

Topic Map cartography

a discussion of Topic Map authoring

Colin Baird

STEP UK Ltd., Swindon, United Kingdom

ctb@stepuk.com

<http://www.stepuk.com>

Abstract:

Topic Maps, implemented through the ISO/IEC 13250 standard, are designed to facilitate the organisation and navigation of large collections of information objects by creating meta-level perspectives of their underlying concepts and relationships. This paper will examine the issues involved in using the standard to create Topic Maps that enable this objective. As a so far unproved new technology, the presentation aims to begin the process of establishing 'good practice' methods for creating and maintaining these meta-level perspectives. It asks some key questions: How do I differentiate between Topic concepts? Is there such a thing as a bad and obstructive Topic? What is the best way to make my Topic associations make sense? How should I organise my topics, occurrences, scopes, themes and maps? What is a good way of preserving the longevity of my Topic Map?

Topic Maps may well develop as an organisation's fundamental perspective of their data, ranging from their core knowledge to their website. We can imagine Topic Map perspectives being used to organise, understand, present and drive any facet of their activity such as their research and development, management, services and marketing initiatives. In reality, the conceptualisation of meta-data from any given source is boundless, but it is inevitably prone to subjectivity either through direct human participation, or by the human creation of rules and patterns in automatic processes. Therein lies both its strength and weakness. One of the purposes of this paper is to examine the standard for mechanisms that may support a regularized and unambiguous approach to creating these perspectives. Where these mechanisms are absent or deficient, there needs to be some thought and discussion concerning additional means to support the authoring of them.

The paper will therefore seek to identify mechanisms within the standard that facilitate the creation of effective Topic Maps, ones that can withstand the rigors of multiple authorship, amendment and merging, yet still provide the author with the conceptual flexibility needed to create an effective representation of their data. Does the standard provide ways of answering the questions outlined above? If it does not, then we need to develop an additional framework to guide the creation of a good map and which enables the author to make that crucial differentiation between concepts, or that crucial expression of a relationship when and where they need to. What's more, this framework needs to be understood and preserved by any subsequent author and possibly even by the application providing the interface to it. There is no doubt that the Topic Map standard has raw power, but if an organisation cannot see how to encapsulate it effectively as a means of expressing their data at a useful level, this power will be wasted. This presentation will endeavour to begin the discussion that should attempt to address this important aspect of Topic Map implementation.

Introduction

The Topic Map ISO standard essentially defines a syntax that allows someone to create a strongly typed, linked model of an area of knowledge that they are familiar with. This model is a representational device, separate to any number of individual information objects that actually constitute part of that knowledge domain. It can be used to provide navigational access to that knowledge domain and help to describe the routes, or links, that connect together related parts of it. Also, because the syntax of the model is defined by SGML and XML DTDs, it is an ‘open’ model that can be shared with others.

Topic Maps are a tool for creating links between ‘things’ or concepts based on how they can be typed, named and associated together. One of the hardest things about Topic Maps is trying to understand the simplicity of the concepts that underlie the technology in tandem with the complexity of the ‘big picture’ that it is possible to derive from it. A Topic Map author starts from this very general platform but ultimately wants to use it to describe very specific instances of their knowledge.

An initial read-through of the ISO Standard can be misleading. Although it appears that there are only a few constructions used to define Topic Maps, it becomes clearer on closer inspection that they are extensively inter-dependent and strongly overloaded. The important lesson for the Topic Map author is this: whilst the Standard exposes an extremely powerful, generalised and open syntax, it consequently does not provide any specific structures or hierarchies into which their information models can be immediately, or obviously, moulded. As a result, a degree of planning and preparation is involved when approaching the task of creating a topic map.

Cartography

Landmarks

The first issue that a Topic Map author faces is the identification of the concepts that lie within their understanding of the knowledge that they wish to model. Any given item of information exudes concepts and ideas, either directly and explicitly from within the data, or from the understanding of the information within the author’s mind. There may be many different categories and types or concepts, some which are fundamental and some which are ancillary and of relatively low significance. The ‘topic’ syntax of the Standard provides the basis for the embodiment of these concepts. In the words of the Standard:

“In the most generic sense, a ‘subject’ is anything whatsoever, regardless of whether it exists or has any other specific characteristics, about which anything whatsoever may be asserted by any means whatsoever.”

This is an interesting starting point. Effectively, this statement means that there is no definition to which an author’s concepts have to adhere in order to be incarnated as topics within their topic map. What’s-more, all topics are defined by the same shell of syntax which lacks any external means of classifying its contents within a global hierarchy, cascade or tree. Instead, topics vary in terms of the characteristics that they are assembled from, the way they and their characteristics are typed and they way that they participate in associations with each other. Mastery of these elements enables the author to create a richness of definition that is lacking in other technologies.

It is from the consideration of these elements that a good start can be made in beginning the topic map. In contrast to many traditional approaches to information modeling, where an author begins at the top of a hierarchy or structure and works progressively deeper, a topic map author can benefit from taking a more holistic approach. This means that an author concentrates on some of the ways they might type and

define topic characteristics and associations first, and they would do this because the syntax enables topics themselves to be used for these purposes.

