign the scopes “France”, “USA”, and “Greek mythology” to the three topics.

Other concepts which extend the expressive power of the topic map model are *public subject* (merging of topic maps) and *facet* (assigning property-value pairs to resources).

1.2 Selected technical issues

The summary of the topic map model as defined in ISO/IEC 13250 shows that the standard 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 ISO working group had always in mind that the model has to be implementable. Therefore they tended towards a more general model because of both implementability and applicability reasons. That explains why the standard does not define more semantic details.

This paper will present three technical issues which are either not covered by the standard (see chapter 2 *Topic map templates* and chapter 3 *Consistency checking using constraints*) or are needed to use existing “legacy data” in topic maps (see chapter 4 *Automatic generation*).

2 Topic map templates

Topic maps are a well-designed standard to model semantic information networks. It defines the basic concepts whereby mostly everything in the map is 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 this recursive declarations makes perfect sense if “limit the concepts to sensible minimum” and “a topic map has to be self-contained” are the goals.

But the standard does not provide a name or definition for the list of declarative “objects” of a map. Thus mixing up “declaring” topics with “regular” topics happens very often and very easily during discussions. In addition to that the different tasks of topic map design, creation, and maintenance are hard to distinguish and to separate.

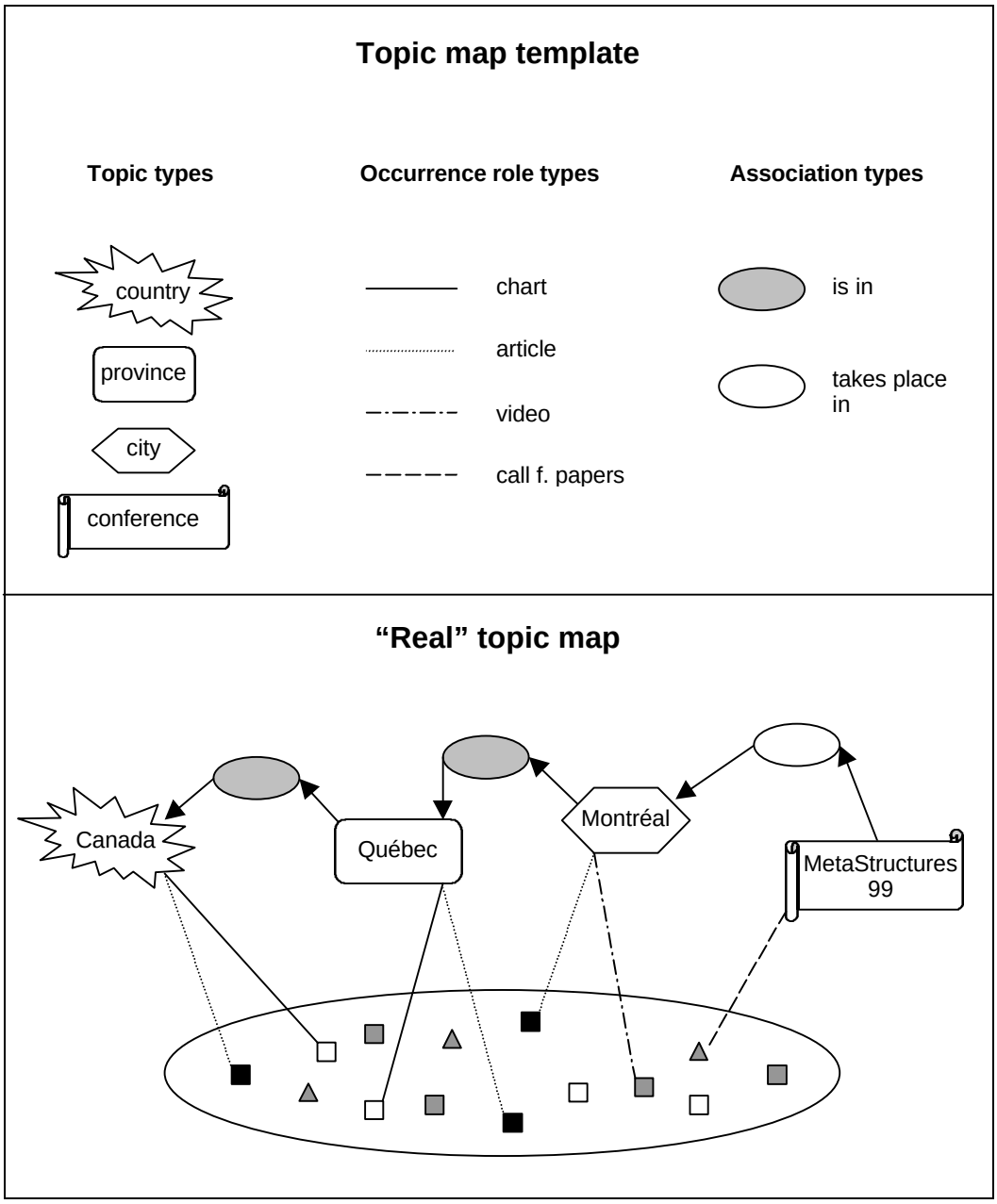

Figure 1 – Topic map template

The ISO working group reacted already on the need for the separation of the declarative part of a topic map. It introduced the term *topic map template*. At the moment, this term is "semi-official", because it did not found its way into the standard, which was approved by the national ISO member bodies before.

