
From Implicit Patterns to Explicit Templates

Next step for Topic Maps interoperability

Bernard **Vatant** <bernard.vatant@mondeca.com>

Abstract

Claims for Topic Maps interoperability have been standing so far on formal representation of subjects by topics, and identification of topics by controlled vocabularies or Published Subjects. Those features allow topics to be merged on the basis of common identifiers, leading to aggregation of relevant information or knowledge expressed by characteristics of the merged topics.

But what should happen to associations at merging time remains an open issue in many cases. According to current standard processing requirements, only exactly identical associations will be merged. Default standard definition of association templates, there is so far no standard rules for discovering and dealing with associations having similar patterns of roles, or representing the same association type with slightly different patterns. And default proper association merging rules, Topic Maps are at risk to be cluttered up with redundant or non-consistent associations, after several successive merging.

A review of how current Topic Map applications deal with those questions is presented. It's shown that they generally use explicit association templates mechanism to identify, create or check patterns. Although there is no official standard, proprietary solutions are quite similar, and the road to adoption of standard simple rules could be fairly easy. An extension of XTM syntax to express templates, conformant to current practices, is proposed.

An use case is presented, where a Topic Map repository in an integrated information system is updated using a text mining tool, a controlled vocabulary and a workflow of semi-structured documents. Similar, redundant or overlapping associations are likely to be generated, and the update process has to deal with them

Further applications are considered, like checking for template consistency before merging Topic Maps, extracting templates from existing Topic Maps, or developing libraries for specific application domains, extending the notion of Published Subjects to association templates.

Table of Contents

1. Topic Maps and Knowledge Interoperability	2
1.1. A standard XML interchange format	2
1.2. Topic Naming Constraint: interoperability of topics inside a scope	2
1.3. Published Subjects: interoperability of topics beyond scope	2
1.4. And what about associations?	3
2. Conditions for associations merging	3
2.1. Identical associations	3
2.2. Introducing association patterns and templates	3
2.3. Similar associations	4
3. Minimal Requirements for Templates	5
3.1. An Association Template binds an Association Type to a set of Role Types	5
3.2. Cardinality of members for each Role Type	5
3.3. Class Constraints for Role Players	5
3.4. Example	5
3.5. Proposed syntax for templates	6

4. Some use cases	7
4.1. Checking consistency of Topic Maps	7
4.2. Workflow extraction and creation of associations matching existing templates	7
5. Further Perspectives	7
5.1. Towards Published Templates	8
5.2. Integration in TMCL	8
Bibliography	8

1. Topic Maps and Knowledge Interoperability

Topic Maps early adopters are now in the process to prove that the technology is a robust and efficient backbone for scalable knowledge and content management solutions, providing users with intuitive, subject-centric, semantic access to distributed resources. Examples and use cases are emerging in various domains: industry, science, education, government ...[\[DMOZ\]](#)

Nevertheless, all the above benefits have been proven in more or less closed environments, standing on controlled vocabularies and ontologies. But Topic Maps fathers and supporters have long ago and constantly claimed that one main benefit of the Topic Maps model, along with the above, is its potential capacity to support "global knowledge interoperability."

We'll try first to figure if that is more than a marketing buzzword, what supports so far such a claim, what pitfalls have been identified, what remains to be done, and what could be the next steps in that direction. We'll see that association patterns and templates could play a critical role on the road towards full interoperability.

1.1. A standard XML interchange format

XTM 1.0 [\[XTM\]](#) provides a syntactic interchange format allowing to exchange and merge topic maps whatever their origin. XTM makes for syntactic interoperability of Topic Maps, but the very generic nature of its tags does not allow to check if merging "makes sense". Note that "merging" will be used throughout this paper in a very broad sense, and may occur not only in the canonical case of merging two XTM documents, but also when loading of an XTM document in a Topic Map engine, updating a Topic Map through an editing tool or automatic feeding ... and whatever process in which a topic map has to be modified and processed to take into account new topics and associations.