The Topic Map Standard provides for the creation of typed links, and we have mentioned that the way in which topics, topic characteristics and associations can be typed is an important aspect of the technology. When there is provision to add type information on topic map linking constructs it takes one of three forms. At the basic level, the element Generic Identifier is used, on top of that an author can provide a mnemonic such as a string literal, and in ascendancy of that the author can use a topic. This latter specification has great potential and its usage creates circuitry within the topic map. The benefits in comparison to the former approaches are arguably so tangible that the use of this typing mechanism should be strongly encouraged. At any rate, its use supports the authoring approach that is mentioned above whereby an author begins by identifying concepts that can be used for defining topic types, occurrence types, topic association role types and association types.

These sorts of topics can then be used as ‘landmarks’ around which an author’s thoughts can be organised. As a simple example, a collection of topics defining concepts such as: village, town, city, would be useful in helping the author to organise their approach in a context where many instances of these ‘types’ existed within the information. Indeed, the Standard refers frequently to the idea of “class-instance” relationships between topic map constructs when topics have been used to provide type information on links. However, it is correct to use this phrase guardedly, as it is consequently tempting to mistake this pattern as part of class hierarchy.

The topic map author may therefore benefit by beginning the task of creating a map with this approach, where some of the initial topics to be created are those that signpost intentions for adding type information to other topics, topic characteristics, topic association roles and associations. Topics to be used to define scope may also be considered at this stage. It is suggested that this is a beneficial approach because of the absence of inherent hierarchical structure within the syntax and the sense of ‘formlessness’ that this initially evokes. All topics are created equally, it is the subsequent characteristics they define and associative roles they play that give them their meaning, scope and variance. A topic map author therefore can define their own templates into which further organisation of concepts can take place and in doing so they can express their knowledge more effectively. Indeed these may in part define hierarchies, or other forms of classification. Moreover, the ‘templates’ they define are an integral, fundamental part of the topic map itself and this has major consequences for subsequent deployment to users. The following section helps to explain this point.

Routes

People need to have routes through information in order to be able to navigate and understand it. The more information that a user can obtain about a route, the more they can understand about what it means, why it has been created and what relevance it has to them. Topic Maps facilitate this goal; we have already mentioned how heavily typed the constructs can be and how the use of topics as type information creates circulatory routes through the information model. In addition, the correct use of associations is paramount to the creation of good routes that allow a user to understand the meaning behind the relationships that an author has created between topics. This meaning is dependent on the use of a clearly defined semantic for topic associations and topic association roles.

The semantics that underlie these are not always obvious and the ISO Standard must be read thoroughly in order to build up a good picture of them. What is often surprising when reading topic map instances, is the

frequency with which people are tempted to make meaning dependent on inference, as opposed to being a true property of a well defined semantic. It is apparent with topic map associations that people don't always mean what they say, or say what they mean. This is most obvious with simple examples where the concepts involved are well known and understood, and consequently where it is easy to fall into the trap of allowing a user's knowledge to do the work of assembling the meaning. For example:

```
<tmx:assoc ID="a1" tmx:type="#THE-JONES-FAMILY">
  <tmx:assocr1 tmx:href="#JOHN" tmx:type="#HUSBAND" />
  <tmx:assocr1 tmx:href="#MARY" tmx:type="#WIFE" />
  <tmx:assocr1 tmx:href="#CLARE" tmx:type="#SISTER" />
  <tmx:assocr1 tmx:href="#HOWARD" tmx:type="#BROTHER" />
</tmx:assoc>
```

The deficiencies of the association in this simple example may or may not be obvious, what is important is that it might be easy for a user to view this association and assemble an idea of the 'Jones Family' based on inference rather than on the explicit meaning contained therein. If we apply the strongly typed semantic that helps us define association roles: "Topic A plays the role 'Topic B' in the association 'Topic C'", then we see that this association may not really represent the relationship of ideas that the author intended. This was to represent the members for the family and give meaning to their membership. In the case above, to state that "Clare plays the role of 'sister' in the association 'The Jones Family'" is arguably ambiguous and misleading. An alternative model might be:

```
<tmx:assoc ID="a1" tmx:type="#FAMILY">
  <tmx:assocr1 tmx:href="#JONES" tmx:type="#FAMILY-SURNAME" />
  <tmx:assocr1 tmx:href="#JOHN" tmx:type="#FATHER" />
  <tmx:assocr1 tmx:href="#MARY" tmx:type="#MOTHER" />
  <tmx:assocr1 tmx:href="#CLARE" tmx:type="#DAUGHTER" />
  <tmx:assocr1 tmx:href="#HOWARD" tmx:type="#SON" />
</tmx:assoc>

<tmx:assoc ID="a2" tmx:type="#MARRIED-PARTNERS">
  <tmx:assocr1 tmx:href="#JOHN" tmx:type="#HUSBAND" />
  <tmx:assocr1 tmx:href="#MARY" tmx:type="#WIFE" />
</tmx:assoc>

<tmx:assoc ID="a3" tmx:type="#SIBLINGS">
  <tmx:assocr1 tmx:href="#CLARE" tmx:type="#SISTER" />
  <tmx:assocr1 tmx:href="#HOWARD" tmx:type="#BROTHER" />
</tmx:assoc>
```