2.1 What is a topic map template?

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

- other "regular" topics,

- occurrence roles,
- associations,
- association roles,
- facets, and
- facet values.

As we will see later the consistency constraints should also become part of a topic map template.

The topic map designer shall mark the topics in the template for which kind of type they could be used in the "real" map. This can be done by either grouping the topics (see below *Template modules*) or by assigning attribute values. The latter gives more freedom to mark topics for more than one kind of type.

In any case it seemed to be important that the topics of the template can be distinguished somehow from the topics of the "real" topic map and that the template becomes a "manageable" object with its own (public) identifier, owner, version number, etc.

2.2 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.

2.3 Template modules

It might be meaningful that a template consists 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 shall be used as topic types, or
- the consistency constraints.

But this is only one possibility. How the declarations will be clustered in modules highly depends on the application specific requirements.

2.4 Splitting up the tasks for design and creation

The design and creation of topic maps can now be split up into subtasks because of the availability of templates and template modules. 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 of type the could be used for,
- definition the consistency constraints.

The tasks for the editor might be:

- definition of the “real” topics,
- definition of the 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).

2.5 Role of topic map templates for ISO/IEC 13250

The concept of templates offers the ISO working group the possibility to define various templates which are specific for different application areas. These templates could be published as annexes to the standard or as separate standards like already done with SGML DTDs (e.g. ISO 12083).

3 Consistency checking using constraints

Real life topic maps will consist of millions of topics and associations. A manual check of a map of such a size is impossible, but necessary for proof-reading and quality assurance. It is obvious that both the designer and the editor need an automatic process at hand which validates a topic map against a set of consistency rules.

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

Unfortunately, the standard states almost nothing about 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 following excerpt from the standard proves:

> 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 for the definition of the consistency constraints.

3.1 Consistency constraints

The topic map standard provides the architectural element types which can be used in a derived DTD. Modeling semantics in a DTD and its content models is possible to a certain degree only. 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 rules modeled as architectural element types which "cooperate" with the topic map architectural forms. The constraining elements should be part of the topic map template, as explained above.

3.2 What shall be constrained?

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.

3.2.1 Associations

The most important validation candidate are the associations. This is obvious because they are the key concept and carry a large number of parameters which might be "misused".

The association type is the starting point. It 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 role type itself controls which topic types shall be referenced.

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

An example:

- Association type: *is in* (geographical containment)
- Association role types: one *containee*, one *container*
- Topic type for *containee* and corresponding topic types for *container*:
  city – country, state, county
  county – state, country
  state – country

3.2.2 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*
- Occurrence role types: *biography*, *portrait*
- Information resource types for biography:
  SGML/XML instance with public identifier "-//STEP//DTD biography//EN"

- Information resource types for portrait:
  object types TIFF, GIF, JPEG

3.2.3 Scopes

Furthermore the correct use of the 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.[1]

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 shall be checked also because the association type is in close relation with 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 *definition*s of mathematical *axiom*s.

3.2.4 Topic names

For completeness reasons checking of the topic names should be possible also. Topic names might be checked against text patterns or against database entries. The topic type controls the constraint.

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.

4 Automatic generation

When someone decides to start working with topic maps it is quite likely that the topic map will be applied to already existing information resources. We can compare this process with applying SGML/XML structures to legacy data.

The conversion of legacy data to SGML/XML structured instances is often called an *up translation*. The naming makes sense because the legacy data is not

[1] 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 single names.

structured properly in most cases and the required structuring information has to be extracted from formatting information and/or text patterns.

The automatic generation of topic maps will also use existing information as hints for the creation of the map. After the map has been generated initially further developments and maintenance will be done manually. The process of automatic updating of an existing map might override manual changes or improvements and will therefore be rarely applicable.

4.1 The basic setting

The available input information for the automatic generation is:

- the topic map template with its topic types, occurrence role types, association types, and association role types,
- the address for information resources in the file system, a database, or the Web,
- metadata about information resources (e.g. name, format, classification code),
- structure of information resources which are available as SGML/XML instances.

The automatic generation is controlled by a set of rules collected in a script. The conversion program interprets the script and gets the available information (see list) as input. The result will be a topic map in the exchange syntax of the standard (see figure 2).