1.2. Topic Naming Constraint: interoperability of topics inside a scope

Topics can be compared and merged on the basis of their name in a given scope. That is the controversial Topic Naming Constraint, that future versions of ISO 13250 [\[ISO\]](#) are about to relax somehow, but which is in fact a very interesting feature in closed environments using controlled vocabularies. It allows Topic Maps applications to leverage existing vocabularies and thesaurus, ensure semantic interoperability of XTM documents in a community of users using the same vocabulary, aggregating content around well-identified subjects defined by controlled terms. But the limits of such an interoperability are of course the limits of the controlled environment.

1.3. Published Subjects: interoperability of topics beyond scope

By setting topic identity on a URI subject identifier and human-readable subject indicator, Published Subjects are binding points that extend Topic Maps interoperability beyond controlled environments, since identity is no more linked to a name.

Development of Published Subjects is still in its infancy, since recommendations about them are still in draft stage in OASIS Technical Committees [\[PUBSUBJ\]](#). Published Subjects are likely to be developed first in wider environments than controlled vocabularies, but still rather closed, like communities or industries using multiple vocabularies for the same subjects. Multilingual environments are particularly interested, e.g., European Community could use Published Subjects to provide citizens uniform access to legal information about identical subjects, across national and linguistic borders.

1.4. And what about associations?

Be it based on Names or on Published Subjects, the resulting interoperability deals only with comparison and merging of Topics, and consequently binding or merging nodes representing the same subject in the Topic Map network. This is fine, but just a first step towards full interoperability, because Topic Maps are generally not only sets of topics (although they could be), but also networks of associations. And what remains to be seen is how associations are affected by merging, and how interoperable they can get.

2. Conditions for associations merging

2.1. Identical associations

XTM 1.0 in Annex F: XTM Processing Requirements (Informative)[[XTM](#)] defines precisely what are identical associations, and how they should be merged at processing time.

A conformant XTM processor must consider two associations to be equal if the following are true:

1. The associations are comprised of the same set of roles.
2. The set of topics playing each role in the associations are equal.
3. The associations are instances of the same class.
4. The scopes of the associations are equal as defined by the scope equality principle.

If two identical associations are found at merging time, one of them has to be removed, so a fully processed topic map should not contain two identical associations.

The current draft of Topic Maps Standard Applications Model[[SAM](#)] basically does not seem to change that viewpoint.

2.2. Introducing association patterns and templates

To go further into association interoperability, let's introduce the distinct notions of association patterns and templates.

- *Pattern* is to be understood hereafter as the structure of an individual association, be it explicitly declared outside the association itself, or implicit. It includes the association type and role types used, and possibly cardinality of members for each role, role type ordering, and class of role players for each role. Pattern is independent of value of individual role players topics, but can include the class of those topics.
- *Template* is to be understood as the formal declaration of a required pattern for a given association type. This declaration can be found either inside the Topic Map itself, either by an external reference, or built-in specific application feature. It could also be called *Schema*. This term has not been used here, first to avoid confusion with XML schemas, and second because "template" is widely used in topic maps community, even if it is not in the standards.

Note that above definitions are author's view. Neither of them is defined or even used by current Topic Maps specifications [[ISO](#)] [[XTM](#)], nor current draft of Topic Maps Standard Application Model [[SAM](#)]. "Pattern" has been introduced here for sake of clarity of presentation. Although it is rarely used so far in Topic Maps, a notable exception being the constraint language AsTMA! developed by Robert Barta[[ASTMA](#)], the word is widely used in related domains like graph theory.

"Association Template" is widely used in Topic Maps community, for example in Topicmaps.net's Processing Model[[PMTM4](#)], in TM4J API[[TM4J](#)], and in various vendors literature.

In fact, major vendors software use templates one way or another, for associations and other topic map objects. Mondeca <http://www.mondeca.com> and empolis <http://www.empolis.com/home/home.asp> use explicitly both

concept and name. Ontopia Schema Language [OSL] includes association templates (called here "association classes") in a general representation of constraints against which a topic map can be validated. Current Draft Requirements for Topic Map Constraint Language [TMCL] provides also an extended notion of template:

The expression of the constraints to which instances of a class of topic map object must conform.

The focus will be put in this paper on associations, because it is certainly at this level that lack of templates will trigger the biggest interoperability issues. Nevertheless, its rationale could be extended to other objects. We'll be back to that in the last section.