Whilst the above definitions require a more extensive outlay of structure, it can be argued that the meaning based on topic association semantics is clearer, less ambiguous and as a result, more powerful. It certainly tells us more about topic concepts such as 'Clare' and there are better-defined routes through the model. What's more, this model can be extended so that association templates can be created for 'Family', 'Married-partners' and 'Siblings' that describe these relationships in a more general sense. For example, the author may define an association where the constituents of a generalised 'family' relationship were defined. This would be helpful where there were large sets of 'instances' of associations that were of similar formation and it also fits with the generalised approach to topic map authorship discussed earlier.

The creation of well-defined, strongly typed routes is therefore dependent on the creation of effective and unambiguous topic map associations. In the example demonstrated above, it is easy to make inferences about the meaning of the association because the related concepts within the idea of a 'family relationship'

are already well understood by users. This may well not be the case in other examples and therefore the author must ensure that they take care to review the associations they create with the proper semantics in mind.

Boundaries

Another important issue that a topic map author should consider at an early stage is the possibility to define scope for many topic map constructs. This enables the author to create definitions in which the validity of topic characteristics can be ascertained. Again, it is possible to use topics to provide the information that is applied as scope upon other topic map constructs and therefore the author may benefit by considering the topics they would use for this at an early stage.

Scope (and the identity attribute) is important for the long-term evolution of the topic map. They can help to differentiate topics and associations that have consistent structure within different topic maps, thereby facilitating topic map merging. Where a group of authors are working within related information domains, it may be essential that a preliminary step in the process be concerned with planning the use of scope and identity. As the topic map community grows, and the number of in-use maps increases, it is likely that scope will become very important. It is suggested that a discussion of consistent definitions of how scope and other similar constructs could be applied is initialised within the topic map community at an early stage.

Automation

Depending on circumstance, it might be possible to generate certain amounts of topic map constructs by automatic process. For example, indexes or databases may already provide enough information to create parts of topic maps or parts of topics such as lists of occurrences or alternative names. The possibility of automatic generation should not be ignored, especially where clear patterns can be identified within existing information structures that translate into the topic map paradigm. In some cases where the information set is large, the use of an automatic process may well be required at some point.

It may be possible to identify areas within the Topic Map syntax where it would be easier to create the rules needed for automatic generation than others, depending on the nature of the data at hand. Some initial experiments have found that in certain conditions, sets of occurrences could be generated from relational databases, or sets of topic names could be expanded to include alternative and foreign spellings. It may be that an automatic process could generate parts of topics or some characteristics that would then need refinement from a human author. This would at least remove some overheads of manual authoring, although inevitably there would consequently be issues of feedback.

So far, it seems that automatic processes are most successful where the available information structure is already highly typed and heavily marked-up. This makes sense of course, we have already stated that one of the most powerful aspects of topic map technology is the ability to provide strongly typed links and associations. What is important when considering automatic processes, is to make sure that the information model encapsulated by the topic map represents the knowledge of its author, rather than the ability to perform the automation. The quality of a map suffers greatly if the potential to create structures with good semantics and strong typing is compromised.

Perhaps the most effective automatic processes created for Topic Maps in the near future will therefore be the creation of authoring environments and interfaces that support the human role and help the author to capture the knowledge that exists for an information set. Such applications could be used to maintain maps and for verifying the link structures as they change over time. This may be a more useful strategy in the long

term, because topic map technology is so generalised and also because automatic generation is likely to be based upon very specific conditions.

Conclusion

This paper has discussed several areas of topic map authorship where guidelines and good practices may help authors to produce better information models. It is very much a preliminary discussion and it is hoped that some more concrete guidelines can be discussed and produced as the topic map community grows and more people become involved within it. Topic map technology has great potential in enabling information architects to create more meaningful representations of their knowledge and making the dissemination of that knowledge more effective.

Bibliography

- [1] Michel Biezunski, Martin Bryan, Steve Newcomb (Editors), ISO/IEC 13250 Topic Maps

Author

Colin Baird

IT Consultant

STEP UK Ltd.

Postal Address:

Unit B, Dorcan Complex

SN3 5HQ Swindon

Wiltshire

United Kingdom

Telephone: +44 (0) 1793 485465

Fax: +44 (0) 1793 485451

E-mail: ctb@stepuk.com

Web: www.stepuk.com

Colin Baird - Colin Baird is an IT Consultant at STEP UK with several years experience in Web technology and development. An archaeology graduate of Exeter University, and a MSc graduate in Electronic Publishing from City University London, Colin has recently been involved with the development of Topic Map technology at STEP UK.