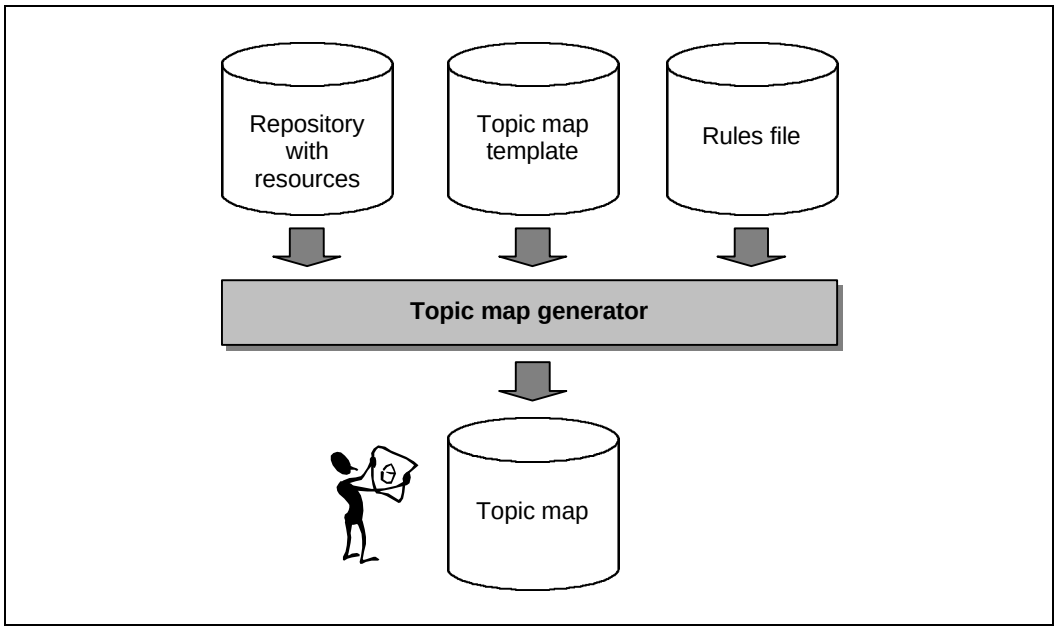

Figure 2 – Automatic generation

But as with all conversions the automatic generation of topic maps has to be checked manually too.

4.2 Identifying topics and occurrences

The first phase of the automatic generation is the identification of information resources which are candidates for topics of a certain type. If such a resource has been detected and the topic is created by the process an algorithm is needed extracting the topic name from the resource or its metadata. Finally the occurrence to the resource is established for the current topic.

The search process could be controlled by a rule like this:

If resource fulfills metadata <condition> and/or
contains structure <element> in <context> containing <content>
then create topic of <type> with name derived from metadata <field> or
name derived from <element> in <context> and
create occurrence to resource with <role>.

Scopes can be assigned to the topics, names, or occurrences also if the resources or their metadata provides suitable scope information.

The result of the first phase is a topic map containing all topics with names, all their occurrences, and maybe all the scopes. Everything which is immediately extractable from the legacy data.

4.3 Identifying associations

The second phase of the automatic generation is the creation of associations between the topics established in the first phase. This is a very complex task because associations model a kind of knowledge structures and the extraction of knowledge from existing data is a task for AI[2] algorithms.

But nevertheless the usage of the consistency constraints lead to a first list of candidates for associations which can afterwards be generated manually. The automatic process interprets the constraints and extracts the information which topic types shall play which association role in which association.

Having these facts, the process is able to generate the lists containing the association types, the association role types, and the candidates for referenced topics.

More sophisticated association generation algorithms cannot be generalized. They will rely on application specific information about or in the resources (metadata, structure and/or text content of instances).

But the value of an automatic generation of associations is in question anyway, because every single generated association has to be checked manually anyway. Thus the work with candidates lists might be more appropriate and more efficient.

---

[2] AI: Artificial Intelligence. Research field in computer science trying to model the cognitive human behavior.

5 Conclusions

The new topic map standard ISO/IEC 13250 provides the concepts and an SGML architecture for 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 a way, that a limited but implementable set of core concepts express the necessary semantics.

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 invented by the ISO working group. Templates can be used as starting points for new maps or can be referenced providing all themes and types the map needs. Standardizing topic map templates will offer base topic maps for specific application areas.

The second technical issue covered by the paper is the validation problem. Topic maps might become rather big with millions of topics, occurrences, and associations. A manual consistency check will be impossible. The paper proposed a couple of rule-based consistency constraints which control the validation process.

Finally, the automatic generation of topic maps from existing data was discussed. Legacy data will become the resources of a map. Again, a rule-based approach controls the process. The available information from the legacy data is used to create topics, their names, and occurrences as well as suggestions for associations.

The presented technical issues about topic maps will – hopefully – improve the understanding of the standard and sensitize for the needed software developments around topic maps.

Acknowledgements

The author would like to thank his colleagues Geir Ove Grønmo, STEP Infotek, Rafal Ksiezyk, STEP Poland, Graham Moore, STEP-dpsl, and Steve Pepper, STEP Infotek as well as all the members of STEP's Reference Works Module Club – the leading European reference works publishers – for their input and open discussions about topic maps.