Even when they do not stand on explicit templates, most Topic Maps use consistent patterns throughout, which means templates exist at least implicitly in the Topic Map's author mind and knowledge model. But, if they are declared nowhere inside or outside the Topic Map, it's up to applications and users to figure how to parse, process or interpret those patterns in the most relevant and consistent way.

2.3. Similar associations

Situations are likely to occur where the same information is expressed in similar but not completely identical associations, in topic maps to be merged. Not pretending to be exhaustive, let's see some typical and relatively simple cases. Things can of course get much more complex than in those examples.

2.3.1. Identical patterns, but different declaration of association type

Example: two different Topic Map sources use the same pattern employer-employee, but first Topic Map declares those associations as instances of "employment", and the other as instances of "contract". No automatic merging would occur in such a case. But if the two association types are considered to be in fact identical, the merging of the association types will trigger the merging of instances. So this case is not difficult to handle.

2.3.2. Same declaration of association type, but different patterns used

To figure what it means, let's take a simple example.

Three Topic Maps about mechanical components in a given industry have used "system-component" associations. The first Topic Map, TM1, is published by a component integrator, and uses associations linking one system to all its components. The pattern for associations in TM1 is (system, component+). The second, TM2, is published by a component producer, and uses association linking a component to systems it's used in. The pattern here is (component, system+). A third one TM3 is using only binary associations (system, component).

Let's assume the best case, in which the three Topic Maps use the same controlled vocabulary and/or published subjects and are able to merge the topics representing the role types "system" and "component", and individual topics representing components and systems. What will happen on merging of the following, for example?

- From TM1 : (System: S1; Component: C1, C2, C3)
- From TM2 : (Component: C1; System: S1, S2, S3)
- From TM3 : (System: S1; Component: C1)

Even if merging can identify the nodes S1 and C1 in three sources, the information "C1 is a component of system S1" will be expressed three times in different forms, with no obvious way to get rid of the redundant representations. So, if two of the above associations are declared as instances of the same "system-component" association type, what should happen in processing is unclear, and every other application would try to reduce redundancies by some "ad hoc" processing. Otherwise, one can easily figure that accumulation of redundant information will happen after several merging operations.

In such examples, sticking to binary associations would of course avoid the problem, and that best practice could be set on pragmatic principle of keeping associations as simple as possible, which means binary whenever possible. But, to quote Einstein, "Things should be made as simple as possible, but not simpler". Some associations are non-binary by definition, and the Topic Maps paradigm would lose much of its expression power if it tried to reduce all associations to binary ones. For a Mozart string quartet interpretation, the pattern could be either "cello (one

player)", "alto (one player)", and "violin" (two players), or more accurately make distinct the roles "violin 1" and "violin 2", but would be difficult to reduce to binary associations.

2.3.3. Identical associations declared in different scopes

Last but not least, scope can add to the complexity of the situation. What is to be done when two associations are found identical, except for scope? The general principle should be that the associations should be merged, and the resulting scope for the merged association would follow the same rules as the ones applied when merging topics ... but those are in fact complex and not yet completely sorted out, and debate is ongoing about it.

3. Minimal Requirements for Templates

On content and structure of templates, even if there are various proposals on syntax and debates on terminology, and where and how the templates are to be declared, it seems that a minimal consensus is found between vendors practice and draft TMCL specification wording, even if the latter is a little less advanced than the former, on the following points.

3.1. An Association Template binds an Association Type to a set of Role Types

This is the minimal feature of an association template. Of course, topics representing the Association Type and Role Types are to be identified in the most interoperable way, which means certainly by Published Subjects. We'll be back to that point in the last section.

A real issue is to know if a given association type might be used in more than one template. The quoted TMCL draft [\[TMCL\]](#) considers the eventuality of multiple templates for a given association type.

There may be multiple templates for the same class of association. An individual association need only conform to one of them.

All the above examples tend to show that the most difficult situation to handle is exactly that one, so allowing several templates for an association type is certainly a viewpoint to reconsider. A reasonable requirement should be to have a unique template for a given association type. The advantage of such a constraint would be that it would not be necessary to check for template details to ensure that two associations have the same one, as long as they have declared the same type. And this is in fact the general practice. It is even an absolute requirement in software like Mondeca ITM.

3.2. Cardinality of members for each Role Type

An association template contains optional constraints on cardinality of members for each role type, expressed for example in terms of minimal and/or maximal number of topics allowed to play each specific role type.

3.3. Class Constraints for Role Players

A common requirement for templates is that a player for a given role type should belong to a specific class. This feature can in fact be used in two ways: either checking that a role player has the required type, or the other way round, have a topic inherit a class from its role type (this latter feature is implemented by MondecaITM) .

3.4. Example

The association template for an association of type "team" could contain the following constraints.

- Role Type = "unit"; card min = 1, max = 1 ; Player Type = "organizational unit"
- Role Type = "manager"; card min = 1, max = 1 ; Player Type = "person"
- Role Type = "member"; card min = 1 ; Player Type = "person"

3.5. Proposed syntax for templates

The following example shows an association of type "team", conformant to the previous template, expressed in XTM.

```
<association>
  <instanceOf>
    <topicRef xlink:href="team" />
  </instanceOf>
  <scope>
    <topicRef xlink:href="mondeca" />
  </scope>
  <member>
    <roleSpec>
      <topicRef xlink:href="unit" />
    </roleSpec>
    <topicRef xlink:href="mondeca" />
  </member>
  <member>
    <roleSpec>
      <topicRef xlink:href="manager" />
    </roleSpec>
    <topicRef xlink:href="jdelahousse" />
  </member>
  <member>
    <roleSpec>
      <topicRef xlink:href="member" />
    </roleSpec>
    <topicRef xlink:href="bvatant" />
  </member>
  <member>
    <roleSpec>
      <topicRef xlink:href="member" />
    </roleSpec>
    <topicRef xlink:href="bcarcenac" />
  </member>
</association>
```

The template itself could be declared as following.

```
<template>
  <assocType>
    <topicRef xlink:href="team" />
  </assocType>
  <roleType max="1" min="1">
    <topicRef xlink:href="unit" />
  </roleType>
  <roleType max="1" min="1">
    <topicRef xlink:href="manager" />
    <playerType>
      <topicRef xlink:href="person" />
    </playerType>
  </roleType>
  <roleType min="1">
    <topicRef xlink:href="member" />
    <playerType>
      <topicRef xlink:href="person" />
    </playerType>
  </roleType>
</template>
```

The above code could be added to the XTM Topic Map itself, with a corresponding backward-compatible extension of the XTM 1.0 DTD, adding the optional element <template> as child of XTM root element <topicMap>.

4. Some use cases

Whatever the way they are declared, templates could be used in various ways. Some examples are presented below.

4.1. Checking consistency of Topic Maps

A consistent Topic Map could be defined as a Topic Map where patterns are conformant to declared constraints. This can be limited to association templates, but could be extended to all possible constraints (see last section). Consistency can be checked at any time: before processing an XTM document, during edition of a Topic Map via an authoring tool, or before merging two XTM files. The latter case is the more interesting in terms of interoperability. Before merging the information of a Topic Map B into a master Topic Map A, patterns used by B are checked against templates declared in A. Or, more simply, only templates declared by B could be checked, if B has already been checked as consistent. If some templates declared in B are not found in A, two main options are possible.

- Add the templates declared in B, to the template list of A. That is blind merging, and likely to produce unexpected and weird results.
- Add to A only the associations from B found conformant to the templates of A. That seems like the best practice: relevant information from B will be added to A, without changing the model of A.

4.2. Workflow extraction and creation of associations matching existing templates

In many cases, a Topic Map will be maintained and updated through automatic feeding in an integrated environment. For example a Topic Map data base is built and updated for the Research Department of a major chemical company, for technological and strategic survey purposes. A text mining tool is used to parse and analyze news from various sources, looking for information like: "Company X has registered a patent by authority A, under reference K, for application W of product Z ". Since news about patents are released in quite standard ways, they are quite easy to parse by the text mining tool. Patterns found in news are transformed in associations matching pre-defined templates like "patent-product-application" "product-company" and "patent-authority". In a typical news item like above, the three following associations are created:

- Patent K, Product Z, Application W
- Product Z, Company X
- Patent K, Authority A

The partial Topic Maps, created on the fly from each news item, are merged to the master Topic Map data base, on a daily basis, for example. Topics are checked against the existing ones and merged as necessary, using either controlled vocabulary or other relevant identifiers. New topics are saved. What is very likely to happen is that the same news will be found from various sources. But since information is added only using pre-defined templates, any redundant information will lead to exactly identical associations that will be merged.

News sources are used as scopes, so the scope is augmented whenever the news is confirmed. The Topic Map can then be filtered by various criteria as "show news confirmed by such sources" or "show me news confirmed by at least three sources".

5. Further Perspectives

Beyond above applications, two roads are to be explored towards extended interoperability: creation of library of templates in the form of Published Subjects, and integration of templates in the more general framework of Topic Map Constraint Language.

5.1. Towards Published Templates

To ensure interoperability in a given industry, pre-defined templates should be available as Published Subjects. Since Published Subjects are used to identify topics only so-far, the identification mechanism should be supported by a one-to-one matching between association types and association templates. Since building an efficient set of templates for a given industry is costly high-level knowledge engineering, both interoperability and development cost considerations should push towards the development of such Published Templates.

5.2. Integration in TMCL

TMCL is still in draft stage of development, but there is a general consensus in Topic Map community that definition of association templates is just part of it, that has to be extended to similar constraints on other topic map objects: occurrences, names and identifiers. That's why some voices in the Topic Map community are pushing towards postponing any specification on association templates until TMCL is achieved. That makes sense, and it's indeed difficult to argue why association templates need a specific treat, and in what they are more important and urgent to deal with than other topic maps constraints.

What this paper has tried to show is that association templates are maybe more easy to grasp than some other constraints like e.g. those concerning topic names and identifiers or data types, or more tricky ones, like associations required for a given topic class, and that they have more obvious immediate and critical applications. Another argument is that there might be still a long way towards complete TMCL standardization, since the process has not even completed the requirements stage, and many issues are pending.

In that perspective, a "low bar" solution based on a simple extension of XTM for association templates (and maybe occurrences templates) would allow to implement quickly and test simple constraints in an easy and backward-compatible way. Such a "low bar" solution could be a test bed allowing to figure what the general TMCL could look like. As a matter of fact, a working meeting on TMCL will take place during this XML 2002 conference. One hope of the author is to have brought some useful contribution to this process.

Bibliography

- [ISO] **ISO/IEC 13250 Topic Maps** <http://www.y12.doe.gov/sgml/sc34/document/0322.htm>
- [XTM] **XML Topic Maps (XTM) 1.0** <http://www.topicmaps.org/xtm/1.0/>
- [SAM] **The Standard Application Model for Topic Maps** <http://www.isotopicmaps.org/sam/sam-model/>
- [PUBSUBJ] **OASIS Topic Maps Published Subjects Technical Committee** <http://www.oasis-open.org/committees/tm-pubsubj/>
- [PMTM4] **Topicmaps.net's Processing Model for XTM 1.0** <http://www.topicmaps.net/pmtm4.htm>
- [TM4J] **TM4J API Documentation : Association Template** <http://tm4j.org/docs/apiDocs/org/tm4j/topicmap/AssociationTemplate.html>
- [ASTMA] **Asymptotic Topic Map Languages**<http://astma.it.bond.edu.au/>
- [TMCL] **Draft requirements, examples, and a "low bar" proposal for Topic Map Constraint Language (TMCL)**<http://www.y12.doe.gov/sgml/sc34/document/0226.htm>
- [OSL] **The Ontopia Schema Language** <http://www.ontopia.net/omnigator/docs/schema/tutorial.html>
- [DMOZ] **Open Directory: Topic Maps** http://dmoz.org/Reference/Knowledge_Management/Knowledge_Representation/Topic_Maps/

Biography

Bernard Vatant

Mondeca

Paris

France

bernard.vatant@mondeca.com

Bernard Vatant is a former high school mathematics teacher, graduated in 1975 from ENSET (Cachan, France). His research interests have long ago been in knowledge representation and organization, singularly applied to science popularization (astronomy). He's been working since the end of Y2000 as a consultant for Mondeca, where he participates in the development of Topic Maps and vocabularies, and coordinates the SemanTopic Map project. He has been a participating member in the XTM Authoring Group, and is founding member and current chair of the OASIS Topic Maps Published Subjects Technical Committee.